Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance

Luyan Ji & Gilles Pourtois

Department of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium

Word count: 9730

Correspondence to: Luyan Ji

Department of Experimental-Clinical and Health Psychology

Ghent University

Ghent 9000, Belgium

Email: Luyan.Ji@Ugent.be

Abstract

We examined the processing capacity and the role of emotion variance in ensemble representation for multiple facial expressions shown concurrently. A standard set size manipulation was used, whereby the sets consisted of 4, 8, or 16 morphed faces each uniquely varying along a happy-angry continuum (Experiment 1) or a neutral-happy/angry continuum (Experiments 2 & 3). Across the three experiments, we reduced the amount of emotion variance in the sets to explore the boundaries of this process. Participants judged the perceived average emotion from each set on a continuous scale. We computed and compared objective and subjective difference scores, using the morph units and post-experiment ratings, respectively. Results of the subjective scores were more consistent than the objective ones across the first two experiments where the variance was relatively large, and revealed each time that increasing set size led to a poorer averaging ability, suggesting capacity limitations in establishing ensemble representations for multiple facial expressions. However, when the emotion variance in the sets was reduced in Experiment 3, both subjective and objective scores remained unaffected by set size, suggesting that the emotion averaging process was unlimited in these conditions. Collectively, these results suggest that extracting mean emotion from a set composed of multiple faces depends on both structural (attentional) and stimulus-related effects.

Keywords: ensemble representation; facial expressions; processing capacity limitations; set size; amplifying effect; sampling

Introduction

For the last ten years, evidence has accumulated showing that human observers are able to rapidly process multiple emotional faces shown concurrently and extract the average emotion from them (e.g., Elias, Dyer, Sweeny, 2016; Haberman & Whitney, 2007, 2009; Ji, Rossi & Pourtois, in press). The representation which summarizes multiple features or items into an ensemble is referred to as ensemble representation (Alvarez, 2011; Whitney & Leib, 2018), and is thought to allow outlier detection in visual search (Cavanagh, 2001), as well as minimize the impression of being exposed to a visual world that would be too rich and complex to handle (Cohen, Dennett, & Kanwisher, 2016; Rensink, O'Regan, & Clark, 1997).

Like averaging low-level features or stimuli, for example orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001) and size (Ariely, 2001; Chong & Treisman, 2005), the ability of deriving the affective gist from multiple facial expressions has also been shown to be very robust and flexible across different tasks and contexts, occurring implicitly (Haberman & Whitney, 2007), and even on sets containing as many as 24 individual faces shown simultaneously for only 100 ms (Yang, Yoon, Chong, & Oh, 2013). In addition, even when the accuracy of individual representations is very low (e.g., at chance level) because of limited attentional resources, ensemble representation remains surprisingly precise (Fischer & Whitney, 2011; Haberman & Whitney, 2009, 2011; Li et al., 2016).

On the other hand, the underlying perceptual mechanism responsible for creating ensemble representation for higher-level information (such as facial expressions) is still

largely unclear and under debate in the existing literature. An open question remaining pertains to knowing whether ensemble representation could help overcome or bypass limitations in visual processing (Alvarez, 2011; Chong & Treisman, 2005; Cohen et al., 2016; but cf. Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Attarha, Moore, & Vecera 2014; Ji, Chen, Loeys, Pourtois, under review).

One way to assess attention bottlenecks in visual processing is using a classical set-size (i.e., the number of items in the set) manipulation (Theeuwes, 1992; Treisman & Gelade, 1980; Wolfe, 2007). This manipulation has been widely used in visual search studies in the past. For example, searching for a negative (angry) face surrounded by neutral faces used as distractor was found to be less impaired by increasing set sizes and thus more efficient, compared to a control condition where a positive face had to be searched in the set (Horstmann, 2007; Öhman, Lundqvist, Esteves, 2001). Based on capacity models for divided attention (Broadbent, 1958; Kahneman, 1973), if visual processing is capacity unlimited, then each stimulus is analyzed independently, so that the quality of perception does not vary (i.e., decrease) with increasing set sizes. In comparison, if perceptual capacity is limited[1], then there is by definition a limit imposed on the amount of information processed at a given time such that interference (competition) between stimuli occurs, particularly so when the number of stimuli in the set increases.

Different from the earlier psychophysical studies which focused on exploring set size effect in the processing and detection of a single target, studies on ensemble

[1] A limited-capacity parallel model is also possible (Palmer, 1990), but in the current study, we did not distinguish between parallel and serial accounts of limited-capacity processing as it goes beyond its scope.

representation consider all stimuli in the set as target elements that all participate in principle to shape visual processing and eventually determine emotion perception. Using set-size manipulations, Haberman and Whitney (2007, 2009) previously found that the averaging performance was not influenced by increasing set sizes, which provided support for a capacity unlimited process, and was consistent with findings on averaging low-level features or stimuli (e.g., mean size, Ariely, 2001; Chong & Treisman, 2005). Similar to these previous psychophysical studies on mean size representation (Ariely, 2001), Haberman and Whitney (2007, 2009) used a uniform distribution of emotional intensities composed of four unique morph units, and notably, they selected only one single identity. Therefore, the sets used in their study usually remained relatively homogeneous or regular with only four different expressions, no matter whether the set size was 4 or 16. However, these homogenous sets would presumably ease the sampling strategies. As a matter of fact, in these conditions, sampling only one or two items might explain behavioral performance and the resistance to set size manipulations, as demonstrated by simulation methods (Myczek & Simons, 2008).

In a recent study (Ji et al., under review), we found that the perceptual capacity of establishing mean representation for mixed full-blown angry and happy facial expressions was limited, using the extended simultaneous-sequential paradigm (Scharff, Palmer, & Moore, 2011). However, it might be challenging and also uncommon to average multiple facial expressions that convey distinct, and even opposite, emotion categories (i.e., happiness vs. anger), as the variance in the set is necessarily high in these conditions. Further, it has been shown previously with low-level attributes such as size or orientation that the averaging turned out to be easier and more accurate when the variance

in the set was reduced (e.g., Solomon, Morgan, & Chubb, 2011). To overcome this problem, in the current study, we used a standard morphing technique meant to reduce the variance of facial expressions presented within the set, as well as to better control their actual emotion intensity values. Moreover, we also manipulated this factor across different experiments to examine if it reliably influenced the averaging process. On the other hand, in order to reduce the regularity in the set and thus create a situation where a subsampling strategy would be inadequate to perform the averaging task, we decided to use different stimuli in the set invariably, namely having different emotional values each time, as was done previously in the case of mean size perception (Marchant, Simons, & de Fockert, 2013; Utochkin & Tiurina, 2014). However, for emotional facial expressions that have a more limited range than low-level properties, a caveat is that for larger set sizes, they are still rather homogeneous as the different stimuli composing the set are necessarily similar. As a compromise, in the current study, we employed a uniform distribution of four unique morph units, regardless of the varying set size (from 4 to 16), similarly to Haberman and Whitney (2007, 2009), but unlike them, we selected 16 different face identities, to increase heterogeneity in the set. In addition, unlike Haberman and Whitney (2007, 2009), we also collected from the same participants emotion ratings for all the individual (unmorphed) faces used in the main experiment in order to assess whether the objective (i.e., actual morph unit) or subjective (i.e., valence intensity rating) value best accounted for the averaging performance during the task (see Methods for details). This choice was motivated by the results of our previous study (Ji, et al., under review) where we found that the subjective emotion perception of faces was a reliable predictor of performance during the main averaging task since it took into account the

subject-specific perception of the emotional faces used as stimuli that can vary considerably across participants (unlike fixed morph units).

All in all, the current study therefore aimed at exploring the (attention) boundaries for extracting the mean emotion from a set composed of multiple facial expressions and how the emotion variance across them could modulate the processing capacity, using a standard set size manipulation and well controlled face stimuli (by means of a morphing procedure). To this aim, three different experiments were performed. Across them, participants judged the perceived average emotion from each face set on a continuous scale (similarly to Ji et al., under review). The face set consisted of 4, 8 or 16 faces, and was presented for 500 ms. In Experiment 1, we used morphed faces extracted from a continuum going from anger to happiness, hence providing a between-emotion categories manipulation. In Experiments 2 and 3, we used within-emotion continua (either from neutral to happy or from neutral to angry) in separate blocks, to decrease the inter-item (face) variance in the sets in terms of emotional expressions. Further, Experiment 3 differed from Experiment 2 in that the distance between the different morph units was smaller (thus the emotion intensity variance within the face set was smaller) in the former compared to the latter experiment. (i) We predicted that the averaging performance should mainly be capacity-limited (see Ji, et al., under review), in the sense of being influenced by the set size manipulation: a worse performance was expected for large compared to small set sizes. (ii) In addition, we hypothesized that the averaging performance would improve and be less affected by set size when the inter-item (face) variance (in terms of emotion expressions) decreased. Hence, we surmised modulatory

effects of set size and inter-stimulus variance on the ability to extract the mean emotion from a complex set composed of multiple facial expressions.

General Methods

Participants

All three experiments included twenty-four participants from Ghent University (Experiment 1: 18-25 years, 17 females; Experiment 2: 18-25 years, 15 females; Experiment 3: 19-28 years, 19 females). The sample size of 24 was determined a priori to be consistent with our previous behavioral study (see Ji et al., under review). The participants gave written informed consent prior to the start of the experiment and were compensated 10 Euro per hour. They reported to be right-handed and have 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

Sixteen different identities, eight males and eight females, were selected from the NimStim database (Tottenham et al., 2009). Each face identity showed happy, angry, or neutral expression, all with closed mouth. The hair, ears, neck and other external information were cropped. All images were converted to greyscale, and scaled to the same mean luminance and root-mean-square contrast (Bex & Makous, 2002). Each face image subtended a visual angle 4.03° × 4.28°, and was presented against a homogenous black background.

