Journal of Experimental Social Psychology 96 (2021) 104161

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Case Report

The group extremity effect: Group ratings of negatively and positively evaluated groups of faces are more extreme than the average ratings of their members☆

Sara C. Verosky [a,*], Lillian Tyack [a], Joel E. Martinez [b]

[a] Department of Psychology, Oberlin College, United States of America
[b] Department of Psychology, Princeton University, United States of America

ARTICLE INFO

Keywords:
Ensemble perception
Facial appearance
Group attractiveness
Group impressions
Impression formation
Trait judgments

ABSTRACT

When forming impressions of groups, people's impressions tend to reflect the average rating of the group members. However, group impressions have also been found to depart from an unweighted average of member ratings. For example, in the recently reported group attractiveness effect, groups were found to be more attractive than would be expected based on the average rating of the group members. In contrast, other studies have found that groups are rated as less attractive than would be expected. In two experiments, we found evidence for a group extremity effect that can help explain these prior findings. In this group extremity effect, group ratings of negatively and positively evaluated groups of faces are more extreme in either direction than would be expected based on the average ratings of the members. For negatively evaluated groups of faces, group ratings were significantly more negative than would be expected based on the average ratings of the group members. For positively evaluated groups of faces, group ratings were significantly more positive than would be expected based on the average ratings of the group members. The group extremity effect was larger for groups with more variability in the ratings of the group members, suggesting that attention to extreme group members underlies the effect. These data demonstrate how the biases involved in evaluating individuals based on appearance can be amplified when rating groups.

People routinely interact with groups made up of individuals. Being able to form impressions of such groups is important in a variety of contexts (Phillips, Weisbuch, & Ambady, 2014). For example, forming an impression of a group based on appearance can help someone decide whether to approach or avoid the group. Although a large literature examines the dimensions and concepts that people use to think about groups (Cuddy, Fiske, & Glick, 2008; Hamilton, 2007; Hamilton & Sherman, 1996), fewer studies examine how people use visual information to form impressions of groups.

Classic work on algebraic models of information integration has suggested that group impressions tend to reflect the average, and not the sum, of group member ratings (Anderson, 1965). An averaging rule has been shown to describe group impressions based on trait information (e.g., Anderson, 1965), as well as impressions based on appearance, including impressions of attractiveness, tenseness, and goodness (Anderson, Lindner, & Lopes, 1973; Levy & Richter, 1963). Other studies have also generally supported the idea that impressions of groups reflect an average of the member ratings (Luo & Zhou, 2018; Miller & Felicio, 1990; van Osch, Blanken, Meijs, & van Wolferen, 2015; Willis, 1960). However, some of these same studies also find that group ratings depart from what would be expected based on an unweighted average of group member ratings.

In the recently reported group attractiveness effect, people rated groups as more attractive than would be expected based on the average rating of the group members (van Osch et al., 2015). The group attractiveness effect has been observed for naturally occurring groups, but also for artificially created groups, with a larger effect size for groups with more members. Some evidence suggests that a mechanism of selective attention underlies the effect (van Osch et al., 2015). With this mechanism, selective attention to attractive group members is thought to bias overall ratings of the group in the direction of those group members.

☆ This paper has been recommended for acceptance by Dr. Rachael Jack.

* Corresponding author at: Department of Psychology, Oberlin College, 120 West Lorain, Oberlin, OH 44074, United States of America.
E-mail address: sverosky@oberlin.edu (S.C. Verosky).

https://doi.org/10.1016/j.jesp.2021.104161
Received 17 January 2020; Received in revised form 3 May 2021; Accepted 7 May 2021
Available online 29 May 2021

While the group attractiveness effect describes an increase in group ratings relative to the average of member ratings, two older studies that varied the overall attractiveness of groups of faces found different results for unattractive versus attractive groups of faces (Miller & Felicio, 1990; Willis, 1960). For attractive groups of faces, group ratings were more positive than would be expected based on the average of group member ratings, although this difference was only significant in one of the two studies (Willis, 1960). However, for unattractive groups of faces, the opposite was found to be the case: in both studies, unattractive groups of faces were evaluated more negatively than would be expected based on the average of group member ratings.

