The Crowd Emotion Amplification Effect

Amit Goldenberg[1*] & Erika Weisz[2*]

Timothy D. Sweeny[3]

Mina Cikara[2]

James J. Gross[4]

[1]Harvard University, Harvard Business School

[2]Harvard University, Department of Psychology

[3]University of Denver, Department of Psychology

[4]Stanford University, Department of Psychology

[*]Equal contribution

Corresponding Author: Correspondence concerning this article should be addressed to Amit Goldenberg. E-mail: agoldenberg@hbs.edu.

Acknowledgements: The authors wish to thank Maria Reyes, Naila Ebeid, Vincent Rice and Sara Zaher for their assistance in running these studies, and Stephanie McMains for her assistance in developing the eye tracking protocol.

How do people go about reading a room or taking the temperature of a crowd? When people catch a brief glimpse of an array of faces, they can only focus their attention on some of the faces. We propose that perceivers preferentially attend to faces exhibiting strong emotions, and that this generates a *crowd emotion amplification effect*—estimating a crowd's average emotional response as more extreme than it is. Study 1 (N = 50) documents the crowd amplification effect. Study 2 (N = 50) replicates the effect even when we increase exposure time. Study 3 (N = 50) uses eye-tracking to show that attentional bias to emotional faces drives amplification. These findings have important implications for many domains in which individuals have to make snap judgments regarding a crowd's emotionality, from public speaking to controlling crowds.

Imagine yourself pitching an idea to a group of people. As you speak, you quickly scan the audience, your attention jumping from face to face. Are people smiling? Or do they look confused, bored, or even angry? People are often tasked with making split-second judgments about the emotions of a crowd, a skill that is critically important for navigating social interactions (Sanchez-Burks & Huy, 2009; Whitney, Haberman, & Sweeny, 2014). But how do people draw such rapid inferences from complex social landscapes?

One candidate mechanism is *ensemble coding*, the process by which perceivers quickly extract representative summaries of visual information (Alvarez, 2011; Whitney & Yamanashi Leib, 2018). Some evidence suggests that when perceivers compute the ensemble properties of a set of objects, they can include information from all objects in that set through distributed attention (Baek & Chong, 2020; Sun & Chong, 2020). Ensemble representations may instead be arrived at by sampling a subset of items and using them to extrapolate summary statistics about the entire set (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Maule & Franklin, 2016; Sweeny, Haroz, & Whitney, 2013). In this latter case, perceivers appear to preferentially sample the most salient objects in a set, which, in turn, biases the summary percept toward extreme or amplified values (Kanaya, Hayashi, & Whitney, 2018).

This selective summarization yields interesting predictions for the perception of crowd emotions. Because highly emotional faces are more salient than neutral faces (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002), ensemble coding of crowd emotions could involve preferentially attending to highly emotional expressions and using them to draw biased inferences about the entire crowd. If perceivers manifest an attentional bias toward more emotional faces, an estimation of a crowd's overall emotion could therefore be greater than it actually is. We evaluated these predictions in the present research.

Perceiving Emotional Faces

Facial expressions provide vital clues about people's goals and intentions, information that is crucial for successful social interactions (Fridlund, 1991; van Kleef, 2009). But not all expressions are equally captivating. Evidence suggests that people preferentially attend to faces conveying strong emotions over faces with neutral expressions (Eimer & Holmes, 2007; Pessoa et al., 2002). Thus, when inferring a crowd's emotion, attentional bias towards more emotional faces should contribute to an amplification in the average emotion estimate.

Just as emotional intensity can affect attention to faces, so too can the valence of the expression (positive or negative). If amplification occurs in the estimation of crowds' emotions, one question is whether such amplification is larger for negative or positive emotions. Some evidence suggests that people preferentially attend to negative stimuli as compared to positive stimuli (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Cacioppo, Berntson, & Gardner, 1997; Öhman, Lundqvist, & Esteves, 2001; Soroka, Fournier, & Nir, 2019). Preferential attention to negative stimuli should lead to greater amplification for a set of negative compared to positive faces. However, the literature is not united in this conclusion and some studies that have examined attention to single faces suggest that positive emotional faces are more attention grabbing, which should mean that amplification would be greater for positive emotions (Becker, Anderson, Mortensen, Neufeld, & Neel, 2011; Kosonogov & Titova, 2018).

Although these possibilities have not been directly tested, one ensemble coding experiment that examined accuracy in classification of either happy or angry crowds (compared to neutral) found that people were more accurate in classifying positive emotions, suggesting larger amplification in negative emotions (Bucher & Voss, 2018). This study, however, focused on classification without assessing whether error was systematically biased towards

amplification. Based on the evidence available to date, we hypothesized that crowd amplification would be stronger for negative emotions, but anticipated that this effect may be small.

Ensemble Coding

When speaking in front of a crowd or entering a room, people generate rapid evaluations of others' emotions (Whitney et al., 2014). Although the capacity for visual representation is finite in these situations (Alvarez, 2011; Whitney & Yamanashi Leib, 2018), ensemble coding compensates for this limitation by allowing perceivers to form compressed, summary representations of visual information (Whitney et al., 2014). How these summaries are computed, however, is still a matter of debate. While some argue that perceivers encode all items in a distributed manner (Baek & Chong, 2020), others argue that only a subset of objects within a set are encoded (Allik et al., 2013; Maule & Franklin, 2016), recently evaluated as equal to the square root of number of items in a set (Whitney & Yamanashi Leib, 2018).

Whether people encode all items in a set or only a subset, it is plausibly that they preferentially attend to the most salient items in a set (Kanaya et al., 2018; Sweeny et al., 2013). If this is true, an estimation of the average could be more extreme than the true mean, because it would reflect oversampling, or overweighting of the more salient items in the set. A recent experiment explored this hypothesis, examining perceivers' accuracy in estimating the mean size of a set of circles that varied in diameter (Kanaya et al., 2018). Participants reported that the average circle size was larger than it actually was, suggesting that perceivers overweighed the largest, most salient items within a set. Furthermore, this tendency increased as more items were added to the set. According to Kanaya at el., considering the fact that people sample $\sqrt{N}$ items from a set, increasing the set size provided more opportunities for a sampling bias (Kanaya et al., 2018). These experiments provided evidence of amplification in the estimation of multiple

objects (for further support, see Sweeny, Haroz, & Whitney, 2012). However, unlike faces, larger circles necessarily occupy more space in the visual field and are therefore more likely to receive preferential representation in retinotopic visual areas (Schwarzkopf & Rees, 2013). It is thus unclear whether a similar effect of amplification would be evident with faces.