Face images were generated by morphing using FantaMorph 5. In Experiment 1, the morphing was carried out between the negative (Face 1) and the positive expression (Face 50) for each identity separately (Figure 1), which resulted in a total of 800 unique face stimuli (50 faces × 16 identities). The differences in emotion intensity between two adjacent images were denoted as one morph unit. In Experiments 2 and 3, images were morphed between the neutral (Face 1) and the apex of the corresponding expression (Face 50, either happy or angry) (Figure 1). This resulted in a total of 1584 unique face stimuli (1 neutral, 49 angry, 49 happy faces × 16 identities). The differences in emotion intensity between two adjacent images within each emotion category were denoted as one morph unit.

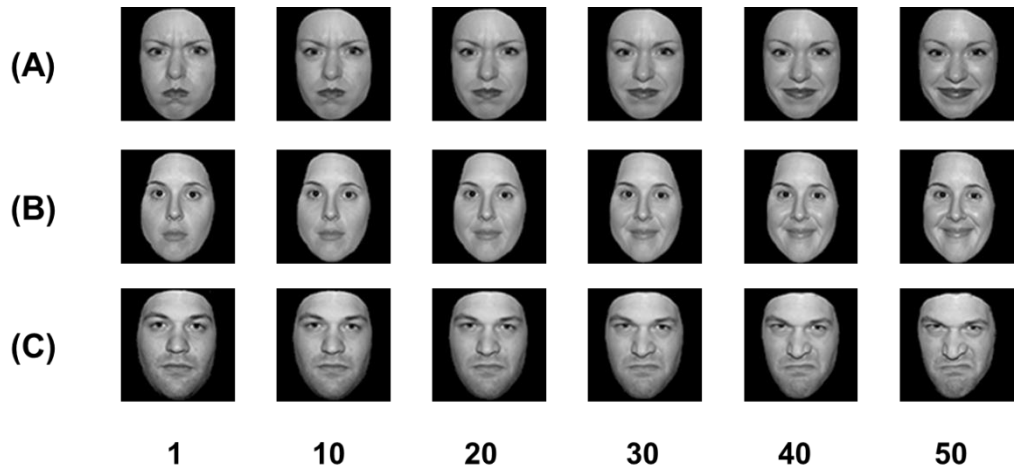


Figure 1. Examples of faces morphed from (A) angry to happy used in Experiment 1, and from (B) neutral to happy or (C) neutral to angry used in Experiments 2 & 3. For each continuum, 50 different images were generated for each face identity.

Each face set consisted of 4, 8, or 16 identities conveying different emotional intensities. The mean emotion of each set was randomly chosen before each trial, and

then four unique morph units surrounding the mean were selected. The smallest distance between each emotion unit was 6 (mean ± 3, ± 9; as used in previous studies, see Haberman & Whitney, 2007, 2009) in Experiment 1 where angry and happy faces were morphed. In Experiment 2 where neutral and emotional faces were morphed, the distance was increased to 12 (mean ± 6, ± 18), in order to match the emotion variance of face sets used in Experiment 1. In Experiment 3, the distance was 6 (mean ± 3, ± 9), but the perceived variance was smaller than in Experiment 1 (see Results). Each face set did not include the extreme emotional values (either Face 1 or Face 50). Thus, the mean was randomly selected from a uniform distribution of morph units ranging from 11 to 40 in Experiments 1 and 3, and from 20 to 31 in Experiment 2, respectively. The mean varied in each trial and was never a member of the face set. In the 8-face set, there were two instances of each morph unit, and in the 16-face set, there were four instances of each morph unit. Since the face identities were different in each face set, although some faces had the same morph unit, their emotion intensity could be perceived differently.

Like in Marchant et al. (2013), we controlled the density of the face set across set sizes. When there were 16 faces in the set, they were randomly located in an invisible 4 × 4 matrix (14.83° × 20.35°) centered on the screen, and their locations in each cell were also random. When there were 8 faces, they were placed in a 3 × 3 subset of the 4 × 4 matrix. It was equally likely that one of the nine cells was empty and three (44.4%) or four faces (55.6%) out of the eight were presented in the central 2 × 2 cells. For the 4-item set, the faces were placed in a 2 × 2 subset of the 4 × 4 grid. It was equally likely that these smaller subsets were present in any of the possible locations within the large 4

× 4 matrix. Therefore, there could be one (44.4%), two (44.4%), or four faces (11.1%) in the central 2 × 2 cells.[2]

Apparatus and procedure

Participants sat at around 60 cm in front of a 17″ CRT screen with a refresh rate of 85 Hz. Participants did the average emotion judgement task first. Speed of response was not emphasized and feedback was not given, but participants were encouraged to rely on their first impression and not to think extensively (similarly to Ji et al., in press, under review). Afterwards, they rated the emotion intensity and arousal of the individual faces. The two tasks were programmed and controlled using the E-Prime Version 2 software (Psychology Software Tools, Inc., 2001). Experiment 1 lasted about 30min, while Experiments 2 & 3 lasted double as long.

Average Emotion Judgment Task. A trial began with a fixation cross which appeared at the center of the screen for 500 ms. Then, a face set, made up of either 4, 8, or 16 faces, was presented for 500 ms, immediately followed by a scrambled face image used as mask and presented for 100 ms. The next trial started automatically 1000 ms-1200 ms after participants gave a response about "what is the average emotion intensity of all the faces", by means of a visual analogue scale (VAS) (Figure 2). The anchors of the scale were labeled *Extremely negative* and *Extremely positive* respectively, and the middle point indicated *Neutral*. The displays of the two labels (negative on the left or the

[2] It is known that acuity declines from fovea to periphery (e.g., Anstis 1974). However, for the 4-face sets, auxiliary results (not shown here) showed that the averaging performance was not worse when there were more faces presented in the periphery (than centrally), with one exception found in Experiment 2 when considering the subjective difference scores (with the opposite direction though). Furthermore, when we compared the 4-face set condition including one central face/three peripheral faces to the 8-face and 16-face conditions, the effect of set-size remained unchanged in all three experiments.

right) were counterbalanced across participants. In Experiments 2 and 3, participants were required to judge the average emotion from neutral to extremely positive (half of the scale) for happy faces, and from neutral to extremely negative for angry faces. Hence, emotion (i.e., valence) was manipulated using a block design in these two experiments.

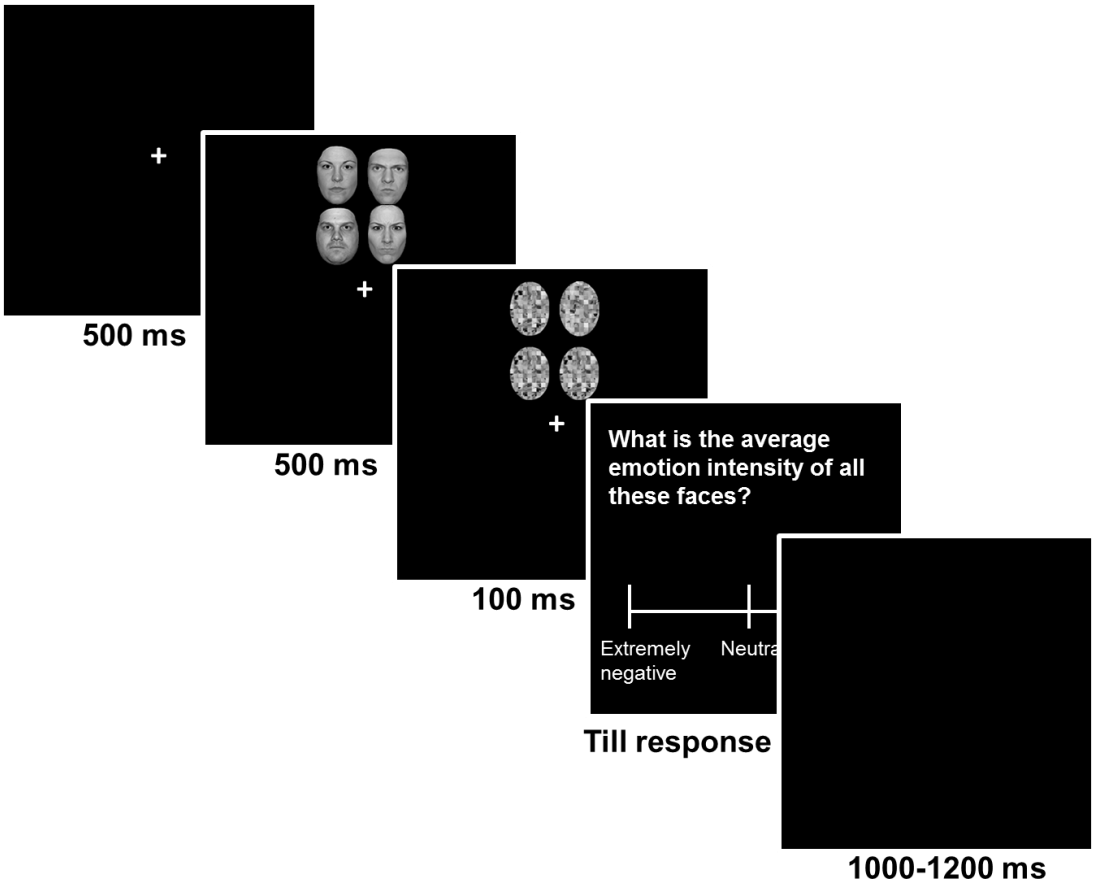


Figure 2. Average emotion judgment task in Experiments 1-3. Participants judged the perceived average emotion intensity from each face set on a visual analogue scale, ranging from extremely negative to extremely positive (these two anchors were counterbalanced across participants). In Experiments 2 & 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 4, 8, or 16 different faces.

