


Averaging multiple facial expressions through subsampling

Luyan Ji, Gilles Pourtois & Timothy D. Sweeny

To cite this article: Luyan Ji, Gilles Pourtois & Timothy D. Sweeny (2020): Averaging multiple facial expressions through subsampling, *Visual Cognition*, DOI: [10.1080/13506285.2020.1717706](https://doi.org/10.1080/13506285.2020.1717706)

To link to this article: <https://doi.org/10.1080/13506285.2020.1717706>

 View supplementary material [↗](#)

 Published online: 23 Jan 2020.

 Submit your article to this journal [↗](#)

 View related articles [↗](#)

 View Crossmark data [↗](#)



Averaging multiple facial expressions through subsampling*

Luyan Ji^{a,b}, Gilles Pourtois^a and Timothy D. Sweeny^c

^aDepartment of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium; ^bDepartment of Psychology, University of Hong Kong, Hong Kong, People's Republic of China; ^cDepartment of Psychology, University of Denver, Denver, USA

ABSTRACT

When perceivers view multiple facial expressions shown concurrently, they can quickly and precisely extract the mean emotion from the set. Yet it is not clear how many faces in the set contribute to summary judgments, and how the variance among them influences this process. To address these questions, we used the subset manipulation and varied emotion variance of faces in the sets across three experiments. Sets containing sixteen faces, or a subset of faces randomly selected from the sixteen-face display were presented, and participants judged the average emotion of each face set on a continuous scale. Results showed that when emotion variance was relatively large (Experiments 1 & 2), only two faces in the set contributed to ensemble representations. In Experiment 3 where the emotion variance was smaller, around three to four faces were likely sampled. However, when directly comparing results from Experiments 2 and 3, there was no strong evidence supporting the impact of variance in averaging efficiency. Altogether, these new results suggest that the process of averaging multiple emotional facial expressions can be explained by capacity-limited subsampling. The claim that ensemble representations are capacity unlimited or can overcome the bottlenecks in visual perception might need to be reconsidered.

ARTICLE HISTORY

Received 23 July 2019
Accepted 2 January 2020

KEYWORDS

Ensemble representation;
facial expressions; sampling;
variance

Introduction


It has been well established that human observers have the remarkable ability to integrate information from multiple visual features or stimuli, and extract summary statistics (e.g., the mean) about these features (Alvarez, 2011; Whitney & Yamanashi Leib, 2018). This kind of ensemble coding, or summary statistical representation provides a rather accurate and stable impression of the visual world (Cohen, Dennett, & Kanwisher, 2016; Corbett & Melcher, 2014). Faces are one example of high-level, multidimensional visual stimuli that can be integrated with one another. For example, mean emotion can be extracted very rapidly and quite precisely from multiple facial expressions shown briefly (Haberma & Whitney, 2007, 2009; Li et al., 2016), even when limited attentional resources are available (Ji, Rossi, & Pourtois, 2018).

However, to date, the mechanisms underlying ensemble perception are still largely unclear and debated, especially when high-level objects such as facial expressions are considered. An open question

remaining in the existing visual cognition literature pertains to the power or efficiency with which processing multiple facial expressions can be achieved. For instance, one could wonder how many faces in a set, if not all, contribute to the perception of that crowd's mean emotion. This question is reminiscent of a current debate in the literature about the actual meaning and function of ensemble representations, and more specifically whether they offer a means to surpass traditional bottlenecks in information processing and attention selection (Alvarez, 2011; Attarha, Moore, & Vecera, 2014; Chong & Treisman, 2005; Cohen et al., 2016; Ji, Chen, Loeys, & Pourtois, 2018). If sampling only 3–4 items (this size being traditionally assumed to correspond to the upper bound for attention capacity; see Pylyshyn & Storm, 1988) or even fewer items in a set could adequately explain human observers' performance when many more items are shown in the visual field, then the claim that ensemble representations are capacity unlimited or can overcome the bottlenecks in visual perception might need to be reconsidered.

In order to examine sampling efficiency, or more specifically, whether participants actually integrate

CONTACT Luyan Ji  luyanji@hku.hk

* Supplemental data for this article can be accessed <https://doi.org/10.1080/13506285.2020.1717706>

© 2020 Informa UK Limited, trading as Taylor & Francis Group

information from multiple objects or sample only a subset of items from the set, a paradigm known as the subset manipulation has been introduced and used successfully in several studies in the past (Chong, Joo, Emmanouil, & Treisman, 2008; Sweeny, Haroz, & Whitney, 2013; Wolfe, Kosovicheva, Yamanashi Leib, Wood, & Whitney, 2015; Yamanashi Leib et al., 2014). This procedure allows researchers to estimate the size of the subset participants use to compute the average of multiple items, for example using a subset of 3 items although 16 items are available and visible in the complete set. Using this procedure, one assumes that seeing a subset of faces largely resembles the act of subsampling these faces from a full set. It allows one to compare how precisely participants are able to estimate the mean of the full set when all items are visible (i.e., whole set) to conditions where only some of the set constituents are visible (i.e., subsets), with a systematic variation of this number across them. In the example above, if a participant's strategy is to extract information from only one randomly sampled item and then use this information to judge the entire crowd, the participant's estimates relative to the mean of the whole set will remain the same regardless of the number of items made visible (i.e., subset size). On the other hand, if this participant samples 3 items, performance will improve with increasing subset sizes (from 1 to 3), but will plateau when larger subsets (i.e., 4 or larger) are used. Hence, averaging performance for the whole set is compared to subsets with variable sizes, with the aim to infer an estimate of the number of items eventually used by the participants in this former case. Computational modelling has confirmed the logic of this framework (Sweeny & Whitney, 2014)—when performance in the subset condition matches that in the whole-set condition, it is assumed that a roughly equivalent number of items in the subset has been used in the estimate of the average for the whole set.

Using this methodology, Chong and colleagues (2008) previously found that the accuracy of mean size perception was better when an entire display of 16 items was presented than when only subsamples of two to four circles were shown to participants, suggesting that at least more than four items were integrated during averaging. Similarly, some earlier studies reported that ensemble judgments about facial expressions, identities and gaze directions also

became more precise as the number of displayed faces increased (Sweeny & Whitney, 2014; Wolfe et al., 2015; Yamanashi Leib et al., 2014). For instance, Wolfe and colleagues (2015) showed an overall decrease in response errors (relative to the entire set of 24 faces) for estimates of mean expression when the number of faces presented increased from 1 to 12, indicating that participants likely integrated multiple items if available; with this number clearly exceeding the limitations of attention selection (Pylyshyn & Storm, 1988). By contrast, Haberman and Whitney (2010) found that mean emotion discrimination performance (relative to the whole set) in their subset condition with four faces did not significantly differ from their condition in which the whole set of 12 faces was presented, though the latter was better than the conditions with subset of one to three faces, which suggested that sampling four faces could adequately explain the performance of judging mean emotion from 12 faces.

Thus, discrepant findings have been reported in the literature with regard to how many items are sampled and integrated when people make summary statistical evaluations of crowds. One factor that could potentially impact sampling efficiency, and thus explain this lack of consensus, is the variance of the set. Previously, we found that inter-item variance strongly influenced the formation of mean emotion representations (Ji & Pourtois, 2018). When the variance of emotional facial expressions in the set was relatively large, increasing set sizes led to poorer averaging performance, which suggested capacity limitations of ensemble representations for them (Experiments 1 & 2; Ji & Pourtois, 2018). Moreover, when the emotion variance was small, averaging performance remained unaffected by increasing set sizes (see Experiment 3 in Ji & Pourtois, 2018; see also Im et al., 2017; Marchant, Simons, & de Fockert, 2013). This result could be interpreted as indicating that the emotion averaging process is capacity unlimited, at least under these experimental conditions. Nonetheless, complementary simulation results suggested that sampling a limited number of faces (one to four) could adequately explain averaging performance observed across the different set sizes (see Supplementary Figure 3). As a matter of fact, when the variance was relatively small (see Ji & Pourtois, 2018), redundancy among items in the set was high, and accordingly, if participants sampled just a few items, these were necessarily

representative of the whole set. Yet, the sampled items would have been less representative of the whole set when the variance was large compared to small. The efficiency of sampling thus appears to be intuitively related to the variance of the set, and we examined these two factors simultaneously in the current study.

Notably, the subset manipulation, which we used here, is distinct from the set size manipulation we previously used (Ji & Pourtois, 2018) in two important ways. First, in the subset manipulation, a large set of items with some emotion variance therein is created on each trial, and either the entire set or a subset of items *randomly* selected from the whole set is presented. Participants thus estimate the average based only on the items that are visible to them. In comparison, when the set size manipulation is used, even though sets of different sizes are displayed, the items in these sets are not randomly selected from a larger set, and regularity remains relatively constant across different set sizes. Second, in the subset manipulation, the accuracy of averaging performance is always calculated relative to the whole set, even when that entire set is not visible to participants. In the set size paradigm, accuracy is computed relative to the faces that are actually presented. In essence, the subset approach provides an estimate of how many items participants use to extract the mean of a large set with a fixed number of items, whereas the set size approach reveals how precisely participants form ensemble representations with variable sizes of items used.