The Present Research

We conducted a series of experiments to test three pre-registered hypotheses (Studies 1, 2: https://osf.io/6rzw4, Study 3: https://osf.io/98k26). Our first hypothesis was that participants would show a crowd amplification effect, estimating a crowd's average emotion as more intense than it actually is. Our second hypothesis was that the amplification effect would be stronger as the number of faces within an array increased (following Kanaya et al., 2018). Our third hypothesis was that amplification would be more robust for negative emotions than for positive emotions. However, we were more equivocal about this third prediction in light of the contradictory findings in the existing literature, especially in the second pre-registration when some evidence had already been collected.

We tested these hypotheses in three experiments using a task in which participants evaluated mean emotions of crowds. Study 1 tested all three hypotheses. Study 2 replicated and extended Study 1 to test whether increased exposure time represents a boundary condition to the three effects observed in Study 1. Finally, Study 3 used eye-tracking to examine the particular pattern of attentional bias that gives rise to the crowdemotion amplification effect.

Study 1: Establishing the Crowd Amplification Effect

The goal of Study 1 was to test the crowd amplification hypothesis by asking participants to estimate the mean emotion of sets of faces (set range: one to twelve faces) that appeared on the screen for one second.

Method

Participants. To evaluate the appropriate sample size for the study we used data from a recent study on group categorization that provided initial evidence for the occurrence of amplification (Goldenberg, Sweeny, Shpigel, & Gross, 2019, Study 2). In this study, 30 participants saw 12 faces expressing neutral to angry facial expressions on the screen for one second (50 trials) and were asked to evaluate the crowd's average emotion. Results pointed to a crowd amplification effect. However, this study consistently presented participants with 12 faces, and only examined one emotion (anger). In the present study, we varied both the number of faces (using arrays that contained between 1 and 12 faces), as well as the emotions expressed by these faces (angry, happy). Based on these modifications we conducted a power analysis which suggested that a sample size of 50 participants completing 150 trials would provide power of 0.97 to support our first hypothesis (see Supplementary Materials for graph). Our final sample therefore included 50 participants (men: 23, women: 27; age: M= 19.52, SD = 1.69), each of whom completed 150 trials. Participants were recruited using the Stanford student pool, and received credit for their participation in the task. All participants were included in analysis.

Materials. In each trial, participants first saw an array containing 1-12 faces expressing different intensities of emotion from either neutral-to-angry (anger condition) or neutral-to-happy (happiness condition) continua (Figure 2A). The group average intensity, the valence of the faces, and the set size were all chosen randomly on each trial. We did not mix the happy and angry faces in the same set for two reasons. First, doing so could undermine our ability to detect

an amplification effect: if participants fixated on one extremely negative and one extremely positive face, then on average, their estimate about the crowd could appear to be relatively accurate despite the fact that they were biased by emotional intensity in their sampling of faces. Second, the most-happy and most-angry faces were not equal in intensity (see Supplementary Materials), thus making the average between the two different from zero.

For the neutral-to-emotion continua for the face arrays, we used a set that was developed for a recent investigation of ensemble face perception (Elias, Dyer, & Sweeny, 2017). Fifty modifications of a face were concatenated, from neutral to extremely angry and from neutral to extremely happy, effectively creating a visual rendering of 50-point anger or happiness scales (Figure 1). This was performed for four exemplar faces from the NimStim face set (Tottenham et al., 2009), creating eight different neutral-to-emotion continua, four positive and four negative, each ranging from 1 - 50. We conducted a pre-test to confirm that the evaluation of a single face was accurate, finding no amplification effect for a single face evaluation (see Supplementary Materials).

On every trial, an array containing between one and twelve faces was presented to participants. Faces could appear in 12 fixed locations on the screen, chosen randomly. In each trial, each fixed location was randomly jittered 15 pixels in one direction (up, down, left, right) to make sure that participants were not tracking the exact locations of the faces. Critically, arrays depicted only one of the four identities (Figure 1A), so that the estimation of average emotion would be as simple as possible, and could be done using a scale with the same facial identity as the set. Furthermore, this ensured that any effect of crowd estimation could not be caused by participants attending to some identities in a crowd more than others.

The group's mean emotional intensity was randomly set to be between 10-40 (based on the 1-50 scale of angry or happy faces, 1 being neutral and 50 very angry or very happy). We limited the range of the group means in order to allow distributions of intensity that were as close to uniform as possible within the 1-50 scale. The sets were engineered so that the standard deviation of a 12-face array was always 10, from which we randomly chose a subset of 1-12 faces. All of the faces were of white males. However, our results were not moderated by participants' race or gender (see Supplementary Materials).

To prevent residual visual processing, we immediately followed each face in the array with a mask at the same location. Each mask was generated by dividing an emotional face from our face space into 70 rectangular pieces and then randomly shuffling the locations of these pieces. This approach ensured that the emotional faces and scrambled masks resembled each other in terms of low-level image characteristics. Each mask appeared for 250ms.

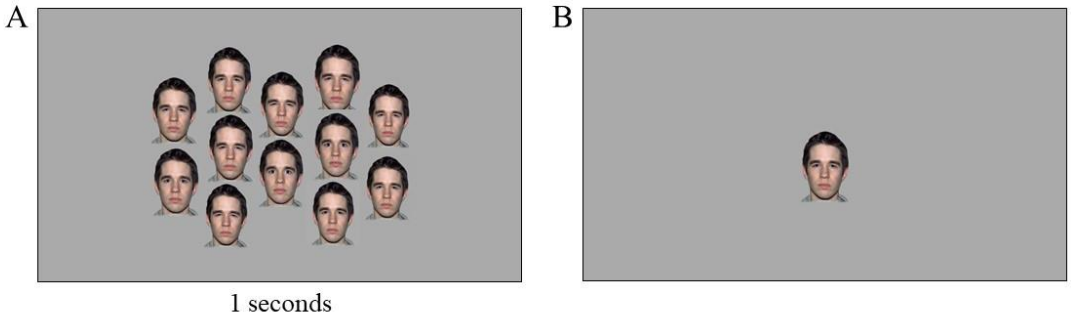

Figure 1. The task used in Study 1. Participants saw an array of 1-12 faces expressing either different degrees of anger or happiness that appeared on the screen for one second (A). Participants were then asked to evaluate the average emotion expressed by these faces by adjusting the intensity of a single morphed face (1-50, B).