The set size (4, 8, 16) and the mean emotion (morph unit from 11 to 40 in Experiments 1 and 3, and from 20 to 31 in Experiment 2) of each face set was 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 and 3, the emotion category (happy, angry) was blocked, and participants performed three experimental blocks of 72 trials (Experiment 2) or 90 trials (Experiment 3) for each emotion category. The happy and angry blocks were performed alternately, and which emotion was used in the first block was counterbalanced across participants. Participants practiced 30 (Experiments 1 and 3) or 24 (Experiment 2) trials to get acquainted with the average emotion judgments task. Practice trials were excluded from all subsequent analyses. Before the practice session, we showed participants several face sets as examples to help them understand and use the verbal labels used as anchors on the VAS. The label of extremely negative, extremely positive (or neutral in Experiments 2 & 3) was presented first, and then the sets containing 4, 8, and 16 original face images all showing the apex of the corresponding expressions (Face 1 or Face 50) appeared. Since all three experiments used the same original face images for morphing, similar references for the scales (not exactly the same though, because the identities in smaller sets were randomly chosen) were assumed to be available to all participants.

Face Emotion Rating Task. Following the main experiment and task (see here above), participants rated the emotion intensity and arousal of each individual face (Face 1 and Face 50 only). One face appeared at a time in the center and had the same size as that in the previous task. Participants used the mouse to click on two different VASes,

one for emotion intensity with two anchors labelled the same as those used in the average emotion judgment task (*Extremely negative* and *Extremely positive*), and another for arousal labeled *Extremely calm* and *Extremely excited*. The labels shown on the left and the right sides were counterbalanced across participants.

Data analysis

Data conversion. The actual positions participants clicked on the VAS in the average emotion judgement task were converted to data ranging from 0 to 100 in all three experiments. After conversion, the larger the value, the more positive the participants judged the average emotion from the face set; and the smaller this value, the more negative the average emotion from the face set was perceived. The morph units of each face stimuli (1-50) were also converted to match the range of the converted average emotion judgments. In Experiment 1, the morph units of every face were multiplied by 2. In Experiments 2 and 3, they were subtracted from 50 for angry faces and added with 50 for happy faces. After conversion, the larger the morph unit, the more positive the face stimuli was, and the smaller the morph unit, the more negative the face stimuli was. We extracted the *objective* absolute difference score by subtracting the average emotion judgment from the averaged morph units of all the faces in each face set.

We also computed the mean emotion of the faces in each set based on the subject-specific emotion intensity ratings obtained for these same faces (see Face Emotion Rating Task here above, as well as Ji et al., under review for a similar procedure). The emotion rating scores for each original face image (Face 1 and Face 50) were converted in the same way as the average emotion judgment data (results see Supplementary Materials).

The subjective emotion intensity of the corresponding morphing faces was extracted by linearly interpolating between that of Face 1 and Face 50. A *subjective* absolute difference score was then calculated as the absolute difference between the converted average emotion judgment and the computed mean emotion intensity.

Data trimming. For the average emotion judgment task, trials with RTs exceeding 2.5 SDs above or below the grand mean RT for each participant (overall 2.5%, 2.5% and 2.6% trials in Experiments 1-3, respectively) were excluded. This standard cutoff was chosen before running data analyses. Another 2.1%, 1.9% and 3.0% of trials with mouse clicks falling excessively far away from the scale (2.5 SDs above or below the mean position of the scale) were excluded in Experiment 1, 2 and 3, respectively. Since participants were required to judge on the scale ranging from neutral to extremely positive or from neutral to extremely negative for the happy and angry blocks respectively in Experiments 2 and 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 excluding 1.9% and 0.7% trials in these two experiments. One, two and one participants in Experiments 1-3 respectively had to be excluded because their subjective or/and objective absolute difference scores exceeded the 2.5SD of the grand mean of all participants in at least one set-size condition. The data of the remaining twenty-three, twenty-two and twenty-three participants were included in the statistical analyses.

Data analysis. To assess whether the average emotion judgments varied with the mean morph units assigned for each face set, we conducted multilevel analyses with random intercepts and random slopes of mean emotion units for each participant using

the lme function in the nlme package for R (Pinheiro, Bates, DebRoy, Sarkar, R Core Team, 2017). The null model with no fixed effects was first built, and then the fixed effects of mean emotion units and set size were added to the model sequentially. In Experiments 2 and 3, the fixed effect of Emotion was also introduced to the model, following the previous two fixed effects. The interaction between mean emotion units and set size was added at the final step. Each model was compared to the previous model by the likelihood ratio tests to examine whether the added component contributed to the average emotion judgments significantly. The coefficients of the final model with the best goodness of fit (smallest Akaike information criterion, Akaike, 1974) were reported (see results). To examine the effect of emotion in Experiments 2 & 3, average emotion judgments (and mean emotion units) were converted to arbitrary units ranging from 50 to 100: the larger was this value, the larger was the emotion intensity perceived by the participants in happy and angry sets.

Objective and subjective absolute difference scores were analyzed using repeated-measure ANOVAs. The common within-subjective factors across all three experiments were Set size (4, 8, 16). Experiments 2 and 3 had an additional within-subject factor, namely Emotion (angry, happy). Greenhouse-Geisser correction was applied when assumptions of sphericity were violated. A Bonferroni correction was used when multiple comparisons were performed. Except for the standard null hypothesis significance testing, we also conducted Bayes factor analyses (Bayesian repeated-measure ANOVAs and Bayesian paired sample *t* tests) for both objective and subjective differences scores using JASP (JASP Team, 2017) on the key main effect of set size and the planned follow-up comparisons. These additional Bayesian analyses helped to quantify the

strength of the evidence in favor of the null hypothesis (i.e., no reliable effect of set size) or alternatively, its rejection and confirmation of the alternative one (i.e., set size influenced performance) (Kass & Raftery, 1995).

Results

Experiment 1

Average Emotion Judgment. There was a significant effect of mean emotion units, χ^2 (1) = 82.10, $p < .001$. Set size or the interaction between mean emotion units and set size did not contribute to the average emotion judgments significantly, χ^2 (2) = 3.98, $p = .14$, χ^2 (2) = 2.62, $p = .27$, and adding these two fixed effects to the model did not improve the goodness of fit, thus they were not retained in the final model. Mean emotion units positively predicted observers' average emotion judgments, $b = 1.03$, $SE = .04$, t (5901) = 28.16, $p < .001$.When the face set contained happier expressions on average, the participants reliably judged more often the average emotion to be more positive (than negative) in this face set, which confirmed that participants' judgments were sensitive to the morph units of happy and angry faces embedded in the set (Figure 3).

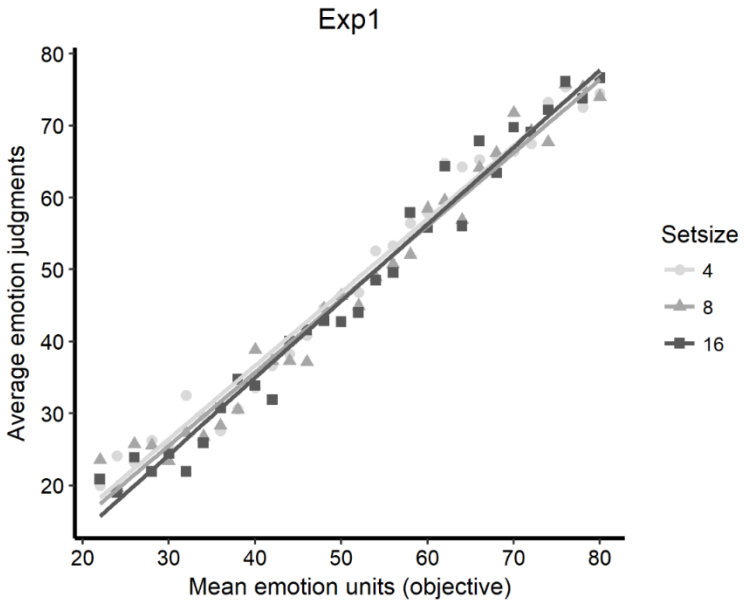


Figure 3. Average emotion judgments (means) of Experiment 1, shown separately for each mean emotion unit and the three set sizes, collapsed across participants. The larger the judgment, the more positive participants perceived the face set; the smaller the judgment, the more negative participants judged it. The regression lines in the graph were fitted for the aggregated average emotion judgments for each set size condition, for illustration purpose.

Objective Difference Scores. The ANOVA revealed a significant main effect of Set size, F (1.59, 35.97) = 10.57, $p < .001$, $\eta_p^2 = .32$, $BF_{10} = 132.30$ (strong evidence for H1). *Post hoc* tests showed that the objective difference scores in the set-size 16 condition ($M = 17.79$, $SD = 3.47$) were larger than both the set-size 4 ($M = 15.89$, $SD = 3.18$) and the set-size 8 conditions ($M = 16.63$, $SD = 3.46$), $p < .001$, $BF_{10} = 30.23$ (strong evidence), $p = .024$, $BF_{10} = 13.76$ (strong evidence); while the latter two did not differ significantly from one another, $p = .25$, $BF_{10} = 1.15$ (anecdotal evidence) (Figure 4).