In their work on the group attractiveness effect, van Osch et al. (2015) did not manipulate the overall attractiveness of the groups. Examination of the ratings for the groups in their experiments reveals that approximately 80% of the groups had ratings that fell above the midpoint of the attractiveness scale. If they had included groups of faces that were less attractive, they might have found that these groups would be rated more negatively than would be expected based on the average of group member ratings. Interestingly, their proposed mechanism of selective attention could explain groups as being rated as more extreme in either direction: while selective attention to attractive members could bias group ratings upward, selective attention to unattractive group members could have the opposite effect.

Although attractive faces have been found to capture attention relative to unattractive faces (e.g., Maner et al., 2003; Sui & Liu, 2009), there is also evidence that unattractive faces are salient to perceivers. For example, when attractive and unattractive faces are matched on their distinctiveness, memory for unattractive faces is better than memory for attractive faces (Wiese, Altmann, & Schweinberger, 2014). Moreover, both negatively and positively evaluated faces have been found to elicit more activity in the amygdala than neutral faces (Mende-Siedlecki, Said, & Todorov, 2013). Face typicality has been offered as an explanation for these results: instead of coding face valence, the amygdala is thought to code distance from the average face, with stronger responses for faces that are further away from the average face (Mattavelli, Andrews, Asghar, Towler, & Young, 2012; Said, Dotsch, & Todorov, 2010).

While the studies discussed so far come from the perspective of social psychology, recent work on perception of groups combines a social psychological approach with insights from ensemble perception in vision science. The world contains an exquisite amount of visual detail, but the visual system has only a limited capacity for processing this information. Researchers believe that the visual system copes with this detail by extracting summary or ensemble information, such as information about the mean, from groups of similar-looking visual objects (Whitney & Leib, 2018). Ensemble perception has been reported for low-level visual features such as orientation and motion, and also for high-level visual features such as the average emotion or gender of groups of faces (Haberman & Whitney, 2007, 2009). With regard to trait judgments, people have been found to extract information about the amount of variance in dominance in facial appearance in groups (Phillips, Slepian, & Hughes, 2018) and information about the average attractiveness of groups of faces (Luo & Zhou, 2018).

When engaging in ensemble perception, research suggests that people sample information from a subset of the items rather than sampling all of the items (Whitney & Leib, 2018). Although multiple factors, such as the stimulus type, influence the number of items that are integrated in ensemble perception, a recent meta-analysis suggests that perceivers generally integrate information from a subset of items that is approximately equal to the square root of the set size (Whitney & Leib, 2018). Therefore, for a set of N items, this would mean that a perceiver would integrate information across a subset of items that is approximately equal to the square root of N.

When people sample items from a larger set, a recent study suggests that they do not do so randomly, but instead that they attend to the most salient items in the set (Kanaya, Hayashi, & Whitney, 2018). Relevant for the current work, attending to the most salient items results in a group amplification effect, whereby group estimates are biased toward the salient items (Kanaya et al., 2018). Although this initial study focused on judgments about properties of circles, a preprint suggests that this is also the case for judgments of the mean emotion of groups of faces (Goldenberg, Weisz, Sweeny, Cikara, & Gross, 2021). Both of these studies also found that the group amplification effects increase with increasing set size. Since perceivers have been shown to sample information from a subset of items equal to the square root of the sample size, increasing the set size means that perceivers are sampling a relatively smaller proportion of the items. Increasing the set size therefore provides more opportunity for the most salient items to bias sample estimates.

While work on ensemble perception examines how people form summary representations of groups, related work explores how being in a group influences the evaluation of the individual group members. In the cheerleader effect, individual faces are evaluated as more attractive when they are presented in a group than when they are presented alone (Walker & Vul, 2014). Walker and Vul (2014) initially proposed that the effect was due to hierarchical encoding: they suggested that people extract an average face from a set of faces, and that this average face, which is seen as more attractive than most of the individual faces that composed it due to its average properties (Langlois & Roggman, 1990), biases the evaluation of individual faces toward it. However, subsequent work suggested that hierarchical coding alone is not sufficient to explain all of the observed results (Carragher, Thomas, Gwinn, & Nicholls, 2019). Instead, a second mechanism, which may rely on social inferences about being a member of a group, is thought to contribute to the results (Carragher et al., 2019). More recently, the perceived emotional valence of scenes has also been found to be shifted toward the average evaluation of other, simultaneously presented, scenes (Alwis & Haberman, 2020).