Following the face array and the mask, participants were asked to evaluate the average emotion expressed in the array. To make their evaluations, a single face bearing a neutral expression was presented on the screen after the array was masked (Figure 2B). The identity of

the scale face matched that of the previous array, and was always anchored on a neutral expression: 1 on the scale from 1 – 50. This was done because previous research indicates that the initial location of the scale in ensemble coding tasks led to an anchoring effect, such that estimations were closer on average to that initial location (see for example Oriet & Brand, 2013). By starting each scale at neutral, our approach provided a conservative measure of amplification. Moving the mouse from left to right transformed the face from a neutral face to an angry or happy face (on a 1-50 scale, see Figure 2). Participants had as much time as they needed to estimate each group's mean intensity of emotion.

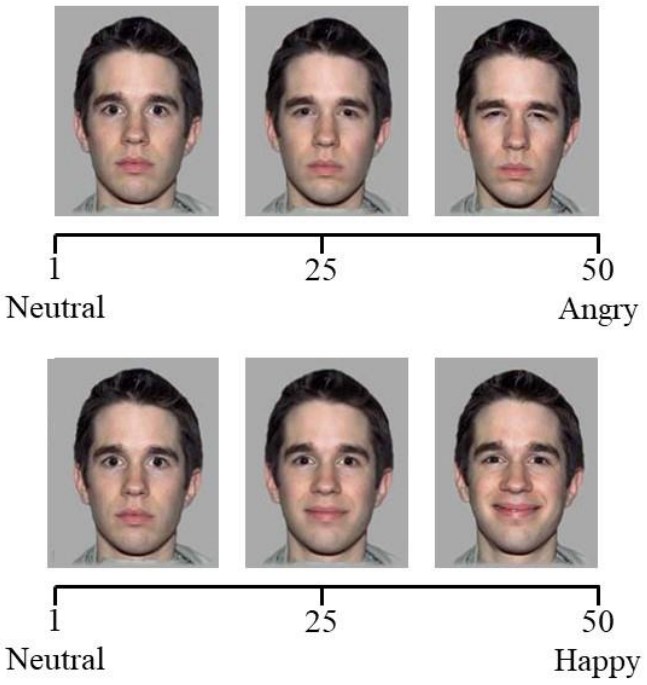

Figure 2. A sample of three faces from the neutral-to-angry scale (top) and from the neutral-to-happy scale (bottom) that were used in the studies. Values of 25 and 50 correspond to 50% and 100% intensities in our morph range, respectively.

After completing the main task, participants filled out a short survey which included an abbreviated version of the social interaction anxiety scale (SIAS-6, Peters, Sunderland, Andrews, Rapee, & Mattick, 2012), a need to belong scale (Leary, Kelly, Cottrell, & Schreindorfer, 2013),

a personality scale (Gosling, Rentfrow, & Swann, 2003), and demographic questions including age, gender, race and education level. These scales were administered for exploratory purposes (i.e., in order to examine potential moderators of the crowd amplification effect), so the corresponding results are reported in the Supplementary Materials.

Procedure. Ethics committee approval was obtained prior to the collection of data. Participants were told that they would take part in a study designed to test whether people can identify the average emotional expressions of crowds. Following instructions and a practice run, participants completed a task adapted from a previous ensemble coding task (Elias et al., 2017). The task included 150 trials. After viewing each face array for one second (following a 250ms mask), participants were asked to estimate the mean emotion expressed by the faces in the set.

Results

We tested three hypotheses (for exploratory analyses see Supplementary Materials). To measure amplification in estimation of the face sets, we conducted a mixed model analysis of repeated measures, comparing the actual mean emotion expressed in each set with participants' estimated mean emotion. As each participant was exposed to four face identities, we added two random intercepts: face-identity and participant. Supporting the crowd amplification hypothesis, estimated mean crowd emotion was 2.87 points higher (1-50 scale) than the actual mean crowd emotion (b =2.87 [2.53, 3.21], SE = .17, t (14533) = 16.8, $p < .001$, R^2= .05[1]).

We used a single model to test whether an increase in the number of faces (hypothesis 2) or the type of emotions expressed by the faces (hypothesis 3) influenced the crowd amplification effect. This allowed us not only to reduce the number of comparisons but also to test the interaction between the two variables. For our dependent variable, we created a difference score

[1] Calculated based on recommendations from (Xu, 2003)

between participants' estimation of the mean crowd emotion and the actual mean crowd emotion for each trial, such that positive numbers indicated amplification. We then conducted a mixed model analysis for repeated measures, with number of faces, valence of the faces, and their interaction predicting the degree of the difference between estimated and actual mean emotions. As in our previous analysis, we included random intercepts of face-identity and participant.

In line with our second hypothesis, number of faces significantly predicted an increase in amplification (b =.42 [.21, .64], SE = .03, t (7233) = 3.86, p < .001, R^2= .08, Figure 3)[2]. Supporting our third hypothesis, amplification was stronger for crowds expressing negative versus positive emotions (b =.24 [.02, .45], SE = .11, t (7253) = 2.18, p = .02, R^2= .08). The interaction between number of faces and the valence expressed by the faces was not significant (b =.08 [-.13, .30], SE = .11, t (7226) = .76, p > .25, R^2= .08).

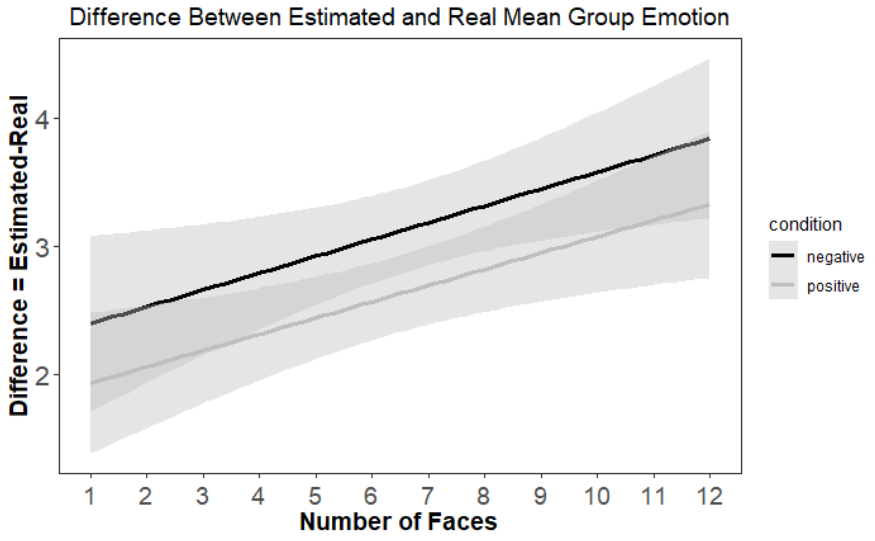

Figure 3. Difference between estimated and real mean for positive and negative face arrays as a function of number of faces in Study 1. Grey areas represent standard errors.