The current study aimed to examine whether sampling efficiency is impacted by the emotion variance of sets when estimating the average emotion of multiple faces. We used the behavioural subset manipulation (Wolfe et al., 2015; Yamanashi Leib et al., 2014) to provide an index of the number of faces likely sampled by participants to establish the mean representation, and manipulated emotion variance across three experiments (Experiment 1: large, Experiment 2: medium, Experiment 3: small), similarly as we did in our previous study (Ji & Pourtois, 2018). In the current investigation, we created whole sets of 16 faces (as in Ji & Pourtois, 2018), and on each trial we presented all of these 16 faces, or a subset of 1, 2, 4, or 8 faces *randomly* selected from the whole set. As in our previous work, averaging performance was calculated by computing the absolute difference between participants' average emotion judgments and the subjective mean emotion of the entire set

(i.e., subjective difference scores), which were calculated based on individual-specific subjective ratings of faces obtained from a separate emotion rating task. In addition, we also calculated the objective mean of all faces in each set (based on morph units) as in many previous investigations (Wolfe et al., 2015; Yamanashi Leib et al., 2014), and then computed the absolute difference between the average emotion judgments and the objective mean (i.e., objective difference scores). Importantly, following the logic of the subset manipulation, we calculated the mean emotion across all 16 faces (the whole set), no matter how many faces were actually presented (subset sizes) to participants. We then compared estimates of the subsets against this mean of the whole set, even though participants could not have based their estimates on the whole set since it was not visible to them (nor could they have known that they were viewing subsets of a larger set). Based on our previous studies (Ji et al., 2018; Ji & Pourtois, 2018), here we hypothesized that only a limited number of faces (e.g., within the capacity of attention, namely 3–4) could be sampled during averaging, and consequently, adding more faces to the display would not further improve averaging performance. Regarding the main question of our study—whether sampling efficiency is impacted by the variance of sets during averaging—if a similar number of faces was sampled across these three experiments that varied in emotion variance, then we could conclude that this process is probably independent from inter-item variance. However, if this number varied across the three experiments, then we could conclude that the efficiency of subsampling depended on the variance of the sets.

General methods

Participants

The three experiments included separate samples of twenty-four participants from Ghent University (Experiment 1: 18–28 years, 20 females; Experiment 2: 18–31 years, 22 females; Experiment 3: 18–27 years, 14 females). The sample size of 24 was determined a priori. We conducted a power analysis based on an effect size from a previous investigation using a similar subset manipulation (Sweeny & Whitney, 2014, Experiment 1, $\eta_p^2 = .578$), with α at .05. This analysis estimated a sample size of three to obtain power of 0.95 (1- β). Here, we decided to enlarge our

sample size to be consistent with our previous behavioural study (see Ji & Pourtois, 2018), and also because we hoped to observe individual differences in emotion ratings which would then allow us to conduct analyses using individual-based subjective indices of averaging performance in addition to objective indices. All participants gave written informed consent and were compensated 10 Euro per hour. They reported to be right-handed and had normal or corrected-to-normal vision. The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee.

Stimuli

All face stimuli were the same as used in our previous study (Ji & Pourtois, 2018), including sixteen different identities, each showing angry, happy and neutral expressions with closed mouths, selected from the NimStim database (Tottenham et al., 2009). Face images were morphed using FantaMorph 5. For each identity, the morphing was carried out between angry (Face 1) and happy expressions (Face 50) in Experiment 1. In Experiments 2 and 3, morphing was carried out between neutral (Face 1) and the apex of the corresponding happy expressions (Face 50), or between the apex of the angry (Face 1) to the corresponding neutral expressions (Face 50) (Figure 1A). The increase/decrease in emotion intensity between two adjacent images was denoted as one morph unit. Here, these morph units were arbitrary and did not necessarily reflect equivalent changes in subjective perception of emotion. However, using similar stimuli, we did show that the changes in morph units correlated with the average emotion judgments (Ji & Pourtois, 2018). Each face image subtended a visual angle $4.03^\circ \times 4.28^\circ$ and was presented against a homogenous black background.

The face sets visible to participants consisted of 1, 2, 4, 8, or 16 identities conveying different emotional intensities. In the 16-face set, we first randomly selected a mean emotional intensity and then four unique morph units surrounding the mean for each trial. There were four instances of each morph unit. The smallest distance between these morph units was six (mean ± 3 , ± 9 ; as used in previous studies, see Haberman & Whitney, 2007, 2009) in Experiments 1 & 3 (Figure 1A, top scale), and 12 (mean ± 6 , ± 18) in Experiment 2 (Figure 1A, bottom scale). The variance

of the face sets was largest in Experiment 1, intermediate in Experiment 2, and smallest in Experiment 3 (variance for Experiments 1 and 3 differed despite equivalent morph distances because we used different sets of face morphs for these experiments; see Supplementary Method and Supplementary Figure 1). Face sets did not include the endpoints of the morph ranges (either Face 1 or Face 50). The mean of each set of faces was randomly selected on each trial from a uniform distribution of morph units ranging from 11 to 40 in Experiments 1 & 3, and from 20 to 31 in Experiment 2. The mean varied in each trial and was never represented by an individual face in the set on that trial. The 16 morph units in each set were randomly paired to the identities within the set with the limitations that (1) there were always the same numbers of female and male faces in the sets, and (2) face identities within the sets were never repeated. For the subset conditions, one, two, four, or eight faces were randomly selected from the original 16-face sets. In the 1-face subset condition, there was an equal probability of presenting a female or a male face. The gender distribution was balanced in the other subset set conditions.

The 16 faces were presented in an invisible 4×4 matrix ($14.83^\circ \times 20.35^\circ$) centred on the screen. The subsets of one, two, four, or eight faces were randomly located in any of the 16 cells evenly distributed across the matrix (Figure 1B), leading to an overall lower spatial density (i.e., more sparse) in the sets with smaller subset sizes than in the sets with larger subset sizes. Clustering the faces from the subset near the centre of the matrix would have introduced a confound of reduced eccentricity for the smaller subsets.

Apparatus and procedure

We used a 17" CRT screen with a refresh rate of 85 Hz, and the viewing distance was roughly 60 cm. We used the same average-emotion judgment task as in Ji and Pourtois (2018), as well as a similar familiarity and practice phase beforehand. A fixation cross first appeared at the centre of the screen for 500 ms¹, followed by a face set which consisted of 1, 2, 4, 8, or 16 faces, presented for 500 ms. The faces in the set were immediately masked by scrambled face images presented for 100 ms. The scrambled image was created by dividing one randomly selected face into 100 square

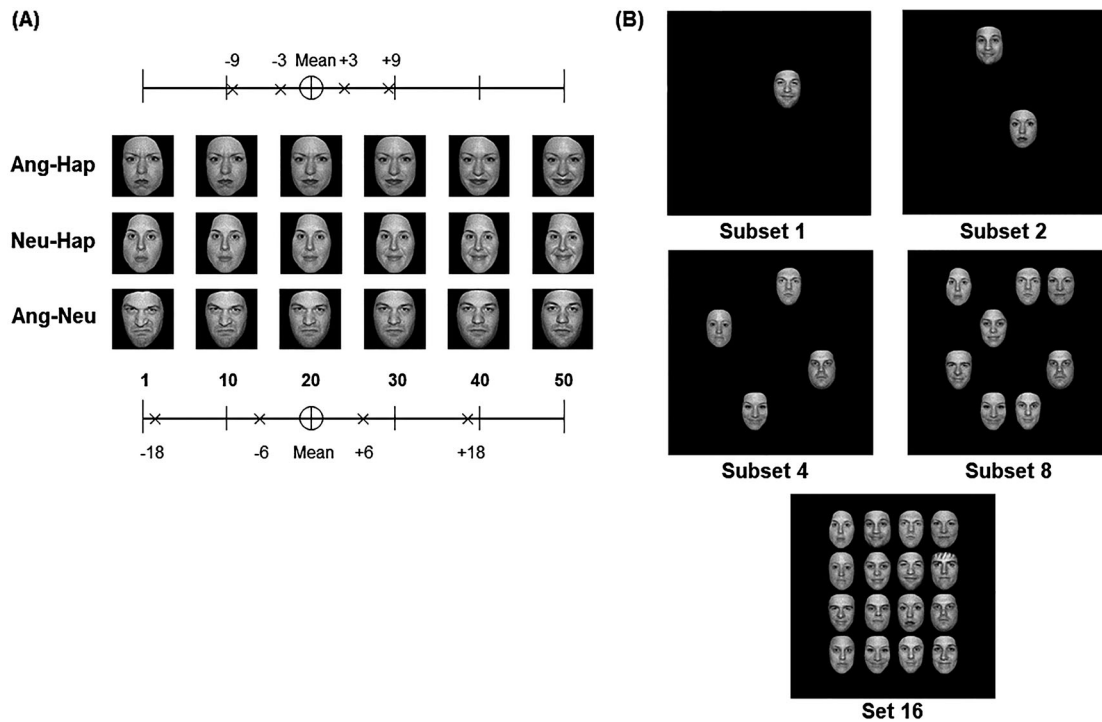


Figure 1. (A) Examples of faces morphed from angry to happy used in Experiment 1, and from neutral to happy or from angry to neutral used in Experiments 2 & 3. For each continuum, 50 different images were generated for each face identity. The scales above and below the face continuums illustrate the construction of face sets with variable emotional expression in Experiments 1 & 2 (mean ± 3 , ± 9) and Experiment 3 (mean ± 6 , ± 18), respectively. (B) Example of face sets used in the three experiments, which consisted of a subset of 1, 2, 4, 8 faces, or a whole set of 16 faces. The position of each face in the subset was randomly selected from 16 possible cells, corresponding to an invisible 4 \times 4 matrix.