Subjective Difference Scores. The main effect of Set size was significant, F (2, 44) = 16.68, $p < .001$, $\eta_p^2 = .43$, $BF_{10} = 3634.12$ (strong evidence for H1). Similar to the

objective difference scores, the subjective difference scores became larger with increasing set sizes (Figure 4). The subjective difference scores were larger in the set-size 16 condition (M = 19.75, SD = 3.55) than those in the set-size 8 condition (M = 18.41, SD = 3.49), p = .009, BF_{10} = 41.96 (strong evidence), and both of them were larger than those in the set-size 4 condition (M = 17.32, SD = 3.68), p < .001, BF_{10} = 270.76 (strong evidence), p = .038, BF_{10} = 4.98 (moderate evidence).

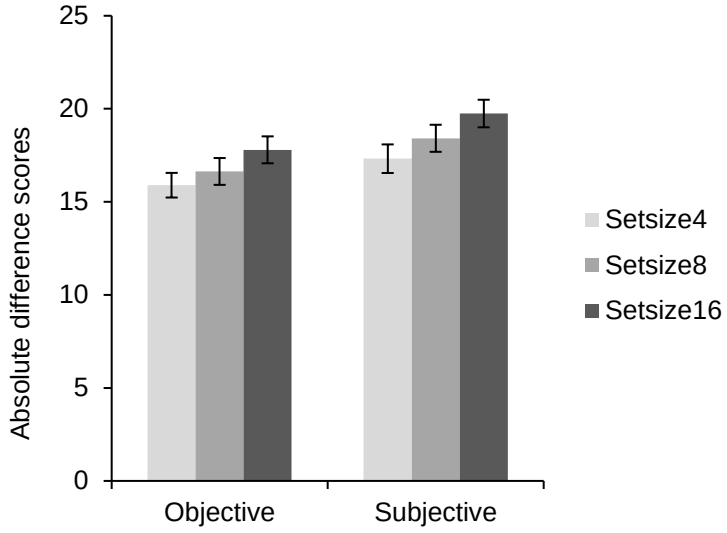


Figure 4. Objective and subjective absolute difference scores (means) of Experiment 1, shown separately for the three set sizes. The larger the value, the worse the averaging ability. The error bar represents one standard error of mean.

Experiment 2

Average Emotion Judgment. Mean emotion units significantly predicted observers' average emotion judgments, χ^2 (1) = 111.19, p < .001. There was also a significant effect of Set size, χ^2 (2) = 95.38, p < .001. Adding the effect of Emotion further improved the model, χ^2 (1) = 100.42, p < .001. The interaction between mean

emotion units and set size was not significant, χ^2 (2) = .55, p = .76, and hence they were not included in the final model. When the face set contained emotionally stronger expressions on average, the participants reliably judged more often the average emotion to be stronger (more positive or negative) compared with neutral in this face set, b = 0.68, SE =.04, t (9781) = 17.45, p < .001, which confirmed a positive relationship between average judgments and the morph units of happy or angry faces in the set (Figure 5). Interestingly, the average emotion judgments were overall larger when set size increased, revealing an amplification effect. The judgments in the set-size 16 condition were larger than the set-size 8 condition, t (9781) = 4.35, p < .001, and both of them were larger than the set-size 4 condition, t (9781) = 9.81, p < .001, t (9781) = 5.50, p < .001. Angry face sets were judged to be stronger than happy face sets, t (9781) = 10.04, p < .001.

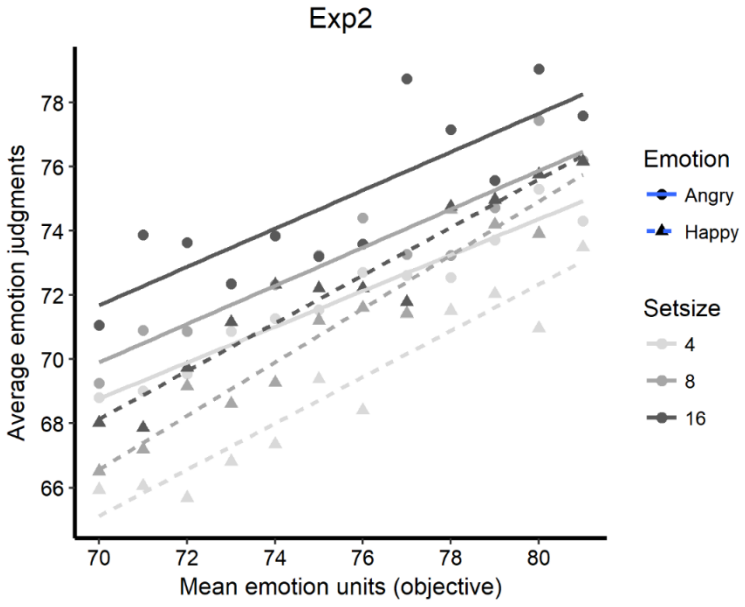


Figure 5. Average emotion judgments (means) of Experiment 2, shown separately for each mean emotion unit, the two emotions and three set sizes, collapsed across participants. The larger the judgment, the stronger emotion (either anger or happiness) participants perceived the face set; the smaller the

judgment, the weaker emotion participants judged. The regression lines in the graph were fitted for the aggregated average emotion judgments for each set size and emotion condition, for illustration purpose.

Objective Difference Scores. The ANOVA showed no significant main effect of Set size, F (1.40, 29.36) = 1.37, p = .27, η_p^2 = .061, BF_{10} = 0.21 (moderate evidence for H0). The main effect of Emotion, F (1, 21) <1, η_p^2 < .001, or the interaction between Set size and Emotion did not reach significance either, F (1.55, 32.47) = 2.67, p = .096, η_p^2 = .11 (Figure 6).

Subjective Difference Scores. There was no significant interaction between Set size and Emotion, F (2, 42) = 1.28, p = .29, η_p^2 = .06. The main effect of Set size was significant, F (1.56, 32.73) = 18.05, p < .001, η_p^2 = .46, BF_{10} = 1724.15 (strong evidence for H1). The subjective difference scores increased when there were more faces in the set (Figure 6). They were the largest in the set-size 16 condition (M = 13.17, SD = 2.44), which were larger than those in the set-size 8 condition (M = 12.05, SD = 1.93), p = .001, BF_{10} = 108.50 (strong evidence), and the set-size 4 condition (M = 11.22, SD = 2.30), p < .001, BF_{10} = 326.45 (strong evidence). The subjective difference scores were also larger when there were 8 faces compared with 4 faces in the set, p = .036, BF_{10} = 4.28 (moderate evidence). The main effect of Emotion was also significant, F (1, 21) = 4.84, p

= .039, η_p^2 = .19. The subjective difference scores for happy faces (*M* = 12.70, *SD* = 3.58) were larger than those for angry faces (*M* = 11.60, *SD* = 2.88).

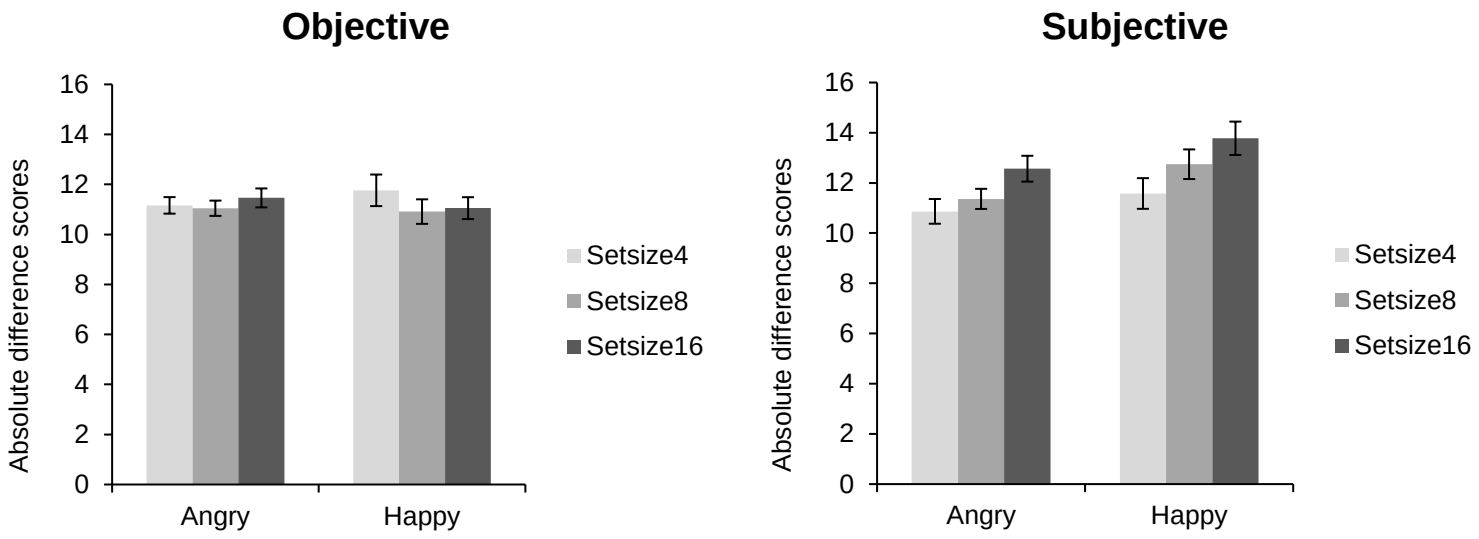


Figure 6. Objective and subjective absolute difference scores (means) of Experiment 2, shown separately for the two emotions and three set sizes. The larger the value, the worse the averaging ability. The error bar represents one standard error of mean.