[2] Somewhat surprisingly, as is evident in Figure 3, we found an amplification effect even in single face trials. Exploratory analyses revealed that this was a carry-over effect (i.e., it was driven by the preceding trials). A pilot study confirmed that a series of single faces does not elicit an amplification effect. Please see Supplementary Materials for details.

In sum, participants estimated that face sets were more emotional than they actually were; amplification increased with set-size; and amplification was stronger for negative, compared to positive, emotions. We hasten to note, however, that this third effect was relatively weak, which is congruent with mixed results in the related literature. By what mechanism does our amplification effect arise? Is it driven by participants running out of time after making only one or two fixations such that the effect would be attenuated if participants had more time? One possibility is that increased exposure would allow participants to look at more faces in the set, thus affording them a larger sample from which to estimate the average, potentially minimizing or eliminating the amplification effect. However, a second possibility is that increased exposure would allow more time to fixate on more intense faces, which would instead *increase* the degree of amplification. We adjudicate between these predictions in Study 2 by varying exposure time.

Study 2: Modifying Exposure Time

The goal of Study 2 was to assess whether the crowd amplification effect varied as a function of exposure time.

Method

Participants. Similar to Study 1, we estimated that 50 participants completing 150 trials would provide more than 80% power to examine the crowd amplification affect. Our final sample therefore included 50 participants (men: 17, women: 33; age: M= 19.07, SD = 1.04). Participants were recruited using the Stanford student pool, and received credit for their participation in the task. No participants were removed from the analysis.

Measures. Measures were identical to those in Study 1.

Procedure. Ethics committee approval was obtained prior to the collection of data. The procedure was identical to that of Study 1 with one difference: we manipulated the exposure to

the face arrays to be either 1, 1.4 or 1.8 seconds (50 trials per each exposure condition within participant).

Results

To test our first hypothesis and measure amplification in the estimation of the face sets, we conducted a mixed model analysis of repeated measures, comparing the actual mean crowd emotion expressed in each set to the estimated mean crowd emotion, as we had done in Study 1 (collapsing across exposure conditions). Results indicated that the estimated mean crowd emotion was 2.39 points higher (in our 1-50 scale) than the actual mean crowd emotion (b =2.39 [2.06, 2.71], SE = .16, t (15,531) = 14.36, $p < .001$, R^2= .03), again supporting the crowd amplification effect hypothesis.

Having replicated the crowd amplification effect, we set out to test hypotheses 2-3, as well as the effect of time on amplification in one model. First, we re-examined whether a greater number of faces would lead to increased amplification. Second, we asked, once again, whether amplification would be stronger for negative versus positive emotions. Third, we considered whether amplification would be larger for longer exposure times. As in Study 1, we created a difference score between participants' estimation of the mean crowd emotion and the actual crowd emotion for each trial, such that a positive number indicated an amplification in participants' estimation of the mean. We then conducted an analysis similar to that of Study 1.

Supporting our second hypothesis (and replicating Study 1), results indicated that the number of faces significantly predicted the difference score (b =.28 [.07, .48], SE = .10, t (7591) = 2.66, p = .01, R^2= .08). Supporting our third hypothesis (and replicating Study 1), results suggested that amplification was stronger for face sets expressing negative versus positive emotions (b =.26 [.05, .46], SE = .10, t (7594) = 2.47, p = .01, R^2= .08). As in Study 1, the

interaction between valence to number of faces was non-significant (b =.07 [-.13, .28], SE = .10, t (7595) = .72, p > .25, R^2= .08). Finally, looking at whether amplification was stronger or weaker depending on exposure time, we examined time as a continuous variable, showing that longer presentation times led to an overall increase in amplification (b =0.84 [.21, 1.48], SE = .10, t (7595) = 2.62, p = .01, R^2= .08).

To learn more about the effect of time on amplification, we examined a model in which exposure time was converted to a categorical variable. To maintain the principle of k-1 comparisons, we decided to compare the 1 second condition to both the 1.4 second and 1.8 seconds conditions. To further understand whether the effect of time on amplification was moderated by the valence or by number of faces, we added an interaction term between exposure time and valence.

Looking first at the main effect of time, amplification was not significantly different between the 1 and 1.4 second trials (b =.02 [-.46, .53], SE = .25, t (7587) = .11, p >.25). However, amplification at the 1.8 second condition was significantly *larger* compared to the 1 second condition (b =.65 [.15, 1.16], SE = .25, t (7586) = 2.54, p =.01). Increasing exposure time led to stronger amplification, at least for the 1.8 seconds condition.

Next, we examined our interaction between categorical exposure time and valence. The interaction between 1 and 1.4 seconds and valence was non-significant (b =.20 [-.30, .71], SE = .25, t (7592) = .80, p > .25), but the interaction between 1 and 1.8 seconds and valence was significant (b =-.53 [-.02, -1.04], SE = .25, t (7593) = -2.06, p = .04). To understand the interaction better, we centered our model on different exposure times, in order to test the difference between positive and negative faces for each duration (Figure 4). Results suggest that both for the 1 second exposure (b =.38 [.02, .74], SE = .18, t (7591) = 2.10, p = .03) and for the

1.4 second exposure (b =59 [.23, .94], SE = .18, t (7591) = 3.23, p = .001), the difference between positive and negative faces was significant. However, when exposed to the face set for 1.8 seconds, there was no difference between the two valence conditions (b =-.15 [-.50, .21], SE = .18, t (7594) = -.82, p > .25). These results suggest that any attentional differences associated with valence disappear when participants can observe the crowd for longer periods of time (but that the overall amplification effect persists).