In the current experiments, we were interested in the possibility that negatively and positively evaluated groups of faces would both be rated in more extreme terms than would be expected based on the average ratings of group members. To test this hypothesis, we manipulated the overall attractiveness of groups of faces and asked participants to rate both the groups and the group members. In Experiment 1, we aimed to establish the existence of a group extremity effect. In Experiment 2, we made changes to the experimental design to increase the strength of the group extremity effect, and we began to explore whether selective attention to extreme group members underlies the effect.

1. Experiment 1

In Experiment 1, we created groups of four faces, and we asked participants to evaluate both the groups and the individual members of the groups on physical attractiveness. We manipulated the overall attractiveness of the groups by varying the number of unattractive relative to more attractive faces in each group. This resulted in groups of faces with five set levels of attractiveness. For the most unattractive groups, we predicted that group ratings would be more negative than the averages of the member ratings. For the most attractive groups, we predicted that group ratings would be more positive than the averages of the member ratings. We expected that groups in the middle of the attractiveness spectrum would show these effects to a lesser degree. The hypothesis and methods for Experiment 1a were pre-registered on the Open Science Framework website https://osf.io/bkm6d).

Experiment 1b was identical to Experiment 1a, except that participants used a different rating scale to enter their responses. In Experiment 1a, participants entered their ratings for the groups of faces and the individual faces using a numerical rating scale ranging from 1 to 9. However, since group ratings include fewer ratings than the averages based on group member ratings, this left open the possibility that group ratings are more extreme simply because they are less precise. To help address this concern, participants in Experiment 1b used a continuous

slider to make their ratings. If group ratings in Experiment 1a are more extreme simply due to the rating scale, then we would not expect to observe a difference between group ratings and the averages based on group member ratings in Experiment 1b.

1.1. Method

All measures, manipulations, and exclusions in these experiments and the subsequent experiment are disclosed. Sample size in this experiment and the subsequent experiment was determined prior to any data analysis.

1.1.1. Participants

Thirty-nine undergraduates participated in both Experiment 1a (24 female, 15 male; $M_{age} = 18.6$, $SD_{age} = 0.8$; racial/ethnic background: 8% Black or African American, 21% East Asian, 8% Other, 3% South Asian, 69% White, 3% did not report, 10% of participants were multiracial and are included in more than one category; 5% Hispanic, 79% Not Hispanic, 13% Other, 3% did not report) and Experiment 1b (20 female, 19 male; $M_{age} = 19.0$, $SD_{age} = 1.0$; racial/ethnic background: 8% Black or African American, 21% East Asian, 5% Other, 67% White, 3% did not report, 3% of participants were multiracial and are included in more than one category; 8% Hispanic, 77% Not Hispanic, 15% Other, 3% did not report). This sample size resulted in 95% power to detect an effect with a standardized mean difference effect size of $d_z = 0.6$ at an alpha level of 5%. This effect size was chosen because the mean group attractiveness effect size for van Osch et al.' (2015) between-subject studies was $d = 0.6$. We used a conservative power level of .95 because van Osch and colleagues found that the group attractiveness effect increased with group size, and the groups in the current study were smaller than the groups in all but one of their experiments (group size: $M = 7.71$, $SD = 2.49$, range: 4–12). Participants were recruited through the Oberlin College subject pool. Participants in this experiment and the subsequent experiment gave consent in accordance with a protocol approved by the Oberlin College Institutional Review Board.

1.1.2. Stimuli

Forty faces with neutral facial expressions were selected from the Karolinska Directed Emotional Faces set (Lundqvist, Flykt, & Ohman, 1998). The faces were selected based attractiveness ratings collected by Oosterhof and Todorov (2008). The ratings are available in the form of z-scores on the Todorov lab website (http://tlab.princeton.edu/). Twenty faces were selected as unattractive faces and twenty were selected as more attractive faces. The sets of unattractive and more attractive faces were matched on the extremity of their z-score ratings relative to zero (unattractive: $M = -0.66$, $SD = 0.32$; more attractive: $M = 0.65$, $SD = 0.33$). We refer to the second set of faces as "more attractive" rather than as "attractive" because inspection of the raw ratings that we collected (as opposed to the z-scores) revealed that the ratings for many of the faces in this set fell below the midpoint of the attractiveness scale.