Experiment 3

Average Emotion Judgment. Similar to Experiment 2, both mean emotion units and set size significantly predicted observers' average emotion judgments, χ^2 (1) = 59.80, $p < .001$, χ^2 (2) = 26.82, $p < .001$. Adding the effect of Emotion further improved the model, χ^2 (1) = 56.11, $p < .001$. The interaction between mean emotion units and set size did not reach significance, χ^2 (2) = 4.85, $p = .09$, and hence they were not included in the final model. Mean emotion units positively predicted observers' average emotion judgments, $b = 0.77$, $SE = .05$, t (11698) = 16.97, $p < .001$, confirming that participants' judgments were sensitive to the emotion intensity of faces (indicated by the morph units)

in the set (Figure 7). When there were 16 or 8 faces in the set, the judgments of setsize16 and setsize8 were both larger than those in the condition of 4 faces, t (11698) = 5.13, p < .001, t (11698) = 3.26, p = .003, while the former two did not differ significantly from each other, t (11698) = 1.88, p = .18. Angry face sets were judged to be stronger than happy face sets, t (11698) = 7.50, p < .001.

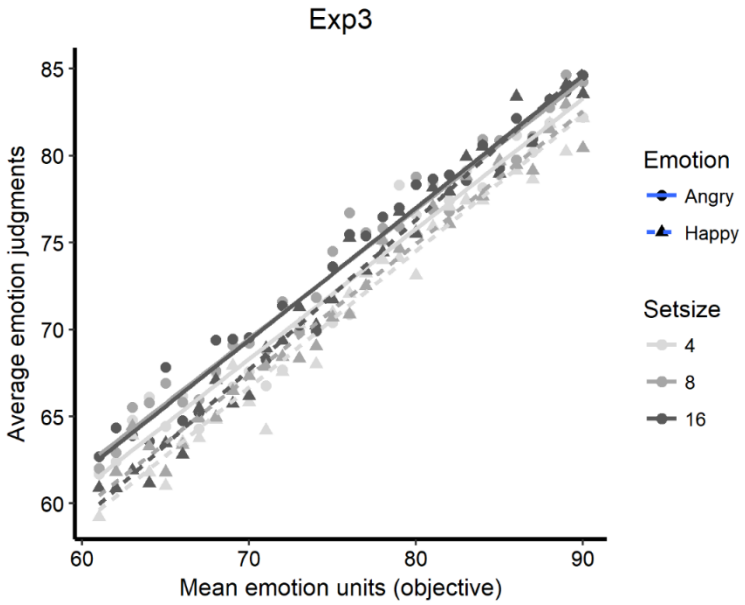


Figure 7. Average emotion judgments (means) of Experiment 3, shown separately for each mean emotion unit, the two emotions and three set sizes, collapsed across participants. The larger the judgment, the stronger emotion (either anger or happiness) participants perceived the face set; the smaller the judgment, the weaker emotion participants judged. The regression lines in the graph were fitted for the aggregated average emotion judgments for each set size and emotion condition, for illustration purpose.

Objective Difference Scores. The main effect of Set size did not reach significance, F (2, 44) = 2.69, p = .079, η_p^2 = .11, BF_{10} = 0.15 (moderate evidence for H0). The main effect of Emotion was not significant either, F (1, 22) = 1.28, p = .27, η_p^2

= .06, nor did the interaction between Set size and Emotion, F (2, 44) = 2.49, p = .095, η_p^2 = .10 (Figure 8).

Subjective Difference Scores. There was no significant main effect of Set size, F (1.54, 33.80) = 2.85, p = .084, η_p^2 = .12, BF_{10} = 0.12 (moderate evidence for H0). The main effect of Emotion or the interaction between Set size and Emotion was not significant either, F (1, 22) = 1.24, p = .28, η_p^2 = .05; F (2, 44) <1, η_p^2 = .01 (Figure 8).

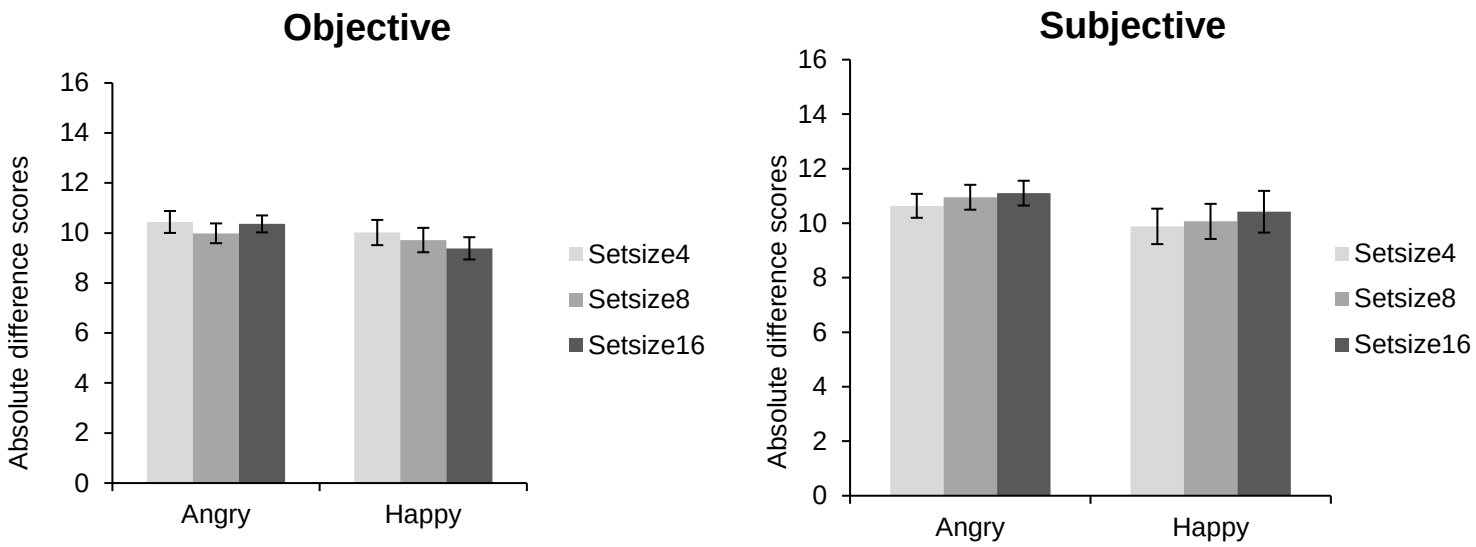


Figure 8. Objective and subjective absolute difference scores (means) of Experiment 3, shown separately for the two emotions and three set sizes. The larger the value, the worse the averaging ability. The error bar represents one standard error of mean.

Comparison of Experiment 2 and 3

In order to examine more directly the possible modulatory effect of variance on averaging performance, we compared the results of Experiments 2 and 3, with

Experiment as between-subject variable, and Set size and Emotion as within-subject variables. Since the range of the selected mean morph units (from 20 to 31) was smaller in Experiment 2 than in Experiment 3 (from 11 to 40), we selected the trials in this last experiment for which the mean values matched the ones used in Experiment 2. To be noted, the variance of emotion intensities in the face sets in the sub-selected trials was still significantly smaller in Experiment 3 (M = 8.27, SD = 1.31) than in Experiment 2, F (1, 43) = 44.58, $p < .001$, η_p^2 = .51. For the objective difference scores, the main effect of Experiment did not reach significance, F (1, 43) = 2.42, p = .13, η_p^2 = .05, nor did the main effect of Set size, F (1.60, 68.69) = 3.19, p = .058, η_p^2 = .07. Other main effects or interactions were not significant either, $ps > .12$. For the subjective difference scores, the main effect of Experiment did not reach significance, F (1, 43) = 3.85, p = .056, η_p^2 = .08, but there was a significant main effect of Set size, F (1.55, 66.44) = 16.57, $p < .001$, η_p^2 = .28, and an interaction between these two factors, F (2, 86) = 4.45, p = .014, η_p^2 = .09. A simple effect analysis revealed that the subjective difference scores did not differ significantly between Experiments 2 and 3 when set size was 4 (Exp2: M = 11.22, SD = 2.30; Exp3: M = 10.60, SD = 2.23) or 8 (Exp2: M = 12.05, SD = 1.93; Exp3: M = 11.00, SD = 2.09), F (1, 43) <1, p = .36, F (1, 43) = 3.04, p = .088, whereas they were significantly larger in Experiment 2 (M = 13.17, SD = 2.44) than Experiment 3 (M = 11.23, SD = 2.38) when the set size was 16, F (1, 43) = 7.28, p = .01. Moreover, the interaction between Experiment and Emotion was also significant, F (1, 43) = 5.46, p = .024, η_p^2 = .11. The subjective difference scores for the happy faces were larger in Experiment 2 (M = 12.70, SD = 2.68) compared to Experiment 3 (M = 10.44, SD = 3.37),

F (1, 43) = 6.15, p = .017; while they did not differ between experiments (Exp2: M = 11.60, SD = 2.00; Exp3: M = 11.45, SD = 1.86) for the angry faces, F (1, 43) <1, p = .79.

General Discussion

Across three experiments, we used a standard set size manipulation to test the processing capacity for extracting mean emotion from multiple facial expressions shown concurrently. Because inter-item variance changed between the three experiments (it was relatively large in Experiments 1 & 2, and smaller in Experiment 3), we could also examine the effect of emotion variance on the averaging performance. The results showed that increasing the number of faces in the set led to a clear impairment of the averaging performance (especially when considering the subjective, as opposed to objective difference scores), no matter the faces in the set showed between- (Experiment 1) or within-categorical emotions (Experiment 2). Hence, we found evidence in favor of capacity limitations to extract the mean emotion from a set composed of multiple facial expressions. Additionally, emotion variance also influenced the averaging performance. When the emotion variance was decreased (Experiment 3), increasing set sizes was no longer accompanied by a significant cost at the behavioral level, suggesting thereby that the averaging process could even become capacity unlimited under some circumstances, pending the actual variability (in terms of emotion intensities) across the different items forming the set was considerably reduced.