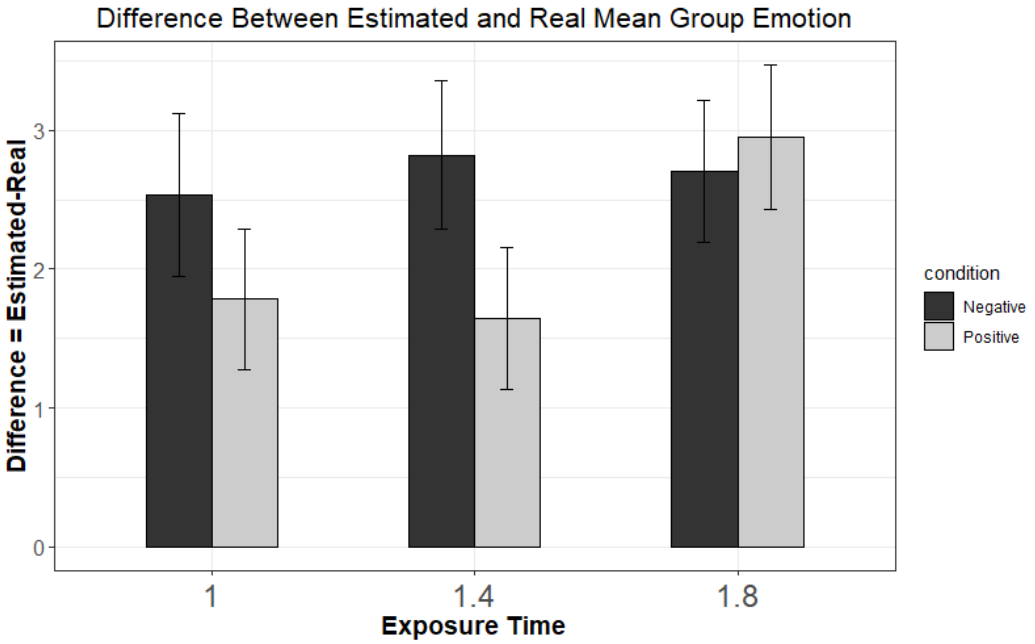

Figure 4. Difference between estimated and real mean for positive and negative face arrays, for three exposure times. Error bars represent 95% confidence intervals.

Overall, Study 2 replicated all of the results of Study 1 and revealed that longer exposure to the face set – within the limit of 1.8 seconds – led to greater amplification (this was driven more by the positive faces). One interpretation of these findings is that longer exposure increased attentional bias, such that participants spent more time looking at the more emotional faces. However, this interpretation can only be supported by a direct examination of fixation durations during the task, which we decided to examine via eye-tracking in Study 3.

Study 3:

Understanding the Mechanism for Amplification Using Eye-Tracking

The goal of Study 3 was to explore the proposed attention-bias mechanism driving the crowd amplification effect using eye-tracking.

Method

Participants. Similar to Studies 1 and 2, we recruited 50 participants who completed 150 trials (men: 11, women: 39; age: M= 22.26, SD = 3.55). Participants were recruited using the Harvard paid pool, and received $15 or course credit for participating in the study for 1 hour.

Measures. Our measures were identical to those in Studies 1 and 2, but also included the measures of eye movement and gaze described below.

Procedure. Ethics committee approval was obtained prior to the collection of data. The task was the same as that of Study 1 with two differences. First, participants underwent eye tracking while completing this task. Second, each array featured twelve faces (instead of a variable number of faces). This was done to maximize our ability to track users' attention when they were looking at a crowd. The twelve faces in each array were generated slightly differently from Studies 1 and 2 in order to allow changes in the array variance, such that each face was randomly chosen from the 1-50 scale, forming a normal distribution of the mean array.

In addition to this study and in order to utilize the hour participants spent in the lab, after completing the 150 trials and taking a short break, participants completed an additional study that was designed for a different publication.

Eye tracking apparatus. Participant eye gaze was recorded using an EyeLink 1000 Plus SR-Research eye-tracker and DataViewer software (Version 3.2). This device measures events related to eye-movement, including fixations (events where participants' gaze is held in one area)

and saccades (events where gaze shifts from one point of focus to another). The EyeLink 1000 Plus features a 500-Hz sampling rate and saccade resolution of .25 degrees. Participants first completed calibration and validation procedures using a 9-point calibration test to ensure that their gaze was being tracked accurately. Saccade onset and offset were determined with the EyeLink saccade algorithm. This algorithm uses a velocity threshold of 30 degrees per second, with a saccade acceleration threshold of 8000 °/s/s, and a saccade motion threshold of 0.1 degrees. Next, participants completed a 150-trial task, with each trial including twelve faces in an array. Arrays were presented on a screen with a display width of 800 x 600. Each face within the array served as a fixation area and subtended an overall visual angle of 2.75 x 2 degrees, and was separated from other faces by a minimum distance of 1.3 degrees with a jitter of .4 degrees. Dwell time was calculated as the time participants spent looking at each part of this fixation area.

Results

To measure amplification, we conducted a mixed model analysis of repeated measures, comparing the estimated mean crowd emotion and the actual mean crowd emotion expressed in each set (as in Studies 1 and 2). Because each participant was exposed to four face identities, we added face-identity and participants' identity as random intercepts. Replicating Studies 1 and 2, estimated mean crowd emotion was .73 points higher (1-50 scale) than the actual mean crowd emotion (b =.73 [.47, .98], SE = .13, t (14945) = 5.65, $p < .001$, R^2= .07), supporting the crowd amplification effect hypothesis. It bears noting, however, that this amplification affect was smaller than the effects found in Studies 1 and 2. One potential reason for this difference may have been the use of an eye-tracking tool, which could have changed how participants sampled the face arrays. It is also important to note that the R^2 of the model was actually higher in this study compared to both of the previous ones suggesting that outcome variance was also smaller.

Congruent with this idea, residual variance for the random effect was 63.05 in this model compared to 107.25 in Study 1 and 107.96 in Study 2.

Was amplification stronger for negative versus positive emotions (Hypothesis 3)? As in Studies 1 and 2, we used the difference score as an outcome variable and the valence of the face set as a predictor, using by-face-identity and by-participant random variables. Results indicated that there was no difference between positive and negative valence (b =-.01 [-.46, .42], SE = .22, t (7363) = -.09, p > .25). These results were different from Studies 1 and 2, which showed a relatively weak but significant difference between negative and positive faces. It is possible that the relatively smaller difference between actual and estimated means made it harder to detect a difference between negative and positive valence given our sample size.