Groups were created by combining the faces into sets of four. To increase the number of groups, each face was shown in two separate groups (with the exception of one face that was accidentally shown in three groups, and another that was only shown in only one). Faces that appeared together in one group did not appear together in a second group. Since there were 40 faces, this meant there were 20 groups total.

We created groups with five set levels of attractiveness by varying the number of unattractive relative to more attractive faces in the group. Unattractive groups were made up of four unattractive faces, less unattractive groups were made up of three unattractive faces and one more attractive face, more neutral groups were made up of two unattractive and two more attractive faces, more attractive groups were made up of one unattractive face and three more attractive faces, and the most attractive groups were made up of four more attractive faces. As with labeling for the initial sets of faces, we used relative labels for the groups of faces that fell toward the upper end of the attractiveness spectrum because the ratings of these groups were not neutral or attractive in terms of the scale itself.

1.1.3. Procedure

1.1.3.1. Experiment 1a. Participants were asked to evaluate the physical attractiveness of the groups of faces and of the individual faces that made up those groups. Participants were instructed not to spend too much time thinking about their decisions, but instead to go with their gut response. Each trial began with a 1-s fixation cross. Following the fixation cross, a group of faces or an individual face was displayed in the center of the screen, along with a rating scale ranging from 1(not at all attractive) to 9(extremely attractive). Each group of faces was arranged in the shape of a square, made up of two rows and two columns. Assuming a viewing distance of approximately 45 cm, each face subtended $7^\circ \times 7^\circ$ of visual angle. The group of faces or individual face remained on the screen until participants entered their rating using the number keys on the computer keyboard.

The trials were blocked together so that participants rated all the groups of faces or all the individual faces before rating stimuli from the other condition. In the group condition, participants rated each of the 20 groups of faces one time. Because each individual face was a member of two separate groups, this meant each face appeared twice in the group condition. To keep the group and individual rating conditions as closely matched as possible, the individual faces were shown twice in the individual rating condition as well. In the individual rating condition, the faces were blocked together so that participants rated each of the faces a first time before rating any of them a second time. Across participants, the average correlation between the first and second ratings of the individual faces was .71 ($SD = 0.17$) in Experiment 1a, .76 ($SD = 0.16$) in Experiment 1b, and .83 ($SD = 0.14$) in Experiment 2. Whether participants started by rating groups of faces or individual faces was counterbalanced across participants. The stimuli in each condition were shown in a different random order for each participant.

1.1.3.2. Experiment 1b. Experiment 1b was identical to Experiment 1a, with the exception that participants used a continuous slider to make their ratings instead of an integer-based scale. The slider was anchored on either side with the labels "not at all attractive" and "extremely attractive". The slider did not have numbers displayed on it, but the responses were coded on a scale from 1 to 9, with values recorded to the hundredths place.

1.1.4. Analysis

The overall design of the experiment was a 5(attractiveness: unattractive, less unattractive, more neutral, more attractive, most attractive) x 2(rating type: group versus individual) x 2(order: individual ratings first or group ratings first) mixed ANOVA, where the first two factors were within-subjects and the last factor was between-subjects. In our pre-registered analysis plan, we indicated that we would analyze the data using a repeated-measures ANOVA, dropping the factor of rating order if it was not significant. However, in order to capture within-participant variability in each condition (instead of collapsing across the four groups in each condition) we later decided to analyze the data using a linear mixed-effects model.

To test whether groups at either end of the attractiveness spectrum would be rated in more extreme terms than would be expected based on the average ratings of group members, we performed a linear mixed-effects analysis predicting attractiveness ratings from group attractiveness, rating type, order, and the interactions between these factors using the lme4 package Version 1.1–21 (Bates, Maechler, Bolker, & Walker, 2015) for R (R Core Team, 2015). For each participant, the rating of each group in the individual rating condition was calculated by taking the mean of the four faces in the group. Since the individual faces were rated twice, the rating for each individual face was itself a mean of two

ratings. Group attractiveness was coded as the number of more attractive faces in a group, ranging from zero to four.