Previously, we already found using the simultaneous-sequential paradigm that averaging mixed full-blown happy and angry faces was in essence capacity limited (Ji et al., under review). The current study based on the set size manipulation, therefore

complemented and extended these earlier results in several directions. First, we showed that when morphed angry and happy faces were used to decrease the inter-item variability, the processing capacity was still limited (Experiment 1). Additionally, when the emotion variance of faces was matched, averaging emotional faces within the same category (i.e. angry or happy expressions with different identities) was capacity limited as well (Experiment 2). By comparison, using similar set-size manipulations, previous studies showed that the averaging performance did not vary with set size (Haberman & Whitney, 2007, 2009; Im, Albohn, Steiner, Cushing, Adams Jr., & Kveraga 2017), which seemed to be compatible with an unlimited-capacity process. The discrepancy between our and these previous results might be explained by the fact that the sets we used here had relatively larger variance than the sets used in these previous studies. Further, we used a continuous scale as response format in the present case whereas binary responses were collected in these earlier studies. This factor too might account for some of the differences found in the averaging ability across existing studies since the use of a VAS probably involves additional processes compared to a simple two-alternative forced choice task (see also Ji et al., under review for a discussion of this issue). Regarding the former issue, Haberman and Whitney (2007, 2009) used only a single face identity in their sets. In comparison, here, we invariably used different face identities in the sets to improve ecological validity and reduce this artificial redundancy or regularity in them. Although we included replications of emotional morph units for the larger set sizes (8 and 16), the faces were more heterogeneous along this specific dimension in the current study compared to Haberman and Whitney (2007, 2009). For low-level features or stimuli, it has been shown previously that the variance or heterogeneity reliably impacted the

precision of ensemble representations (Solomon, et al., 2011), and the averaging performance dropped significantly with increasing set sizes when the variance or range of items was increased (Marchant et al., 2013; Utochkin & Tiurina, 2014). Our new results are consistent with these earlier findings, extending them to the case of averaging multiple emotional expressions. When the variance of emotion intensities in the set was minimized by decreasing the distances of the selected morph units in Experiment 3 compared to Experiment 2, the set size effect disappeared (for subjective difference scores), an effect explained by a better performance for large set sizes in this experiment.

An additional interesting finding resulting from the current study relates to the unexpected amplifying effect found with increasing set size in Experiments 2 & 3, but not in Experiment 1. More specifically, in Experiments 2 & 3, when there were more faces in the set, participants were inclined to judge the average emotion with stronger intensity (e.g., much happier or angrier depending on the condition) on the VAS scale relative to the condition with a smaller set size, even though the actual mean intensity (either computed based on the subjective ratings or the objective morph units) was actually kept constant across the different set-size conditions or even slightly weaker with increasing set sizes (see Supplementary Materials). Noteworthy, previous studies on ensemble representation usually used binary responses, and computed either accuracy or a discrimination threshold (emotion: Haberman & Whitney, 2007, 2009; size: Ariely, 2001; Chong & Treisman, 2005). Alternatively, continuous adjustment responses were collected and the error between the estimated and the actual mean information was calculated (emotion: Haberman & Whitney, 2011; Elias et al., 2016; size: Marchant et al., 2013). To the best of our knowledge, these earlier studies did not analyze or report the

raw average judgment data extracted from a continuous scale however, like we did in the current study. Noteworthy, these continuous data probably provide richer and more complex information about the underlying averaging process than the use of binary responses. When the sets contained only one emotion category (either angry or happy) composed of faces having different intensities (Experiments 2 & 3), participants might use some biased sampling or weighting, where some relatively stronger expressions were selected or had larger weights in the averaging, and at the same time some expressions with relatively weak intensities were down-weighted or even ignored. When there were more faces in the set (larger set size), the number of emotionally stronger faces was also larger and these faces were more easily to be selected or attended, leading perhaps to the observed amplification effect. However, this does not entail that participants did not average multiple faces and only detected the emotionally strongest face, because the range of emotions intensities were identical across set sizes, and selecting only the strongest face could logically not result in such an amplification effect. From an evolutionary perspective, it may even be beneficial to have this sort of biased subsampling or weighting at the perceptual level and the resulting amplifying responses, since it seems important to rapidly discriminate which faces in the crowd are potentially friendly and which ones are threatening or foes, using the ones with clear expression/intensity each time (allowing in turn to establish a weighted average of them) (Cacioppo & Bernston, 1994). Interestingly, a similar bias has been demonstrated in numerical averaging where larger magnitude of numbers were selectively over-weighted compared to smaller numbers, and this effect was assumed to be beneficial somehow to deal efficiently with inherent capacity limitations during the averaging process (Spitzer,

Waschke, & Summerfield, 2017). Alternatively, the observed amplifying effect could result from the fact that the intensity of some faces in the set was perceived "exaggeratedly" because of the use of a rapid and peripheral presentation in the present study. A similar effect was previously reported for low-level visual stimuli and accounted for by a shift of neural representations to extreme channels during population coding (e.g, Mareschal, Morgan, & Solomon, 2008). At any rate, it remains currently unclear what factor(s) eventually caused this amplifying effect in the present case and therefore, additional work is needed to elucidate it.

At the methodological level, our study also adds to previous work on this topic by computing and comparing systematically across three experiments so-called objective to subjective differences scores, and eventually showing some valuable differences between them. Here, we did not only compute and use the objective morph units against which the actual averaging performance was calculated (as in Haberman & Whitney, 2007, 2009), but we also collected for all participants subjective ratings (in terms of intensity) for all the original face stimuli used in our experiments, and could therefore compute subjective differences scores that took into account the subject-specific perception of these faces (as opposed to arbitrarily set morph units) during the averaging process. In Experiments 1 & 3, the objective and the subjective difference scores showed similar results. Interestingly, in Experiment 2, we only found a clear effect of set size using the subjective difference scores, but not the objective difference scores. Moreover, we found that the averaging performance was influenced by the emotion variance when using the subjective difference scores only, as opposed to the objective difference scores, suggesting indirectly that the averaging process actually depended on the subject-specific perception

of these individual emotional expressions. Further, this dissociation confirms that these two dependent variables likely capture different effects or influences in the measurements made, and that for emotional facial expression recognition (and averaging), taking into account the subject-specific perception (in terms of emotional intensity) allows to reveal clearer effects of set-size (attention) and variance. Elias and colleagues (2016) previously collected subjective ratings for the emotional face stimuli they used, but obtained from independent raters. At variance with our results, these authors did not find any reliable difference for the averaging performance when comparing morph units to these subjective ratings. Accordingly, some caution is needed in the interpretation of the difference (or lack thereof) between the objective and the subjective difference scores, and additional empirical work is needed to elucidate under which conditions they converge (as in Experiment 1) or can show dissociable effects (as in Experiment 2). As can be seen in the Supplementary Materials section, it is striking that some of the face sets having the exact same mean morph units were eventually not perceived as conveying the same mean emotional intensity by different participants, emphasizing the individual differences of the muscular-action-intensity expressed in the emotional face stimuli on one hand (Ekman, 1993), and the idiosyncratic perception of emotional intensities on the other hand (Ekman, 1987). More generally, this apparent difference between objective and subjective differences scores might be important to consider in future studies that try to assess the commonalties or differences (and sometimes even independence) between averaging high-level (such as emotional facial expressions) and low-level properties (Haberman, Brady, & Alvarez, 2015). In the case of emotional facial expressions, like some of our results suggest (see Experiment 2), depending on which difference score is

used, a different outcome can be found. We therefore want to raise herewith awareness for this important methodological issue, especially when the focus is on averaging high-level objects or features, such as emotional facial expressions for which it is well known that large inter-individual differences do exist.

It is worth noting that because we sought to control and match density across the three set sizes in the current study, as a result, there were necessarily more faces shown in the periphery when larger sets were used. Accordingly, it might be argued that acuity (Anstis, 1974), as opposed to set size per se, eventually contributed to a drop in averaging performance with increasing set sizes in the present case (see Experiments 1 & 2). However, several arguments allow us to rule out this alternative account. First, in Experiment 3, where emotion variance was reduced, the exact same display was used as in Experiments 1 & 2, but no significant impairment of performance was found with increasing set sizes. Hence, we found that the averaging performance systematically varied across the experiments while acuity presumably did not. Additionally, auxiliary data analyses confirmed that the averaging performance was not worse when there were more faces presented in the periphery than centrally in the 4-face sets (see footnote 2). Although acuity alone is unlikely to explain the present results, in the case of averaging emotional facial expressions, effects of density and spatial configuration/extent on set size manipulations have not been explored systematically yet. Interestingly, such an attempt was made previously by Dakin (2001) with a focus on mean orientation processing. Accordingly, it might be valuable in future studies to adopt a similar methodology and eventually assess at the behavioral level if density and spatial extent

can influence the averaging performance for facial expressions when different set sizes are considered and compared with one another.