We next examined participants' dwell time, or the amount of time participants spent looking at each of the faces in the set to test whether attentional bias could account for the amplification effect. On average, participants looked at ~ 3 faces in each set (M = 3.05, SD = 0.93). Therefore, the dwell time on each face was distributed such that most faces were not looked at (dwell time of 0), while the time spent on the rest of the faces was normally distributed (Figure 5). This type of data called for two separate analyses.

First, we examined whether an expression's emotionality made participants more likely to fixate on it (a binary outcome). This measure assesses attentional engagement: whether participants were initially drawn to faces expressing stronger (versus weaker) emotion. Second, we examined whether dwelling on a more emotional face increased the degree of amplification participants reported in their estimates of average emotions (a continuous measure). This measure assesses attentional disengagement: whether – once looking at a face – participants'

dwell time was longer if the face was more emotional, suggesting that they had a harder time detaching from these emotional faces.

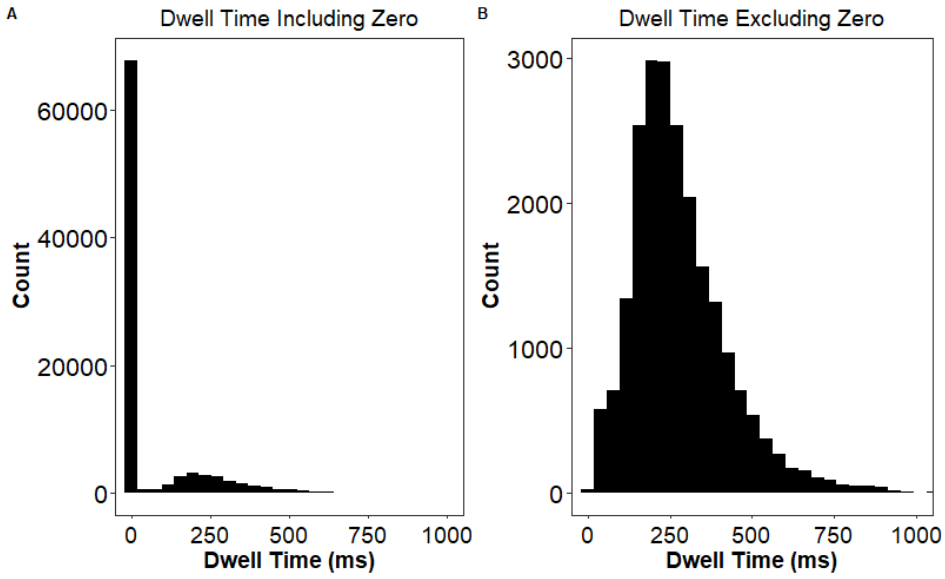

Figure 5. Histograms of dwell time in Study 3. Panel A is a histogram of dwell time to all faces, including interest areas that were not looked at (dwell time = 0). Panel B is a histogram of dwell time only for interest areas that were looked at by participants.

To examine our binary measure, we converted the dwell time variable into a binary outcome of either zero (participants did not look at this face at all) or one (participants looked at this face). We then conducted a generalized mixed-model analysis of repeated measures using the intensity of emotion expressed in each face as a predictor and our binary measure of dwell/not dwell as an outcome variable. Unlike all of our previous analyses, the analysis of dwell time necessarily included the time spent looking at each individual face (rather than on each array as a whole). Therefore, to make sure that comparisons between faces were similar across arrays with different average group emotion and standard deviation, we added the mean group emotion and standard deviation as covariates (though notably, removing these covariates did not change the significance of the results). Finally, we added by-participant and by-face-identity random intercepts as in our previous analysis. Results suggested that the emotionality of a face

did not affect the likelihood that participants would fixate on it or not (b =.00 [-.01,.01], SE = .01, z = .31, p > .25), pointing to the fact that the intensity of emotions expressed by the faces did not influence on whether these faces were looked at or not.

Next, we examined the continuous outcome, looking at whether dwell time was longer on more emotional faces than neutral faces. We focused our analysis on the faces that participants looked at in each trial. For these faces, we conducted a linear mixed model repeated measures analysis predicting time spent on these faces as a function of their emotional intensity. Two covariates were added to the analysis: mean and standard deviation of group emotions, and two random variables (a by-participants random variable and by-face-identity random variable). Results suggested that face emotional intensity significantly predicted dwell time, such that participants spent more time looking at more emotional faces compared to less emotional faces (b =.18 [.05, .31], SE = .06, t (22345) = 2.78, p < .001, R^2= .16).

Having established that dwell time was significantly longer for more emotional faces, we tested whether participants who spent more time looking at more emotional faces showed greater amplification. To examine this possibility, we created a coefficient for each participant of face emotional intensity predicting dwell time: specifically, we ran a linear regression for each participant looking at emotional intensity as a predictor for dwell time across trials (face emotional intensity predicting dwell time, looking only at faces that were observed by participants). We then took the regression coefficient for each participant and used it as an estimate of the tendency to dwell more on emotional faces, such that a positive coefficient indicated that a certain participant spent more time looking at emotional faces and a negative coefficient indicated that a certain participant spent more time looking at neutral faces. We examined the interaction between a dummy coded predictor for each value (i.e., whether the

value was the participant-estimated or actual mean for that trial) with each individuals' dwell amplification score predicting emotionality (b =1.42 [1.04, 1.79], SE = .19, t (14798) = 7.49, p < .001, R^2= .18). In other words, the gap between the estimated and actual mean emotionality values were larger for those participants who had higher dwell coefficients.

One question raised by these results is whether we can use the dwell-time data to predict participants' estimations of the mean, and whether this would be a better predictor of participants' estimation than the objective average of the picture. To answer this question we calculated the average of only the faces that participants looked at, weighted by the relative time participants spent on each face in each trial (the weighted-fixation average model). This model was compared to the actual average of all faces in each array. In order to estimate which of the two best predicted participants' evaluation of the average crowd emotion, we created a regression model that used both our weighted-fixation average model and the actual average of the face predicting participants' estimation of the group mean emotion.