pieces, randomly shuffling the locations of the pieces, and then cropping it to a shape of an average face with a black background, using a custom MATLAB script and Adobe Photoshop. The positions of the faces in the (sub)sets and those of the masks were identical. The next trial started automatically 1000 ms to 1200 ms after participants responded (Figure 2). Participants judged the average emotion of each face set on a visual analogue scale (VAS) by clicking on a unique horizontal location on the scale with the mouse. The VAS had endpoints labelled as *Extremely negative* and *Extremely positive*, respectively, and with the middle point labelled as *Neutral*. The positions of the two labels (negative on the left and positive the right, or the other way around) were counterbalanced across participants. We used a VAS for two reasons. First, this approach allowed us to avoid presenting a test face against which participants had to compare their average, and hence we limited biases/reference issues that might have emerged from presenting different test-face identities. Second, because we also used a VAS for the post-experiment

ratings of the individual faces, using the same VAS during the averaging task considerably eased the computation of subjective differences scores (described in more detail later).

In Experiments 2 & 3, participants were required to judge the average emotion of the set from neutral to extremely positive (half of the scale) for happy faces, or from neutral to extremely negative for angry faces in different blocks. The subset size (1, 2, 4, 8, 16) and the mean emotion (morph units from 11 to 40 in Experiments 1 & 3, and from 20 to 31 in Experiment 2) of each face set were randomized within blocks. Every trial had a unique face set to minimize statistical regularity across trials. In Experiment 1, participants performed three experimental blocks of 90 trials. In Experiments 2 & 3, the emotion category (happy, angry) was blocked, and for each emotion category, participants performed two experimental blocks of 120 and 150 trials, respectively. The happy and angry blocks were performed alternately, and the emotion used in the first block was counterbalanced across participants. Following the main task,

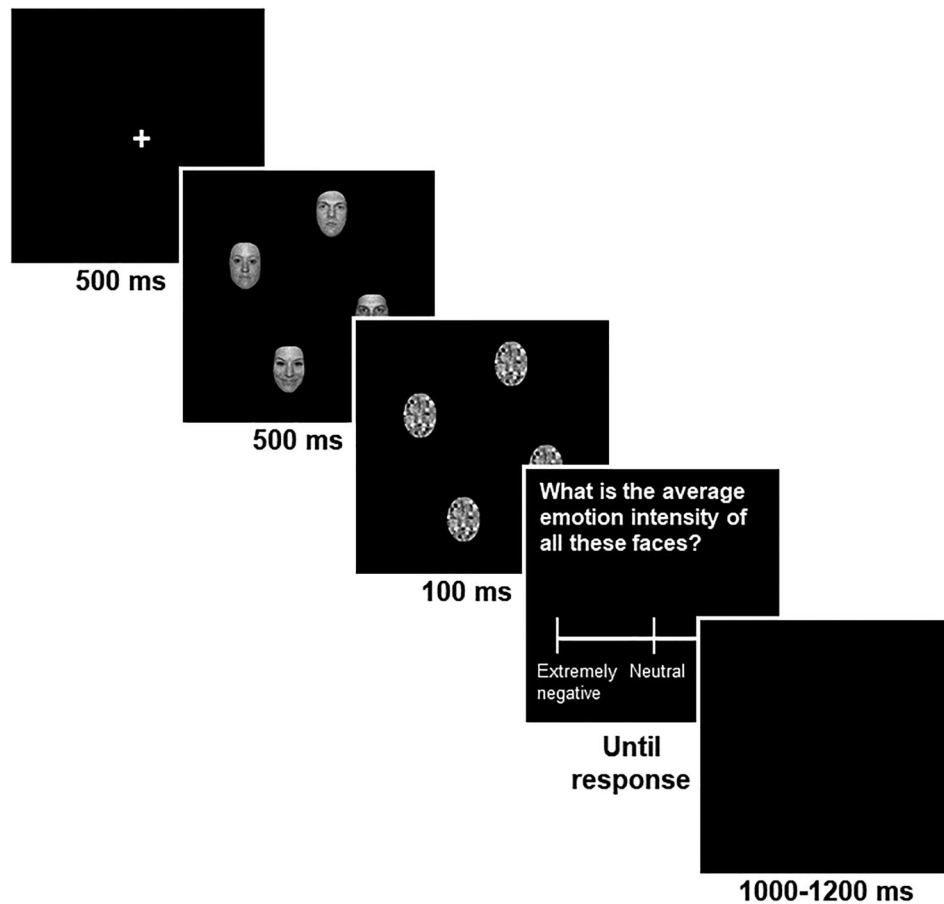


Figure 2. Average emotion judgment task in Experiments 1-3. Participants judged the perceived average emotion intensity from each face set on a visual analogue scale, ranging from extremely negative to extremely positive (the order of the two anchors was counter-balanced across participants). In Experiments 2 & 3, participants were asked to use half of the scale, from neutral to extremely positive in happy face blocks, and from neutral to extremely negative in angry face blocks. The sets contained 1, 2, 4, 8, or 16 different emotional faces.

participants rated the emotion intensity and arousal of each individual face (Face 1 and Face 50 only) on VASes, similarly to Ji and Pourtois (2018). The VAS post-test for emotion intensity had two anchors that were the same as those used in the average emotion judgment task (*Extremely negative* and *Extremely positive*). For arousal, the VAS post-test had the labels *Extremely calm* and *Extremely excited*. The labels shown on the left and right sides of the VAS were counterbalanced across participants. The two tasks were programmed and controlled using E-Prime Version 2.0 software.

Data analysis

Data conversion

The actual positions participants clicked on the VAS in the average emotion judgment task and the emotion rating task were converted to data ranging from 0 to

100 in all three experiments, similarly to Ji and Pourtois (2018). After conversion, the larger the value, the more positive the participants judged the (average) emotion; and conversely the smaller this value, the more negative the (average) emotion was perceived. The morph units of each face (1–50) were also converted to match the range of the converted average emotion judgments; the larger the morph unit, the more positive the face stimulus was, and the smaller the morph unit, the more negative the facial expression was. For each trial, we extracted the *whole-set based objective* difference score by subtracting the average emotion judgment from the mean morph unit of all 16 faces in each face set, no matter how many faces were actually made visible to the participants. A *whole-set based subjective* difference score was calculated by subtracting the converted average emotion judgment from the computed mean emotion intensity of the whole set of 16 faces, based on participant-specific emotion ratings for the

faces (made at the end of the experiments, see above). We also computed the *subset-based objective* and *subset-based subjective* difference scores by subtracting the average emotion judgments from the mean units and mean emotion intensity of the physically presented faces (the subset—the faces participants could actually see) in each set, respectively. This was done in order to examine whether we could replicate our previous findings using the set size paradigm (Ji & Pourtois, 2018). We then computed the absolute values of all the difference scores. Thus, the larger the whole-set based and subset-based absolute difference scores, the worse the averaging performance, relative to the whole set or to the subset, respectively. Notably, the absolute difference scores did not capture potential response biases. An additional analysis on the signed difference scores showed that positive response biases were present in all three experiments, but they were not impacted by variance, and they were similar for happy and angry faces (see Supplementary materials).

Data trimming

For the average emotion judgment task, trials with RTs exceeding 2.5 SDs above or below the mean RT for each participant (overall 2.8%, 2.5% and 2.3% of trials in Experiments 1-3, respectively) were excluded. This standard cutoff was chosen before running data analyses, similar to Ji and Pourtois (2018). Another 1.7%, 1.1% and 1.6% of trials with mouse clicks falling excessively far away from the scale (2.5 SDs above or below the mean position occupied by the scale on the screen) were excluded in Experiment 1, 2 and 3, respectively. Since participants were required to use a scale ranging from neutral to extremely positive or from neutral to extremely negative for the happy and angry blocks respectively in Experiments 2 & 3, the mouse clicks on the wrong part of the scale (e.g., judgment on the scale ranging from neutral to extremely positive in the angry blocks) were also removed from the analyses, leading to the exclusion of 2.5% of trials in Experiment 2 and 1.9% of trials in Experiment 3. In total, 4.4%, 5.9%, and 5.8% of trials were removed from the subsequent analyses in Experiment 1, 2 and 3, respectively.