Last, we have to acknowledge that as is often the case with research on ensemble representation, it remains challenging in the present case to disentangle the contribution of averaging per se from the use of a sampling strategy to the observed behavioral results. In this context, limited capacity sampling strategies (Marchant et al., 2013) or "biased" weighting based on eccentricity (Ji, Chen, & Fu, 2014) might very well account for the observed drop in the averaging performance with increasing set size as well as when the variance in the set was large. When the number of faces in the set and/or the variability of facial expressions in the set increased, it became less likely that the specific stimuli which were sampled or gained additional weight (at the cost of other ones that were perhaps not sampled or processed) resembled the entire set. On the other hand, sampling a limited number of faces might not easily and fully explain the lack of set size effect, when the set was irregular or heterogeneous (Chong, Joo, Emmmanouil, & Treisman, 2008; Utochkin & Tiurina, 2014), as we found in Experiment 3. Nonetheless, even if a subsampling strategy could be assumed, it remains currently unclear whether built-in capacity limitations enforced it, or conversely, this subsampling strategy yielded capacity limitations to extract the mean emotion from a scene composed of multiple facials expressions and shown briefly. Hence, additional empirical work, including simulations, is probably required to clarify the complex link between capacity limitations and subsampling.

To conclude, the results of this study suggest that processing capacities to extract the average emotion from multiple facial expressions shown concurrently is limited.

Further, when the variance in the set (in terms of emotion valence and intensities) is reduced, this process appears to be less limited as increasing set size no longer negatively influenced the averaging performance. Moreover, clearer effects of set size and emotion variances were found on the averaging ability when using the subject-specific emotion perception (based on subjective ratings) compared to fixed or arbitrary morph units. As such, these results confirm that, despite some limitations in processing capacity, human observers can nevertheless perceive and extract with precision the mean emotion from a complex scene composed of multiple facial expressions and shown briefly (especially when they exhibit a limited variability in their emotional intensity), an extraordinary perceptual ability that is probably essential to guide interactions in complex social environments.

Acknowledgement

This work is supported by a China Scholarship Council (CSC) grant ([2014]3026) and a cofounding grant (BOFCHN2016000901) from Ghent University, both awarded to LJ.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

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.

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Research, 14*(7), 589–592.

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

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.

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America A*, *19*(6), 1096-1106.

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

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401.

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

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

Chong, S. C., & Treisman, A. (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.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, 18*(5), 1016–26.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... & Scherer, K. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*(4), 712.

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science, 28*(2), 193-203.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389-1398.

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.

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.

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

Horstmann, G. (2007). Preattentive face processing: What do visual search experiments with schematic faces tell us? *Visual Cognition*, *4532*(49), 1–53.

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.

JASP Team (2017). JASP (Version 0.8.2)[Computer software].

Ji, L., Chen, W., & Fu, X. (2014). Different roles of foveal and extrafoveal vision in ensemble representation for facial expressions. *In International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 164-173). Springer, Cham.

Ji, L., Chen, W., Loeys, T., Pourtois, G. (2018). *Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process.* Manuscript under review.

Ji, L., Rossi, V., Pourtois, G (in press). Mean emotion from multiple facial expressions can be extracted with limited attention: Evidence from visual ERPs. *Neuropsychologia.*

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

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.

Mareschal, I., Morgan, M. J., & Solomon, J. A. (2008). Contextual effects on decision templates for parafoveal orientation identification. *Vision research*, *48*(27), 2689-2695.

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.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*(3), 381.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 332–350.

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.

Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131, https://CRAN.R-project.org/package=nlme.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368-373.

Scharff, A., Palmer, J., & Moore, C. M. (2011). Extending the simultaneous-sequential paradigm to measure perceptual capacity for features and words. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 813.

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

Spitzer, B., Waschke, L., & Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, *1*(8), s41562-017.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Attention, Perception, & Psychophysics, 51*(6), 599-606.

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.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136.

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., & Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 12.1-12.25.

Wolfe, J. M. (2007). Guided search 4.0. *Integrated Models of Cognitive Systems*, 99-119.

Yang, J. W., Yoon, K. L., Chong, S. C., & Oh, K. J. (2013). Accurate but pathological: Social anxiety and ensemble coding of emotion. *Cognitive Therapy and Research*, *37*(3), 572-578.

1

Supplementary Materials

Emotion ratings for original face images

The emotion rating scores for each original face image (Face 1 and Face 50) were converted in the same way as the average emotion judgment data. The neutral faces in both Experiments 2 and 3 were perceived as negative, since the comparison (based on a one-sample *t* test) showed that their ratings were significantly lower than 50 (the smaller the value, the more negative the faces were perceived; see Supplementary Table 1B), $t(15) = -5.93$, $p < .001$, $t(15) = -5.25$, $p < .001$. In addition, neutral faces were rated as less aroused than both happy and angry faces in Experiment 2, $t(15) = -14.28$, $t(15) = -13.64$, *ps* < .001, and in Experiment 3, $t(15) = -13.19$, $t(15) = -16.48$, *ps* < .001 (Table Supplementary 1A & 1B).

To directly compare the emotion intensity of happy and angry faces in Experiments 1 to 3, we subtracted the converted emotion ratings from 100 for angry faces. Thus, the larger the value, the larger the emotion intensity was perceived in the faces for both angry and happy faces (see Table 1A). Paired *t* tests showed that the angry faces were perceived stronger than happy faces in Experiments 1 & 2, $t(15) = 2.42$, $p = .029$, $t(15) = 2.1310$, $p = .053$, but not in Experiment 3, $t(15) = .73$, $p = .48$. The angry faces were judged to be more aroused than happy faces in Experiment 1, $t(15) = 3.24$, $p = .005$, but they were rated as equally aroused as happy faces in Experiments 2 & 3, $t(15) = 1.00$, $p = .33$, $t(15) = -.08$, $p = .94$.

Supplementary Table 1A

Summary of Emotion Ratings for Angry and Happy faces in Experiments 1-3

		Intensity	Arousal
Experiment 1	Angry	84.54(9.22)	62.69(12.67)
	Happy	78.18(6.55)	51.52(12.50)
Experiment 2	Angry	82.55(8.96)	54.67(7.72)
	Happy	77.67(5.56)	52.63(6.32)
Experiment 3	Angry	84.94(8.64)	57.67(8.23)
	Happy	83.24(3.81)	57.85(6.08)

Note. The valence (intensity) and arousal rating (means and standard deviations) for the angry and happy faces in Experiments 1-3. The perceived intensity of angry faces was stronger than that of happy faces in Experiments 1& 2, but not in Experiment 3. The angry faces were judged to be more aroused than happy faces in Experiment 1 only.

Supplementary Table 1B

Summary of Emotion Ratings for Neutral faces in Experiments 2-3

		Valence	Arousal
Experiment 2	Neutral	46.71(2.22)	22.73(5.91)
Experiment 3	Neutral	47.93(1.58)	26.97(3.60)

Note. The valence and arousal rating (means and standard deviations) for the neutral faces in Experiments 2-3. The neutral faces were perceived as negative in both experiments, since their ratings were significantly lower than 50 (the smaller the value, the more negative the faces were perceived). In addition, neutral faces were rated as less aroused than both happy and angry faces in both experiments.

1

Emotion variance of face sets

We compared the variance (i.e. standard deviation) of emotion intensity in face sets between the three experiments based on the subjective emotion rating scores, and conducted a repeated-measure ANOVA with Experiment as between-subject variable, and with Set size as within-subject variable. The factor Emotion was collapsed in Experiments 2 and 3. There was no significant main effect of Set size, F (1.40, 90.69) = 1.26, p = .29, η_p^2 = .02, nor an interaction between Experiment and Set size, F (4, 130) <1, η_p^2 = .01. The main effect of Experiment was significant, F (2, 65) = 36.01, p < .001, η_p^2 = .53. *Post hoc* tests revealed that the variance was the smallest in Experiment 3 (M = 8.41, SD = 1.37), compared with that in Experiments 1 & 2, *ps* < .001. On the other hand, the variance in Experiment 1 (M = 12.47, SD = 1.60) and Experiment 2 (M = 11.83, SD = 2.18) did not differ significantly, p = .67.

Subjective mean emotion intensity of face sets with different set sizes

We examined whether the computed subjective mean emotion intensity in each set varied systematically with different set sizes in all three experiments. After conversion, the larger the value, the stronger the computed mean emotion intensity was, while conversely, the smaller this value, the weaker the computed mean intensity was. A repeated-measure ANOVA was run on these converted values for each experiment separately, with Set size as a common within-subject factor across all three experiments, and Emotion as an additional within-subject factor in the last two experiments only. We found a significant main effect of Set size in Experiment 1, F (2, 44) = 5.753.92, p =.006, η_p^2 = .21. The *post hoc* analysis showed that the subjective mean emotion was