Results indicated that both the actual average (b =.17 [1.04, 1.79], SE = .02, t (7296) = 6.44, p < .001, R^2= .02) and the weighted-fixation average model (b =.47 [1.04, 1.79], SE = .01, t (7221) = 37.97, p < .001, R^2= .02) were significant predictors of participants estimated average. However, the coefficient of the weighted fixation average coefficient was about three-times higher than the coefficient of the actual average, suggesting that in our data the weighted-fixation average was a better predictor of the average group emotion. Correlating each of the models suggests that the correlation between the weighted-fixation model and participants' estimations is .45 (r(7296) = .45 [.43, .37], p < .001), while the correlation between the actual and estimated mean is .26 (r(7296) = .22 [.24, .28], p < .001). We compared the two correlations by conducting a Pearson and Filon's Z test using the package 'cocor' (Diedenhofen, 2016). Results

indicated that indeed the correlation between the weighted-fixation model and participants' estimation was significantly better ($z = 16.70$, $p < .001$).

It is important to note, however, that although our weighted average model did fairly well in predicting participants' estimation, the correlation between the two was .45, which suggests that a contribution from other faces in the periphery could account for some of the residuals. Indeed, participants' estimation of the crowd emotion can be impacted by visual stimuli that appear in the periphery and are not fixated (Wolfe, Kosovicheva, Wood, & Whitney, 2015). One potential adjustment to the weighted-fixation average model would therefore be to include all faces in a crowd, including those in the periphery. However, further optimization is required to determine how to weight the fixated and non-fixated items in such a model, which is beyond the scope of the current investigation. More research should be done to understand the role of peripheral information in crowd-emotion amplification.

Overall, the results of Study 3 provided support for a potential mechanism for amplification. Participants spent more time to attending emotional faces as compared to neutral faces, and participants who were especially attentive to emotional faces showed the greatest degree of amplification in estimating the average group emotions.

General Discussion

How do people rapidly estimate crowds' emotions? Are such estimations accurate or systematically biased? These questions are relevant to many situations in which people must evaluate the emotions of others. Here we found that participants tended to overestimate the intensity of a crowd's average emotion. Furthermore, we observed that amplification grew as the number of faces within an array increased. Finally, results from Studies 1 and 2 (but not Study 3) showed that amplification was stronger for faces expressing negative versus positive emotion.

Mechanisms Underlying the Crowd Amplification Effect

One potential mechanism underlying the crowd amplification effect is attentional bias toward emotional faces, which leads to a corresponding bias in estimating the average emotion of the crowd. In Study 2, we tried to eliminate the possibility that such bias is an artifact of the one-second exposure time that was used in Study 1. We found that extending the exposure time led to an increase in the amplification effects rather than a reduction.

The use of eye-tracking in Study 3 allowed us to examine this mechanism more directly. Although participants did not appear to seek faces conveying stronger emotions (no differences in attention engagement), they did spend more time looking at highly emotional faces once they fixated on them (slower attention disengagement). Furthermore, results suggest that amplification increased as participants spent more time attending to emotional faces, providing support for the idea that this bias in attention disengagement contributes to the crowd amplification effect. However, it remains unclear how increased attention is translated into amplification. It is possible that increased attention contributes to stronger or more accurate visual memory of specific faces, which in turn shapes estimation of the mean (Brady & Alvarez, 2011).

Limitations and Future Directions

These experiments revealed a tendency for amplification that seems to be relevant to many aspects of social behavior. Nevertheless, these studies leave open several questions regarding how amplification is translated to perception and behavior outside the lab.

One limitation relates to the way the arrays were presented and rated. Participants viewed expression arrays that were all derived from the same face, and were asked to evaluate them using a single face. In real life, when people estimate the emotions of a crowd, they analyze

expressions of many different faces of various distributions of emotional intensity. Furthermore, it is not clear that participants were mentally representing the crowd using a single face (Kim & Chong, 2020), or at least would naturally, independent of our instruction to do so. Attempting to form perceptual summaries of sets of faces that vary in certain dimensions could reduce or increase tendencies for amplification. Other measures of crowd emotions that do not require translating a crowd into a single face measure may thus produce estimates of the degree of amplification that differ from what we found here. Further work should be done to examine the amplification effect we have demonstrated here in more naturalistic settings, such as when giving a talk or when observing large-scale gatherings, and using different measures. In addition, a test of how different emotional distributions affect amplification is also important. For example, recent work suggest that people tend to both discount (Haberman & Whitney, 2010) and overweight outliers (Dannals & Miller, 2017) depending on their degree of extremity. Whether strong emotional outliers increase or decrease amplification remains unclear.

A second limitation of this experiment is that participants were exposed solely to faces of white men. Although our findings suggested that the race and gender of the participants did not change amplification, the strength of the effect may be different when varying the gender and the race of the faces in a crowd (see Sanchez-Burks & Huy, 2009). Research on emotion perception has shown that white perceivers are much more sensitive to faces of black people, particularly those who express negative emotion (Ackerman et al., 2006; Hugenberg, 2005). It is therefore likely that amplification would be stronger for black versus white targets, particularly when their faces convey negative emotions. Recent research that compared ensemble coding of Korean and American participants revealed differences between within-culture and between-culture accuracy (Im et al., 2017). Additionally, the literature on gender stereotypes of emotions indicates that

people tend to assume that women are more emotional than men (Lafrance & Banaji, 1992; Plant, Hyde, Keltner, & Devine, 2000), which may increase estimations of amplification. Further studies should therefore examine these candidate moderating factors.

A third limitation of this research is the incomplete assessment of characteristics of the perceiver that may influence amplification. People who are anxious in social contexts spend more time attending to emotional faces, and therefore may be more likely to amplify crowd emotions than people who are not anxious (Bronfman, Brezis, Lazarov, Usher, & Bar-Haim, 2018; Mogg, Philippot, & Bradley, 2004). In our set of studies, only Study 1 revealed a correlation between social anxiety and amplification, but this sample was also more anxious on average than those in Studies 2 and 3 (see Supplementary Material). Further work is needed to examine this effect in a clinical population. If bias in attention drives amplification among individuals with social anxiety, training these individuals to focus their attention on less expressive members of the crowd may be useful in reducing their social anxiety (see for example Bauer, 2009).