Data analysis

To assess whether and how whole set-based and subset-based objective and subjective absolute

difference scores differed with increasing subset sizes (1, 2, 4, 8, 16), we conducted multilevel analyses with random intercepts for each participant using the lme function available in the nlme package for R (Pinheiro, Bates, DebRoy, Sarkar, & Core Team, 2017). The null model with no fixed effects was first built, and then the fixed effect of subset size (treated as a categorical variable) was added to the model. In Experiments 2 & 3, the fixed effect of emotion and the interaction between subset size and emotion was also introduced to the model sequentially, following the previous fixed effect. Each model was compared to the previous model by likelihood ratio tests to examine whether the added component contributed to averaging performance significantly. The models were fit for the four kinds of difference scores separately. The results of the final models with the best goodness-of-fit (smallest Akaike information criterion, Akaike, 1974) are reported (see Results). If one factor was significant or there were significant interactions between factors, we conducted post hoc (paired-samples t tests) and simple effect analyses (Chi-Square tests of one factor on different levels of another factor) for the final model using emmeans andphia (testInteractions) packages in R, respectively. Degrees of freedom were estimated with the containment method. A Bonferroni correction was used whenever multiple comparisons were performed. The reported descriptive results (mean and standard deviation) are based on single-trial data.

Results and discussion

Experiment 1

Whole-set based objective difference scores

There was a significant effect of subset size, $\chi^2(7) = 133.84$, $p < .001$. Post-hoc analyses showed that the whole-set based objective difference scores in the subset1 condition ($M = 23.28$, $SD = 14.87$) were significantly larger than those in all the other subset conditions, $t_s > 8.20$, $p_s < .001$ (Figure 3A). On the other hand, the objective difference scores did not differ from each other in the subset2 ($M = 19.13$, $SD = 14.06$), subset4 ($M = 18.53$, $SD = 13.84$), subset8 ($M = 18.26$, $SD = 13.91$), or the set16 conditions ($M = 18.84$, $SD = 14.02$), $-1.14 < t_s < 1.69$, $p_s > .90$. The finding that the whole-set based difference scores did not decrease beyond the subset2 condition suggests

that participants were perhaps pooling information from only two faces.

Whole-set based subjective difference scores

Subset size contributed to the average emotion judgments significantly, $\chi^2(7) = 161.54, p < .001$. Similar to the objective difference scores, the whole-set based subjective difference scores were largest in the subset1 condition ($M = 24.57, SD = 14.34$), $t_s > 8.73, p_s < .001$. The difference scores in the subset2 ($M = 20.43, SD = 13.54$), subset4 ($M = 19.63, SD = 13.20$), subset8 ($M = 19.22, SD = 13.14$) and set16 conditions ($M = 20.21, SD = 13.18$) did not differ from each other significantly, $-2.12 < t_s < 2.50, p_s > .13$. This finding

again suggests that participants were integrating only a limited number of faces, namely two in the present case.

Subset based objective difference scores

There was a significant effect of subset size, $\chi^2(7) = 12.25, p = .016$. Post-hoc analyses showed that there was only a significant difference between the subset1 ($M = 18.97, SD = 14.10$) and the subset2 condition ($M = 17.59, SD = 13.90$), $t(6857) = 2.81, p = .0495$. Unlike our previous study (Ji & Pourtois, 2018), the subset based objective difference scores did not differ from each other when there were 4, 8 or 16 faces presented (subset4: $M = 17.81, SD = 13.86$; subset8: $M = 18.11, SD = 13.91$; set16: $M = 18.84, SD = 14.02$), $-2.02 < t_s < -.57, p_s > .44$ (Figure 3A).

Subset based subjective difference scores

The effect of subset size was significant, $\chi^2(7) = 25.86, p < .001$ (Figure 3B). The subset based subjective difference scores in the subset1 condition ($M = 20.56, SD = 14.28$) were larger than those in the subset2 ($M = 18.52, SD = 13.38$), subset4 ($M = 19.09, SD = 13.36$), and subset8 conditions ($M = 19.08, SD = 13.20$), $t_s > 2.99, p_s < .028$, but did not differ significantly from those in the set16 condition ($M = 20.21, SD = 13.18$), $t(6857) = .69, p > .99$. Similar to the subset-based objective difference scores, the subjective difference scores did not differ between the subset4, subset8 or set16 conditions, $-2.4 < t_s < .10, p_s > .16$.

In addition, as seen in Figure 3, the errors computed relative to the visible faces (the subset-based difference scores) were on average lower than those computed relative to the full set (the whole-set based difference scores), which makes sense because participants could only see the subset of faces from which they extracted the mean emotion.

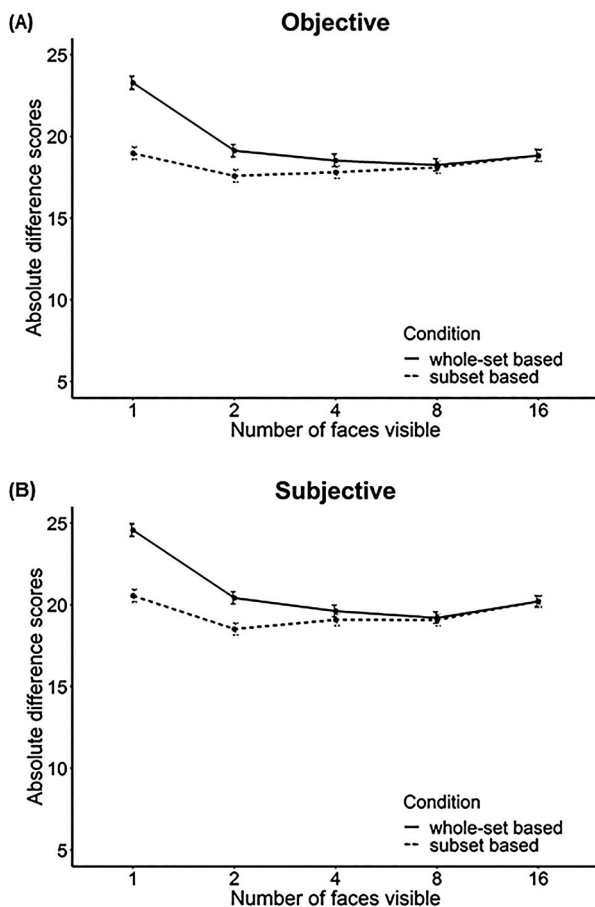


Figure 3. Results of Experiment 1. (A) Objective and (B) subjective difference scores (means) for each subset condition. The whole-set based difference scores (solid line) reflect the absolute difference between the average emotion judgments and the mean emotion of the whole set, regardless of the subset size. The subset-based difference scores (dashed line) were the absolute difference between the average emotion judgments and the mean emotion of the visible faces. Larger values indicate worse performance on the averaging task. Error bars represent one standard error of the mean (computed across trials and participants).

Experiment 2

In Experiment 1, we used faces morphed from angry to happy expressions, and the emotion variance of face sets was relatively large. In order to examine whether sampling efficiency was impacted by the emotion variance of faces in the set, we reduced the variance in Experiment 2 by using faces morphed from neutral to either happy or angry expressions. An additional analysis (see Supplementary results)

confirmed that the emotion variance was indeed smaller in Experiment 2 than Experiment 1.

Whole-set based objective difference scores

There were significant effects both for subset size and emotion, $\chi^2(7) = 759.58$, $\chi^2(8) = 16.93$, $ps < .001$. The interaction between subset size and emotion was also significant, $\chi^2(12) = 11.22$, $p = .02$, and adding this interaction effect to the model improved the goodness of fit. The difference scores were larger for averaging angry faces ($M = 17.69$, $SD = 7.76$) relative to happy faces ($M = 16.07$, $SD = 8.39$) in the subset1 condition, $\chi^2(1) = 23.22$, $p < .001$, but they did not differ significantly in the other subset conditions, $ps > .99$ (Figure 4A). Importantly, for both angry and happy faces, post-hoc analyses showed that the whole-set based objective difference scores in the subset1 condition were significantly larger than those in all the other subset conditions, $ts > 16.10$, $ps < .001$, and the objective difference scores did not differ from each other in the subset2 (angry: $M = 12.44$, $SD = 8.24$; happy: $M = 12.04$, $SD = 8.03$), subset4 (angry: $M = 11.99$, $SD = 7.62$; happy: $M = 11.69$, $SD = 7.68$), subset8 (angry: $M = 11.61$, $SD = 7.41$; happy: $M = 11.19$, $SD = 7.50$), or the set16 conditions (angry: $M = 11.70$, $SD = 7.43$; happy: $M = 11.28$, $SD = 7.68$), $-.30 < ts < 2.61$, $ps > .09$. The whole-set based difference scores bottomed-out when more than two faces were visible, which suggests that only around two faces were integrated to build the mean emotion representation.