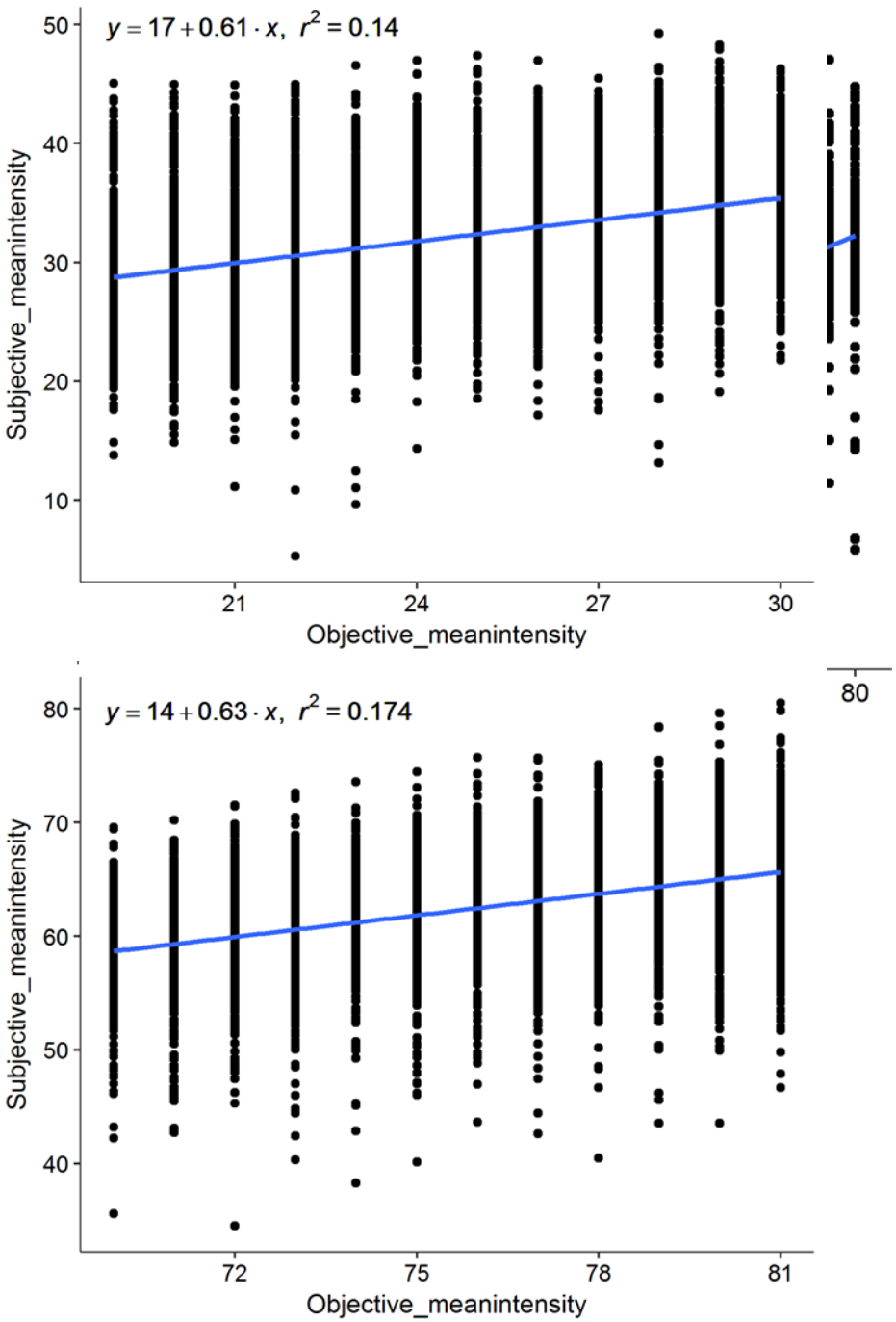
significantly weaker when there were 8 faces (M = 46.48, SD = 3.30) compared with 4 faces (M = 46.76, SD = 3.17), p =.024. The set-size 16 condition (M = 46.66, SD = 3.34) did not differ from the two other conditions, ps >.11. In Experiment 2, neither the main effect of Set size nor the interaction between Set size and Emotion was significant, F (2, 42) <1, η_p^2 = .01, F (2, 42) <1, η_p^2 = .002, indicating that the subjective mean emotion intensity of either the happy or angry faces did not vary across the three set sizes. In Experiment 3, the main effect of Set size was not significant either, F (2, 44) <1, η_p^2 = .02, however and interestingly, there was a significant interaction between Set size and Emotion, F (1.33, 29.34) = 3.92, p =.046, η_p^2 = .15. A simple effect analysis showed that the subjective mean emotion did not change with Set size for happy faces, F (2, 44) <1, η_p^2 = .028. However, for angry faces, it was slightly weaker but not significantly so when there were 16 faces (M = 68.54, SD = 3.70) compared with 4 faces (M = 68.66, SD = 3.72) in the set, p =.096; while it did not differ significantly between the set size 8 (M = 68.56, SD = 3.72) and the other two set size conditions, ps >.28.

Relationship between the subjective and objective mean intensity

We collapsed the three set sizes and used a simple linear regression model to assess if the objective mean intensity could predict the subjective one, for each experiment separately. After conversion, for both subjective and objective mean emotion, the larger the value, the more positive the face set was, and the smaller this value, the more negative the face set was. As can be seen from the Supplementary Figures 1-3, the mean emotion intensity computed based on individual intensity ratings for each face had actually large variance such that the face sets with the same mean morph units were not

necessarily perceived the same by participants. However, simple linear regression analyses showed that the objective morph units significantly predicted the subjective mean emotion; Experiment 1: F (1, 5923) = 20560, p <.001, adjusted R^2 = .78; Experiment 2: F (1, 4900) = 1035, p < .001, adjsuted R^2 = .17 (Happy), F (1, 4905) = 797.7, p < .001, adjsuted R^2 = .14 (Angry); Experiment 3: F (1, 5875) = 9784, p <.001, adjusted R^2 = .63 (Happy), F (1, 5864) = 10250, p <.001, adjusted R^2 = .64 (Angry).

(B)

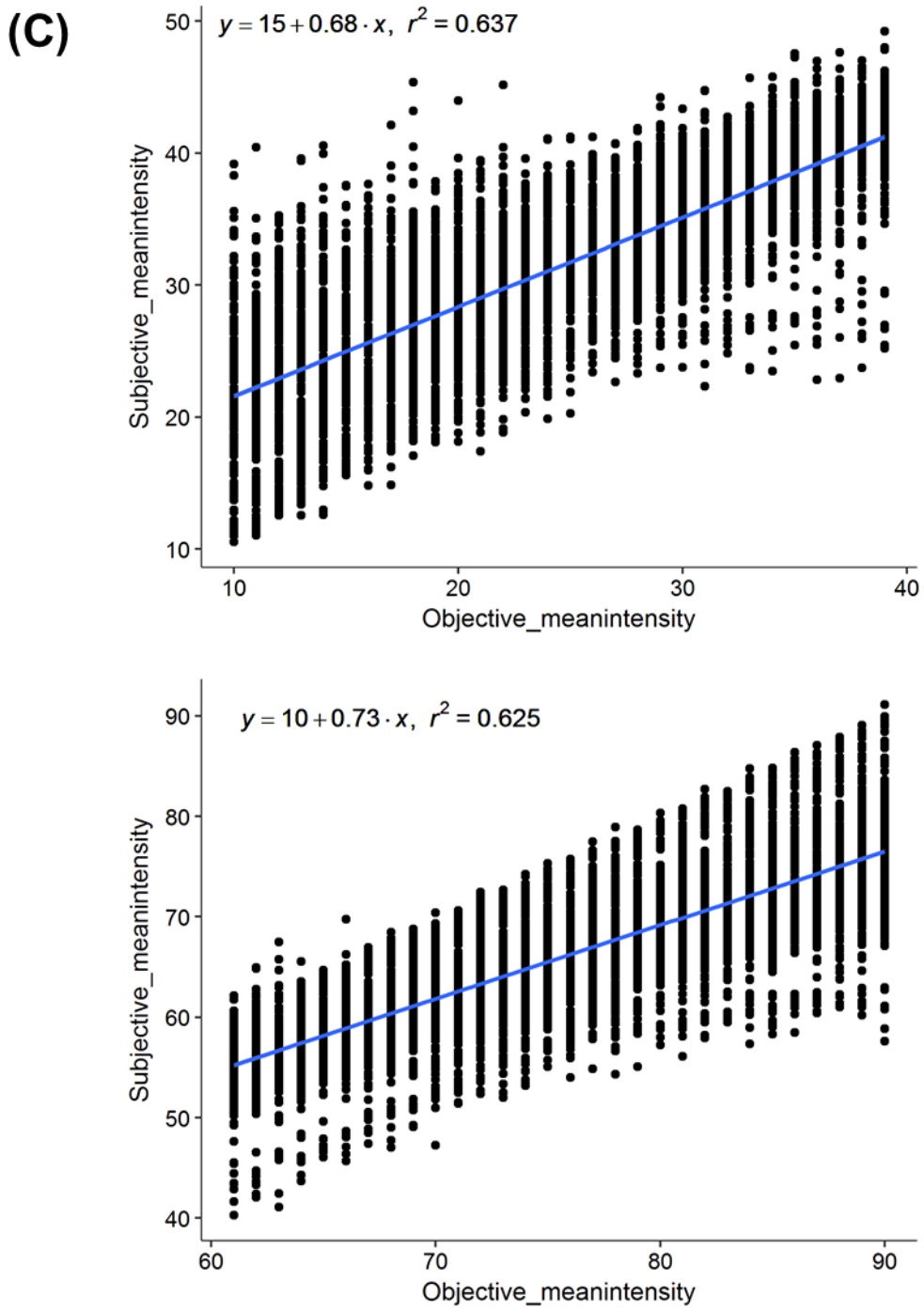

y = 17 + 0.61 · x, r² = 0.14
Subjective_meanintensity
Objective_meanintensity
50
40
30
20
10
21
24
27
30
y = 14 + 0.63 · x, r² = 0.174
Subjective_meanintensity
Objective_meanintensity
80
70
60
50
40
72
75
78
81

1

2

3

4



Supplementary Figure 1. The computed subjective mean intensity collapsed across the three set sizes shown against the computed objective mean intensity in Experiment 1 (A), Experiment 2 (B) and Experiment 3 (C). For Experiments 2 & 3, results of the angry (upper) and happy (lower) faces are shown separately. The larger the value, the more positive the computed mean emotion intensity was, while

conversely, the smaller this value, the more negative the computed mean intensity was. The regression model is provided for each experiment separately.