To conclude, this project extends recent attempts to examine the role of ensemble coding in processes that are important to social functioning (Goldenberg et al., 2019; Lamer, Sweeny, Dyer, & Weisbuch, 2018; Phillips, Slepian, & Hughes, 2018). Learning more about how people rapidly evaluate complex social information may not only explain important aspects of social behavior, but may also open the door to future interventions.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., … Schaller, M. (2006). They all look the same to me (unless they're angry). *Psychological Science*, *17*(10), 836–840. https://doi.org/10.1111/j.1467-9280.2006.01790.x

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, *83*, 25–39. https://doi.org/10.1016/j.visres.2013.02.018

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. Retrieved from https://www.sciencedirect.com/science/article/pii/S1364661311000040

Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. *Attention, Perception, and Psychophysics*, *82*(1), 63–79. https://doi.org/10.3758/s13414-019-01827-z

Bauer, B. (2009). The danger of trail-by-trial knowledge of results in preceptual averaging studies. *Attention, Perception, and Psychophysics*, *71*(3), 655–665. https://doi.org/10.3758/APP

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. (2011). The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General*, *140*(4), 637–659. https://doi.org/10.1037/a0024060

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Bronfman, Z. Z., Brezis, N., Lazarov, A., Usher, M., & Bar-Haim, Y. (2018). Extraction of mean emotional tone from face arrays in social anxiety disorder. *Depression and Anxiety*, *35*(3), 248–255. https://doi.org/10.1002/da.22713

Bucher, A., & Voss, A. (2018). Judging the mood of the crowd: Attention is focused on happy faces. *Emotion*, *19*(6), 1044–1059. https://doi.org/10.1037/emo0000507

Cacioppo, J. T., Berntson, G. G., & Gardner, W. L. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*(1), 3–25. https://doi.org/10.1207/s15327957pspr0101

Dannals, J. E., & Miller, D. T. (2017). General social norm perception in groups with outliers social norm perception in groups with outliers. *Journal of Experimental Psychology: General*, *146*(9), 1342–1359.

Diedenhofen, B. (2016). *Package "cocor."* Retrieved from https://cran.r-project.org/web/packages/cocor/cocor.pdf

Eimer, M., & Holmes, A. (2007). Event-related brain potential correlates of emotional face processing. *Neuropsychologia*, *45*(1), 15–31. https://doi.org/10.1016/j.neuropsychologia.2006.04.022

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science*, *28*(2), 193–203. https://doi.org/10.1177/0956797616678188

Fridlund, A. J. (1991). Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology*, *32*(1), 3–100. https://doi.org/10.1016/0301-0511(91)90003-Y

Goldenberg, A., Sweeny, T. D., Shpigel, E., & Gross, J. J. (2019). Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000651

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, and Psychophysics*, *7*, 1825–1838. https://doi.org/10.3758/APP

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion*, *5*(3), 267–276. https://doi.org/10.1037/1528-3542.5.3.267

Im, H. Y., Chong, S. C., Sun, J., Steiner, T. G., Albohn, D. N., Adams, R. B., & Kveraga, K. (2017). Cross-cultural and hemispheric laterality effects on the ensemble coding of emotion in facial crows. *Culture and Brain*, *5*(2), 125–152. https://doi.org/10.1016/j.physbeh.2017.03.040

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879). https://doi.org/10.1098/rspb.2017.2770

Kim, M., & Chong, S. C. (2020). The visual system does not compute a single mean but summarizes a distribution. *Journal of Experimental Psychology: Human Perception and Performance*.

Kosonogov, V., & Titova, A. (2018). Recognition of all basic emotions varies in accuracy and reaction time: A new verbal method of measurement. *International Journal of Psychology*, *54*(5), 582–588. https://doi.org/10.1002/ijop.12512

Lafrance, M., & Banaji, M. (1992). Toward a reconsideration of the gender-emotion relationship.

In M. S. Clark (Ed.), *Review of personality and social psychology* (pp. 178–201). Thousand Oaks, CA: Sage Publication.

Lamer, S. A., Sweeny, T. D., Dyer, M. L., & Weisbuch, M. (2018). Rapid visual perception of interracial crowds: Racial category learning from emotional segregation. *Journal of Experimental Psychology: General*, *147*(5), 683–701. https://doi.org/10.1037/xge0000443

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the need to belong scale: Mapping the nomological network. *Journal of Personality Assessment*, *95*(6), 610–624. https://doi.org/10.1080/00223891.2013.819511

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A*, *33*(3), A22–A29. https://doi.org/10.1364/josaa.33.000a22

Mogg, K., Philippot, P., & Bradley, B. P. (2004). Selective attention to angry faces in clinical social phobia. *Journal of Abnormal Psychology*, *113*(1), 160–165. https://doi.org/10.1037/0021-843X.113.1.160

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), 381–396. https://doi.org/10.1037/0022-3514.80.3.381

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research*, *79*, 8–16. https://doi.org/10.1016/j.visres.2012.12.004

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11458–11463. https://doi.org/10.1073/pnas.172403899

Peters, L., Sunderland, M., Andrews, G., Rapee, R. M., & Mattick, R. P. (2012). Development of

a short form Social Interaction Anxiety (SIAS) and Social Phobia Scale (SPS) using nonparametric item response theory: The SIAS-6 and the SPS-6. *Psychological Assessment*, *24*(1), 66–76. https://doi.org/10.1037/a0024544

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology*, *114*(5), 766–785. https://doi.org/10.1037/pspi0000120

Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly*, *24*(1), 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x

Sanchez-Burks, J., & Huy, Q. N. (2009). Emotional aperture and strategic renewal: The accurate recognition of collective emotions. *Organization Science*, *20*(1), 22–34. https://doi.org/10.2139/ssrn.931492

Schwarzkopf, D. S., & Rees, G. (2013). Subjective size perception depends on central visual cortical magnification in human V1. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0060550

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 1–5. https://doi.org/10.1073/pnas.1908369116

Sun, J., & Chong, S. C. (2020). *Power of Averaging : Noise Reduction by Ensemble Coding of Multiple Faces*. *149*(3), 550–563.

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. https://doi.org/10.1152/jn.00061.2010

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. *Vision Research*, *64*, 26–34. https://doi.org/10.1016/j.visres.2012.05.008

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. https://doi.org/10.1037/a0028712

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. https://doi.org/10.1016/j.psychres.2008.05.006

van Kleef, G. A. (2009). How emotions regulate social life the emotions as social information (EASI) model. *Current Directions in Psychological Science*, *18*(3), 184–188.

Whitney, D., Haberman, J., & Sweeny, T. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J. Werner & L. M. Chalupa (Eds.), *The new visual neuroscience* (pp. 695–709). Boston, MA: MIT Press.

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

Wolfe, B. A., Kosovicheva, A. A., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision*, *15*(4), 11–11. https://doi.org/10.1167/15.4.11

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, *22*(22), 3527–3541.