Whole-set based subjective difference scores

Both effects of subset size and emotion were significant, $\chi^2(7) = 592.56$, $\chi^2(8) = 11.84$, $ps < .001$. The interaction between subset size and emotion was not significant, $\chi^2(12) = 7.71$, $p = .10$, and adding this interaction effect to the model did not improve the goodness of fit. Similar to the objective difference scores, the whole-set based subjective difference scores in the subset1 condition ($M = 17.54$, $SD = 9.20$) were significantly larger than those in all the other subset conditions, $ts > 18.52$, $ps < .001$. In addition, the subjective difference scores did not differ from each other in the subset2 ($M = 12.66$, $SD = 8.76$), subset4 ($M = 12.29$, $SD = 8.36$), subset8 ($M = 12.31$, $SD = 8.39$), or the set16 conditions ($M = 12.78$, $SD = 8.45$), $-1.84 < ts < 1.42$, $ps > .65$. On the other hand, the difference scores were larger for estimates of average emotion on angry

faces ($M = 13.77$, $SD = 8.62$) relative to happy faces ($M = 13.22$, $SD = 9.10$), $t(10812) = 3.44$, $p < .001$ (Figure 4B). The results confirm that around two faces were integrated and contributed to mean emotion perception.

Subset based objective difference scores

There were significant effects of subset size and emotion, $\chi^2(7) = 29.02$, $\chi^2(8) = 21.72$, $ps < .001$. The interaction between subset size and emotion was not significant, $\chi^2(12) = 7.18$, $ps = .13$. The subset based objective difference scores were smaller in the subset1 condition ($M = 10.66$, $SD = 8.40$) than in the set16 condition ($M = 11.49$, $SD = 7.56$), $t(10812) = -3.53$, $p = .004$. Difference scores were smaller in the subset2 condition ($M = 10.30$, $SD = 7.63$) than in the subset4 ($M = 11.01$, $SD = 7.51$), subset8 ($M = 11.00$, $SD = 7.50$) and set16 conditions, $ts < -2.99$, $ps < .028$. Similar to our previous study (Ji & Pourtois, 2018), the subset based objective difference scores did not differ from each other when there were 4, 8 or 16 faces presented, $-2.11 < ts < .11$, $ps > .35$ (Figure 4A). The other comparisons between different subset size conditions were not significant, $ps > .99$. The subset based objective difference scores were generally larger for averaging angry faces ($M = 11.25$, $SD = 7.92$) relative to happy faces ($M = 10.54$, $SD = 7.53$), $t(10812) = 4.66$, $p < .001$ (Figure 4A).

Subset based subjective difference scores

The effect of subset size was significant, $\chi^2(7) = 57.68$, $p < .001$ (Figure 4B). Neither the main effect of emotion nor the interaction between subset size and emotion reached significance, $\chi^2(8) = 0.54$, $p = .46$, $\chi^2(12) = 8.90$, $p = .06$. Thus, only the effect of subset size was kept in the model. Similar to the subset-based objective difference scores, the subjective difference scores in the subset1 condition ($M = 11.57$, $SD = 9.19$) were smaller than those in the set16 condition ($M = 12.78$, $SD = 8.45$), $t(10813) = -4.70$, $p < .001$. They were also smaller in the subset2 ($M = 10.89$, $SD = 8.18$) than in the subset4 ($M = 11.63$, $SD = 8.40$), subset8 ($M = 12.00$, $SD = 8.33$) and set16 conditions, $ts < -2.88$, $ps < .040$. The subset based subjective difference scores in the set16 condition were significantly larger than those in the subset4 and subset8 conditions, $ts > 3.09$, $ps < .02$, which is consistent with our previous findings (Ji & Pourtois, 2018), although the latter two

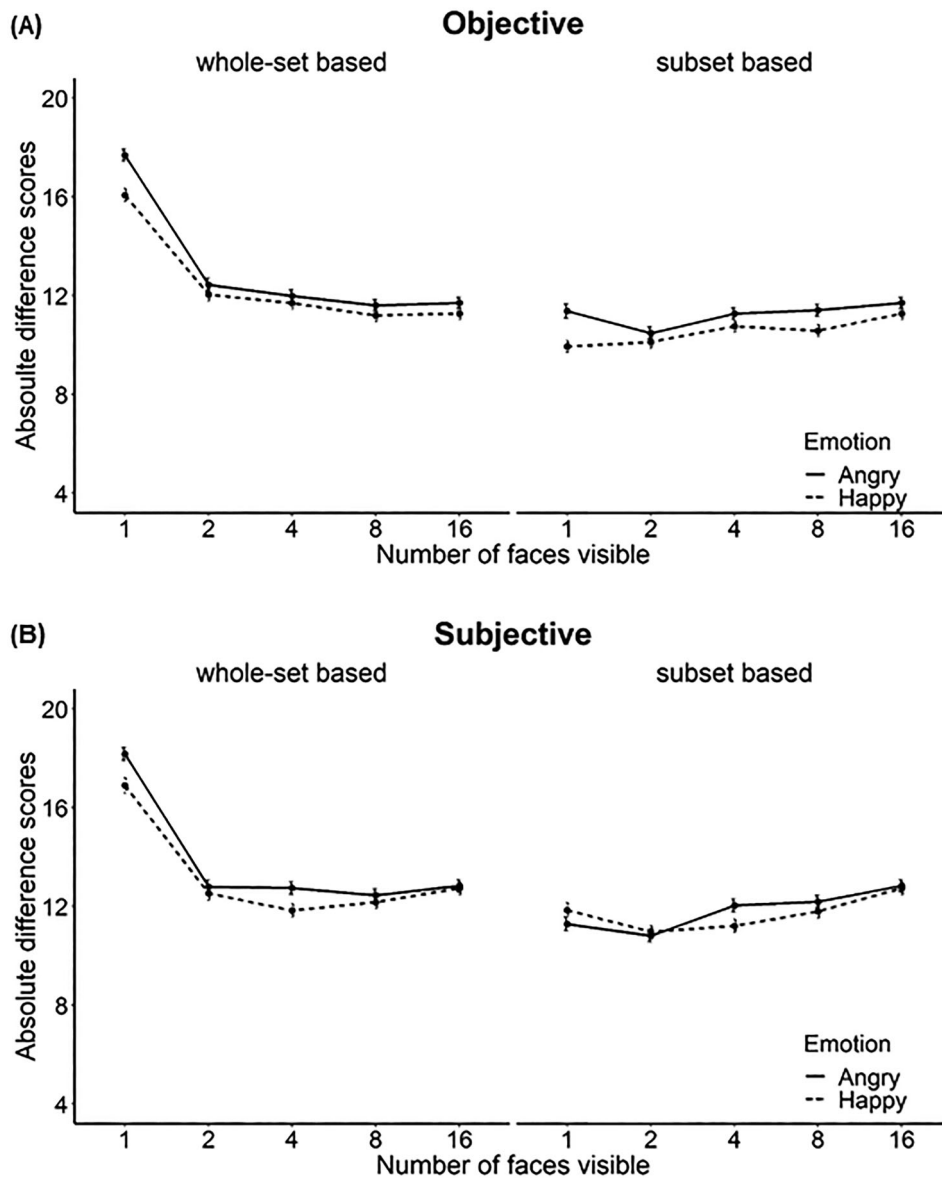


Figure 4. Results of Experiment 2. (A) Objective and (B) subjective difference scores (means) for each subset condition, shown separately for judgments of angry and happy faces. Larger values indicate worse performance on the averaging task. Error bars represent one standard error of the mean (computed across trials and participants).

conditions did not differ from each other, $t(10813) = -1.41$, $p > .99$.

Experiment 3

The emotion variance was smaller in Experiment 2 than Experiment 1, however, different face stimuli were used which made it difficult to directly compare results from these two experiments. In Experiment 3, we used faces morphed from neutral to either happy or angry expressions, as in Experiment 2, but the minimal distance between each emotion unit was smaller in this experiment

compared to Experiment 2, leading to a further reduction of variance here and more comparability between the two experiments, which helped to further examine the effect of emotion variance on sampling efficiency.

Whole-set based objective difference scores

There was a significant effect of subset size, $\chi^2(7) = 258.46$, $p < .001$. Emotion also contributed to average emotion judgments significantly, $\chi^2(8) = 11.91$, $p < .001$. The interaction between subset size and emotion was not significant, $\chi^2(12) = 4.70$, $p = .32$, and adding this interaction effect to the model did

not improve the goodness of fit. As can be seen in Figure 5, the whole-set based objective difference scores were largest in the subset1 condition ($M = 13.39$, $SD = 8.30$), $t_s > 8.97$, $p_s < .001$. The difference scores were also significantly larger in the subset2 ($M = 11.58$, $SD = 7.64$) compared to the subset4, subset8, and the set16 conditions (subset4: $M = 11.00$, $SD = 7.55$; subset8: $M = 10.55$, $SD = 7.43$; set16: $M = 10.68$, $SD = 7.48$), $t_s > 2.86$, $p_s < .043$; while the latter three conditions did not differ significantly from each other, $-.64 < t_s < 2.22$, $p_s > .27$. In addition, the difference scores were generally larger for averaging angry faces ($M = 11.66$, $SD = 7.99$) relative to averaging happy faces ($M = 11.20$, $SD = 7.50$), $t(13535) = 3.45$, $p < .001$ (Figure 5A). Based on the observed whole-set based objective difference scores which decreased when the subset size increased from two to four and then levelled off for larger subset sizes, it seems that around three to four faces were sampled during averaging in Experiment 3.

Whole-set based subjective difference scores

There were significant effects of subset size, $\chi^2(7) = 147.45$, and emotion, $\chi^2(8) = 93.27$, $p_s < .001$. The interaction between subset size and emotion also contributed to average emotion judgments significantly, $\chi^2(12) = 27.20$, $p < .001$, and adding the interaction effect to the model improved the goodness of fit. The difference scores for averaging angry faces were larger compared to those for averaging happy faces, $p_s < .028$, except in the set16 condition, $p = .54$. When only one face was visible (subset1; angry: $M = 14.39$, $SD = 8.56$; happy: $M = 12.11$, $SD = 9.44$), the whole-set based subjective difference scores were larger than all the other conditions for both angry and happy faces, $t_s > 4.88$, $p_s < .001$, except that those in the subset1 condition did not differ significantly from those in the set16 condition for the happy faces ($M = 11.53$, $SD = 8.88$), $t(13535) = 1.85$, $p = .65$. For the angry faces, the difference scores in the subset2 condition ($M = 12.41$, $SD = 8.33$) were also significantly larger than those in the subset4 and subset8 conditions (subset4: $M = 11.48$, $SD = 7.88$; subset8: $M = 11.48$, $SD = 7.89$), $t_s > 3.09$, $p_s < .020$, but the difference scores in the subset2, subset4, and subset8 conditions did not differ significantly from the set16 condition ($M = 12.01$, $SD = 8.00$), $p_s > .50$. On the other hand, for the happy

faces, the difference scores in the subset2 condition (subset2; $M = 10.34$, $SD = 8.16$) were not significantly different from those in the subset4 or the subset8 conditions (subset4; $M = 10.52$, $SD = 8.49$; subset8; $M = 10.60$, $SD = 8.03$), $p_s > .99$, and the difference scores in these subset conditions were even smaller than those in the set16 condition ($M = 11.53$, $SD = 8.88$), $t_s < -3.04$, $p_s < .024$ (Figure 5B).

The whole-set based subjective difference scores generated somewhat complicated results when emotion variance was relatively small. It seemed that three to four faces were likely sampled when averaging angry expressions, while around two faces were integrated when averaging happy expressions. Notably, the whole-set based subjective difference scores decreased and then increased to some extent (i.e., averaging performance improved and then deteriorated) when the number of visible happy faces increased from 1 to 16. This pattern was unexpected based on our original hypothesis and previous findings (Sweeny & Whitney, 2014; Yamanashi Leib et al., 2014) which predicted that the averaging performance should plateau at a certain point. One reason for these unexpected results might be that noise levels were actually set-size dependent rather than stable, as we originally assumed. When we allowed our estimates of noise to increase with increasing set sizes in our ideal-observer analyses (similar to Tokita, Ueda, & Ishiguchi, 2016; also see Sweeny, Grabowecky, Kim, & Suzuki, 2011; although note that noise values were arbitrarily chosen in these analyses, see Supplementary materials), we observed a U-shaped curve for whole-set based difference scores (Supplementary Figure 3).

Subset based objective difference scores

The effect of subset size and emotion both contributed to the model significantly, $\chi^2(7) = 73.17$, $\chi^2(8) = 20.45$, $p_s < .001$. The interaction between subset size and emotion was significant as well, $\chi^2(12) = 10.90$, $p = .028$. The comparisons between subsets were similar for the two emotion blocks, $p_s > .23$, except between the subset1 and set16 condition, where the former was significantly larger than the latter for angry faces (subset1: $M = 12.68$, $SD = 8.64$; set16: $M = 10.68$, $SD = 7.60$); this difference between these two subsets being smaller for happy faces (subset1: $M = 11.43$, $SD = 7.85$; set16: $M = 10.68$, $SD = 7.38$), $p = .03$. The subset-based objective difference scores were

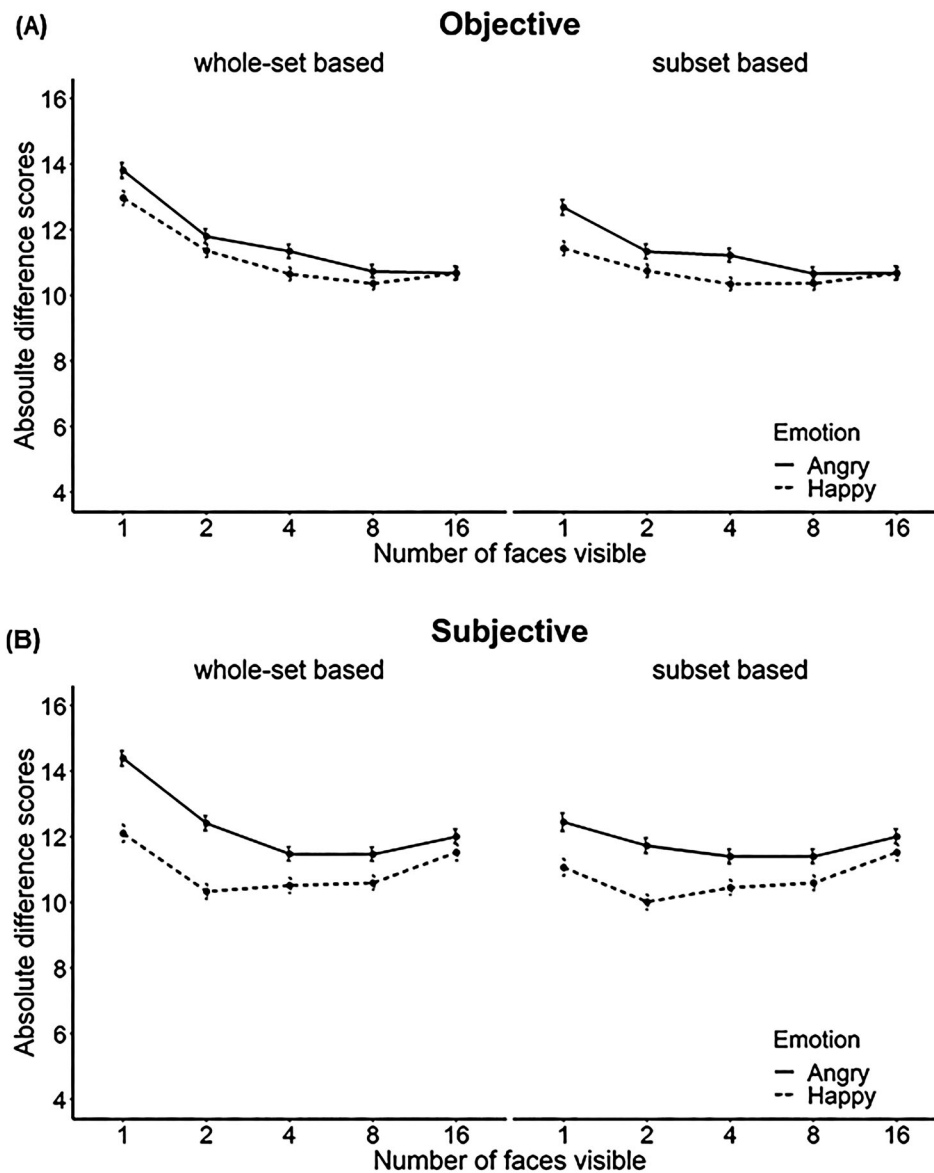


Figure 5. Results of Experiment 3. (A) Objective and (B) subjective difference scores (means) for each subset condition, shown separately for angry and happy faces. Larger values indicate worse performance on the averaging task. Error bars represent one standard error of the mean (computed across trials and participants).

the largest in the subset1 condition, $t_s > 2.95$, $p_s < .032$, except that the difference between the subset1 and subset2 condition did not reach significance for happy faces, $t(6763) = 2.95$, $p = .09$. Scores did not differ from each other when there were 2, 4, 8 or 16 faces presented in both angry and happy blocks, $-1.12 < t_s < 2.25$, $p_s > .24$ (Figure 5A), which was consistent with our previous study (Ji & Pourtois, 2018).

Subset based subjective difference scores

Both effects of subset size and emotion were significant, $\chi^2(7) = 35.98$, $\chi^2(8) = 57.17$, $p_s < .001$. The

interaction between subset size and emotion did not reach significance, $\chi^2(12) = 9.28$, $p = .055$. As shown in Figure 5B, the subset-based subjective difference scores were larger in the subset1 condition ($M = 11.76$, $SD = 9.65$) than the subset2 ($M = 10.88$, $SD = 8.45$), subset4 ($M = 10.93$, $SD = 8.19$) and subset8 ($M = 11.00$, $SD = 7.99$) conditions, $t_s > 3.44$, $p_s < .006$, but the latter three conditions did not differ significantly from each other, $-.72 < t_s < -.35$, $p_s > .99$. Meanwhile, difference scores were significantly larger (i.e., worse performance) when all 16 faces were made available ($M = 11.77$, $SD = 8.45$), compared to the conditions when 2, 4, or 8 faces were presented, $t_s > 3.55$, p_s

< .004. It was different from the null set-size effect found in our previous study where the spatial density of faces was matched across set sizes (Ji & Pourtois, 2018). On the other hand, similar to subset-based objective difference scores, the subjective difference scores were larger for average judgments of angry faces ($M = 11.80$, $SD = 8.54$) compared with happy faces ($M = 10.73$, $SD = 8.57$), $t(13535) = 7.57$, $p < .001$.

Comparison of experiments 2 and 3

We also directly compared the results of Experiments 2 and 3 where we used the same face stimuli but with different amounts of variance, in order to examine the potential impact of emotional variability on sampling efficiency. We constructed a null model with no fixed effects first, and then added fixed effects of subset size, emotion and experiment to the model sequentially. The interactions between each pair of factors (subset size and emotion, experiment and subset size, and experiment and emotion) and their three-way interactions were also introduced to the model sequentially, following the previous fixed effect. Each model was compared to the previous model by likelihood ratio tests to examine whether the added component contributed to averaging performance significantly. The models were fit for the whole-set based objective and subjective difference scores separately. Similar to Ji and Pourtois (2018), the range of the mean values (from 20 to 31) in Experiment 2 was smaller than in Experiment 3 (from 11 to 40), thus we selected the trials in Experiment 3 for which the mean units matched those in Experiment 2.

Both effects of subset size and emotion contributed to the whole-set based objective difference scores significantly, $\chi^2(7) = 896.85$, $\chi^2(8) = 18.51$, $ps < .001$. The interaction between subset size and emotion, and the interaction between subset size and experiment were significant as well, $\chi^2(13) = 13.55$, $\chi^2(17) = 47.86$, $ps < .009$. The effect of experiment did not contribute to the model significantly, $\chi^2(18) = .74$, $p = .039$, nor did the interaction between experiment and emotion or the three-way interaction between the three factors, $\chi^2(18) = 1.45$, $\chi^2(22) = 3.35$, $ps > .23$. Notably, the whole-set based objective difference scores only differed in the subset1 condition between Experiments 2 and 3, $\chi^2(1)$, $p < .001$. Moreover, the comparisons between each pair of subset-

size conditions did not differ significantly between Experiments 2 and 3, $ps > .99$, except the comparisons between subset1 and the other size conditions, $ps < .001$. The whole-set based objective difference scores were different between angry and happy faces only in the subset1 condition, $\chi^2(1)$, $p < .001$.

The comparisons of whole-set based subjective difference scores between two experiments were similar to the objective ones. There were significant effects of subset size and emotion, $\chi^2(7) = 638.93$, $\chi^2(8) = 40.97$, $ps < .001$, and significant interactions between subset size and emotion, $\chi^2(13) = 12.71$, $p = .013$, between subset size and experiment, $\chi^2(17) = 57.45$, $p < .001$, and between experiment and emotion, $\chi^2(18) = 8.90$, $p = .003$. Neither the effect of experiment nor the interaction between the three factors was significant, $\chi^2(9) = 2.75$, $\chi^2(22) = 7.98$, $ps > .097$. Similar to the objective scores, the whole-set based subjective difference scores were different between Experiments 2 and 3 in the subset1 condition only, $\chi^2(1) = 18.36$, $p < .001$, and the comparisons between the subset1 and any other size conditions were significantly different between two experiments, $ps < .001$, but not the other pairs of comparisons, $ps > .99$. The whole-set based subjective difference scores were larger for angry than happy faces in the subset1, subset2, and subset4 conditions, $\chi^2(1) = 58.39$, $\chi^2(1) = 26.47$, $\chi^2(1) = 14.74$, $ps < .001$, but not the other set-size conditions, $ps > .10$.

General discussion

The current study examined whether subsampling during averaging of multiple facial expressions depends on emotion variance in the set. Using the subset manipulation, different subsets of faces with variable intensities of expression were presented to participants across three experiments, who made estimates of the average emotion of these sets. Averaging performance for each subset size was compared to a condition in which all faces ($n = 16$) were available, in order to infer the number of faces eventually sampled by participants to perform this task. Our results suggest that extracting mean emotion from multiple faces can be explained by subsampling with a limited perceptual capacity. However, we did not find strong evidence that sampling efficiency was modulated by the inter-item variance in the sets, at least in this study where the emotion variance of

faces was roughly in the range of 8 and 25 (see Supplementary Figure 1).

Using the subset manipulation, we tested the hypothesis that during averaging, more than one face would eventually be sampled and used to form an ensemble representation. If only one item was randomly selected from the set, then the averaging performance, when compared to the mean of the whole set ($n = 16$), should remain constant regardless of how many items were made visible to the participants (Wolfe et al., 2015; Yamanashi et al., 2014; see also our simulation results in Supplementary Figure 2; and those from Sweeny & Whitney, 2014). In all three experiments, results showed that the whole-set based difference scores were significantly smaller (i.e., better performance) when two faces were presented relative to a single face, thereby ruling out the mere use of a one-face sampling strategy when extracting mean emotion from multiple faces shown concurrently. Integrating more than one face is consistent with the defining feature of ensemble coding recently put forward by Whitney and Yamanashi Leib (2018). Accordingly, our results lend support to the assumption that genuine ensemble representations of facial expressions were formed by the participants in our study.

In Experiments 1 and 2, where the emotion variance was high and medium, respectively, the whole-set based difference scores remained stable when the subset size increased from two to sixteen faces, suggesting therefore that probably only two faces were actually integrated with one another. Interestingly, when the emotion variance was further reduced in Experiment 3, the whole-set based objective difference scores dropped substantially from one to two faces and continued to decrease from the subset2 to subset4 condition, before they levelled off with further increasing subset sizes, which suggested that around three to four faces might have been sampled and averaged together. However, we are cautious about concluding that averaging efficiency seemed to improve when emotion variance decreased, since the whole-set based subjective difference scores showed a complex pattern of results and direct comparisons between Experiments 2 and 3 did not provide evidence that variance impacted sampling efficiency. Notwithstanding this caveat, we suggest that limited-capacity sampling (i.e., no more than four faces or four faces' worth of

information) could satisfactorily explain our new results. A similar interpretation was made previously for the averaging of size information by Myczek and Simons (2008).

By contrast, using a similar subset method, Wolfe and colleagues (2015) showed that at least twelve faces in a 24-face set were sampled. They found that averaging performance continued to improve when the number of visible faces increased from 1 to 12. Moreover, under gaze-contingent foveal occlusion where only peripheral vision could be used by participants, their results remained the same (Wolfe et al., 2015). Tentatively, this discrepancy between these and our new results might be explained by several methodological factors. For example, a 1500 ms stimulus presentation time was used in Wolfe et al. (2015), whereas the faces in the current study were presented for 500 ms only and immediately masked. It has been shown previously that increasing exposure time improved average emotion perception (Haberman & Whitney, 2009; Li et al., 2016). Although some studies have reported an invariance in the sampling efficiency with changing durations for some stimulus categories (see Sweeny et al., 2013 for biological motion), it remains unclear whether the same holds true when different facial expressions have to be averaged. Moreover, unlike Wolfe et al. (2015), we used face stimuli from different identities. Previously, Im et al. (2017) showed that implicit average emotion perception did not differ when identical or different facial identities were used. However, it remains to be determined whether different identities might impair performance when explicit averaging is required, considering that perception of facial expression and identity are inter-dependent (e.g., Calder & Young, 2005).

Furthermore, Wolfe et al. (2015) used a normal distribution for selecting the emotional intensities of set members, whereas we selected emotion units from a uniform distribution for the full set, always with four repetitions of each unique unit, resulting in a generally smaller amount of emotion variance compared to Wolfe et al. (2015). Indeed, it has been shown that when a set has no variance (i.e., all items are identical), only one item may be sampled (Allard & Cavanagh, 2012). Taking these previous studies together, it is reasonable that more items may have been sampled when the variance increased in our sets, but the range of emotion variance we employed was not sensitive enough to yield a systematic effect on the

averaging efficiency. The actual relationship between emotion variance, sampling and averaging is not entirely clear yet, and needs to be elucidated further. Several unanswered questions remain. (i) Below which variance level does the averaging process cease, and as a result, only a single face actually contributes to the percept? (ii) Symmetrically, above which level must emotion variance be held preferably to yield an averaging process that includes many more faces (i.e., more than 4)? (iii) Within a certain range of variance, is there a linear relationship between variance and averaging efficiency or might more complicated relationships be at work? Additional empirical work is needed to answer these questions.

At a theoretical level, our new results also have important implications for understanding the perceptual mechanisms underlying ensemble representations for multiple facial expressions. First, although sampling efficiency was not unambiguously impacted by variance in the current study, our results do suggest a potential interaction between mean and variance in averaging performance (e.g., interaction between subset size and experiment). These two kinds of summary statistics may be processed in parallel, and may involve different cognitive processes (Khvostov & Utochkin, 2019; Yang, Tokita, & Ishiguchi, 2018). Alternatively, these two important summary statistics could also interact with each other. In agreement with this view, in a previous study (Tong, Ji, Chen, & Fu, 2015), variance perception was found to be impacted by the stability of the mean, whereas in other investigations, perception of the mean was clearly modulated by variance (Ji & Pourtois, 2018; Marchant et al., 2013; Utochkin & Tiurina, 2014; but not in Ying & Xu, 2017, perhaps because implicit temporal ensemble coding was involved in their adaptation task). Second, in both Experiments 1 & 3, extraction of mean emotion was significantly worse (i.e., larger subset based difference scores, namely the error was larger relative to the mean of the faces that were actually presented) when there was only one face in the set (the subset 1 condition) compared to the other conditions where a variable number of multiple faces was presented. This outcome is consistent with what was previously referred to as the “power of averaging” in the existing literature (Alvarez, 2011). Perceivers appear to be able to compute an accurate representation of the scene that allows reducing and even cancelling out the noise carried by each

individual item. However, as our results show, the power of averaging is not unlimited. When many faces were presented (i.e., two or more), perception of the average emotion did not improve further. There were some exceptions, as in Experiment 2, for example, where averaging fewer faces in the set turned out to be better than sampling more faces. This result is unexpected and awaits replication. Moreover, in Experiment 1, subset-based difference scores did not differ when four, eight or sixteen faces had to be averaged. As such, this null subset-size effect for averaging angry and happy facial expressions is somewhat at odds with what we found in our previous study (see Experiment 1 in Ji & Pourtois, 2018, where a clear set-size effect was reported). At this point, we can only speculate on the possible causes of this discrepancy, but two factors are worth considering. First, the subsets of faces were randomly selected from the 16-face sets in the current study, and the sets containing smaller numbers of faces actually had smaller variance than the larger sets (see Supplementary Figure 1), which was different from Ji and Pourtois (2018) where four unique emotion units were always selected and the emotion variance remained constant across different set sizes. Smaller variance might have contributed to the smaller difference scores (i.e., better averaging performance) in the subset4 and subset8 conditions (see also Marchant et al., 2013; Utochkin & Tiurina, 2014). Second, here we used random locations of faces in the subset, and thus did not control spatial density across the different set sizes, something we explicitly did in our previous study (Ji & Pourtois, 2018). It has been suggested previously that mean orientation and mean size perception is generally robust and resistant to changes in density (Chong & Treisman, 2005; Dakin, 2001). However, it remains an open question for future research whether sparser displays with smaller numbers of faces could artificially impair emotion-averaging ability, which would in turn lead to worse performance in smaller subset conditions and thus could explain the lack of differences between these conditions and the set16 condition in the current study.

Notably, the lack of a set-size effect (apart from the one-face condition) in averaging performance in the current study does not necessarily confirm unlimited-capacity for, or parallel processing of, multiple faces. Based on the results of our ideal observer analyses (see Supplementary results, simulations

section), the subset-based difference scores appeared to be flat when the number of faces increased from four to sixteen, even though only two faces were randomly sampled and averaged, as our empirical results suggest. In other words, sampling a limited number of faces (e.g., less than three or four) could result in an unchanged averaging performance across the different set sizes. Accordingly, the conclusion of a parallel- or capacity-unlimited processing for multiple items based on findings revealing a null set-size effect might need to be reconsidered, and eventually amended somehow (Haberman & Whitney, 2007, 2009; Im et al., 2017; Marchant et al., 2013; Utochkin & Tiurina, 2014). In this context, Allik and colleagues previously proposed an intriguing “Noise and Selection” model for mean size perception (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013). It might be equally valid for averaging high-level multidimensional stimuli, such as emotional expressions. The measurement of each face necessarily contains noise, and only a limited number of faces are presumably integrated in the computation of mean emotion. Although we included different identities to increase the heterogeneity of face sets and considered the inter-individual differences in emotion perception by computing subjective difference scores, a few faces likely represented the whole set adequately to some extent already. When similar items are grouped together, any item selected in the set necessarily provides a representative estimate of the whole set, which might explain the limited-capacity subsampling on one hand, and constant (noisy) averaging performance with increasing set sizes on the other hand.

Using the subset manipulation, we assumed that perceptual experience of judging a subset of faces largely resembled the experience of subsampling and pooling these faces from an entire set. Admittedly, however, these two kinds of experience could never be identical. For example, in subset conditions (e.g., two faces), the sampling process has already been done for participants, whereas if two faces are to be sampled in the whole set condition, this process has to unfold organically. Factors like crowding (Whitney & Levi, 2011), for example, could also differentially impact performance in the whole set condition. Adding variable levels of internal noise to our ideal-observer models begins to account for these types of differences (see Supplementary Figure 3). Although

our ideal observer simulations only served an illustrative purpose and did not provide precise estimates of how many faces were actually sampled by participants, they do provide a valuable and complementary tool for gaining insight into the nature of subsampling used by participants to perform ensemble coding. The ideal observer simulations used in this study, when considered together with the behavioural results obtained using the subset method, provide a more comprehensive perspective on the sampling strategies participants may have used to compute and establish ensemble representation for multiple facial expressions. In future studies, it would be useful to manipulate variance within an experiment and apply the “external noise” technique (e.g., Allard & Cavanagh, 2012; Solomon, Morgan, & Chubb, 2011) or some other sophisticated simulations or models (e.g., de Gardelle & Summerfield, 2011; Li, Herce Castañón, Solomon, Vandormael, & Summerfield, 2017), to further explore the effect of variance on the computational efficiency of high-level ensemble coding.

Across three experiments, we ruled out the use of an extremely economical one-face sampling strategy and demonstrated ensemble coding for facial expressions (because at least two faces were integrated with one another), yet our results also point to an averaging process which lacked efficiency. As our new results suggest, it is likely that a limited number of (i.e., four or less than four) noise-perturbed faces, as opposed to a whole set including 16 faces, nor even a majority of a set, contributed to the averaging process in the present case. To conclude, our new findings corroborate the assumption that averaging multiple facial expressions could be achieved by limited-capacity subsampling. Whether or not its strength and efficiency are determined by factors like the variance of a set remains an open question to be examined in future studies.

Note

1. Due to the limitation of the refresh rate of the screen, the actual presentation of a 500 ms display was 493–494 ms. Similarly, for the 100 ms display, the actual presentation time was 93–94 ms.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by a China Scholarship Council (CSC) grant ([2014]3026) and a cofounding grant from Ghent University, both awarded to LJ.

Data availability

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/ufjmk/?view_only=b0cba61caed248b7-b01243ec2a4fc66a, DOI 10.17605/OSF.IO/UFJMK.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Allard, R., & Cavanagh, P. (2012). Different processing strategies underlie voluntary averaging in low and high noise. *Journal of Vision*, 12(11), 1–12.
- 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.
- Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. *Journal of Experimental Psychology: Human Perception and Performance*, 40(4), 1440–1449.
- Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8), 641–651.
- Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, 70(7), 1327–1334.
- Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900.
- Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, 20(5), 324–335.
- Corbett, J. E., & Melcher, D. (2014). Stable statistical representations facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 40(5), 1915–1925.
- 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.
- de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, 108(32), 13341–13346.
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), R751–R753.
- 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.
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics*, 72(7), 1825–1838.
- Im, H. Y., Albohn, D. N., Steiner, T. G., Cushing, C. A., Adams, R. B., & Kveraga, K. (2017). Differential hemispheric and visual stream contributions to ensemble coding of crowd emotion. *Nature Human Behaviour*, 1, 828–842.
- Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of Vision*, 18(3), 1–19.
- Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Vision Research*, 145, 39–48.
- Ji, L., Rossi, V., & Pourtois, G. (2018). Mean emotion from multiple facial expressions can be extracted with limited attention: Evidence from visual ERPs. *Neuropsychologia*, 111, 92–102.
- Khvostov, V., & Utochkin, I. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, 19(9), 1–18.
- Li, V., Hecce 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.
- Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, 7, 1332.
- 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.
- 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.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Core Team, R. (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131. Retrieved from <https://CRAN.R-project.org/package=nlme>
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3), 179–197.
- Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, 11(12), 1–11.
- Sweeny, T. D., Grabowecky, M., Kim, Y. J., & Suzuki, S. (2011). Internal curvature signal and noise in low- and high-level vision. *Journal of Neurophysiology*, 105(3), 1236–1257.
- 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.
- Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science*, 25(10), 1903–1913.
- Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a global sampling process in extraction of summary statistics of item sizes in a set. *Frontiers in Psychology*, 7, 711.
- Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, 15(4), 1–12.
- Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, 168(3), 242–249.
- 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(1), 7–18.
- Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15, 160–168.
- Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, 69, 105–129.
- Wolfe, B. A., Kosovicheva, A. A., Yamanashi Leib, A., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision*, 15(4), 11.
- Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, 14(8), 1–13.
- Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is there a Common summary statistical process for Representing the mean and variance? A study using Illustrations of Familiar items. *i-Perception*, 9(1), 204166951774729.
- Ying, H., & Xu, H. (2017). Adaptation reveals that facial expression averaging occurs during rapid serial presentation. *Journal of Vision*, 17, 1–19.