Within the linear mixed-effects model, group attractiveness was treated as a continuous predictor, and rating type and order were treated as categorical predictors. Participants were modeled as random effects with varying intercepts, and with varying slopes for the interaction between group attractiveness and rating type. Face group was also modeled as a random effect with varying intercepts. *P* values were obtained using the Satterthwaite degrees of freedom method from the package lmerTest Version 3.1–1 (Kuznetsova, Brockhoff, & Christensen, 2017). Planned comparisons and confidence intervals (CIs) were computed using the emmeans package Version 1.4.5 (Lenth, 2020). Effect sizes and CIs for the effect sizes were estimated using the "eff_size" function in the emmeans package (Lenth, 2020). Multiple comparisons were corrected for using the Benjamini and Hochberg method. The data and the script for the analysis of this experiment and the subsequent experiment are available on the Open Science Framework https://osf.io/swtqj/.

1.2. Results

1.2.1. Experiment 1a

Group attractiveness, quantified as the number of unattractive relative to attractive faces in a group, had a significant effect on group ratings, $F(1, 24.13) = 19.43$, $p < .001$. The interaction between group attractiveness and rating type, meaning whether the rating was based on a group rating or the average of the ratings of individual group members, was also significant, $F(1, 374.78) = 5.45$, $p = .02$. None of the other effects reached significance (rating type: $F(1, 42.38) = 3.14$, $p = .08$; order: $F(1, 38.97) = 0.004$, $p = .95$; attractiveness by order: $F(1, 38.36) = 0.07$, $p = .79$; rating type X order: $F(1, 42.38) = 0.15$, $p = .70$; attractiveness X rating type X rating order: $F(1, 374.78) = 0.20$, $p = .66$).

As the attractiveness of the groups increased, both the group ratings ($B = 0.31$, $SE = 0.06$, 95% CI [0.18, 0.44], $t(26.79) = 4.79$, $p < .001$), and the ratings based on the average of the ratings of the individual members of the groups increased ($B = 0.24$, $SE = 0.06$, 95% CI [0.11, 0.38], $t(26.02) = 3.80$, $p < .001$; Fig. 1a). However, as indicated by the interaction between group attractiveness and rating type, the influence of attractiveness was larger for group ratings than for ratings based on the average ratings of the group members (difference $B = 0.07$, $SE = 0.03$, 95% CI [0.01, 0.12], $d = 0.04$, $SE = 0.02$, 95% CI [0.005, 0.07]).

For unattractive groups of faces, group ratings were more negative than the average ratings of the individual group members, although this difference did not reach significance (difference = −0.21, $SE = 0.12$, CI [−0.52, 0.11], $t(42.38) = 1.77$, $p = .42$, $d = -0.12$, $SE = 07$, 95% CI [−0.26, 0.02]). As the attractiveness of the groups increased, the difference between the group ratings and the average of the ratings of the individual members systematically decreased (less unattractive: difference = −0.14, $SE = 0.11$, CI [−0.44, 0.16], $t(39.53) = 1.28$, $p = .52$, $d = -0.08$, $SE = 0.07$, 95% CI [−0.22, 0.05]; more neutral: difference = −0.07, $SE = 0.11$, CI [−0.37, 0.22], $t(39.00) = 0.67$, $p = .83$, $d = -0.04$, $SE = 0.07$, 95% CI [−0.18, 0.09]; more attractive: difference = −0.01, $SE = 0.12$, CI [−0.33, 0.31], $t(39.28) = 0.07$, $p = .95$, $d = -0.005$, $SE = 0.07$, 95% CI [−0.15, 0.14]). For the most attractive groups of faces, group ratings were slightly more positive than the average ratings of the individual group members, although again this difference was not significant (difference = 0.06, $SE = 0.13$, CI [−0.30, 0.41], $t(40.69) = 0.44$, $p = .83$, $d = 0.03$, $SE = 0.08$, 95% CI [−0.12, 0.19]).

1.2.2. Experiment 1b

The attractiveness of the group, $F(1, 28.66) = 34.21$, $p < .001$, rating type, $F(1, 39.00) = 17.56$, $p < .001$, and the interaction between attractiveness and rating type, $F(1, 39.00) = 11.40$, $p = .002$, all had significant effects on group ratings. None of the other effects reached significance (order: $F(1, 38.95) = 0.002$, $p = .97$; attractiveness X order: $F(1, 38.64) = 0.01$, $p = .94$; rating type X order: $F(1, 39) = 0.02$, $p = .88$; attractiveness X rating type X order: $F(1, 39) = 2.32$, $p = .14$).

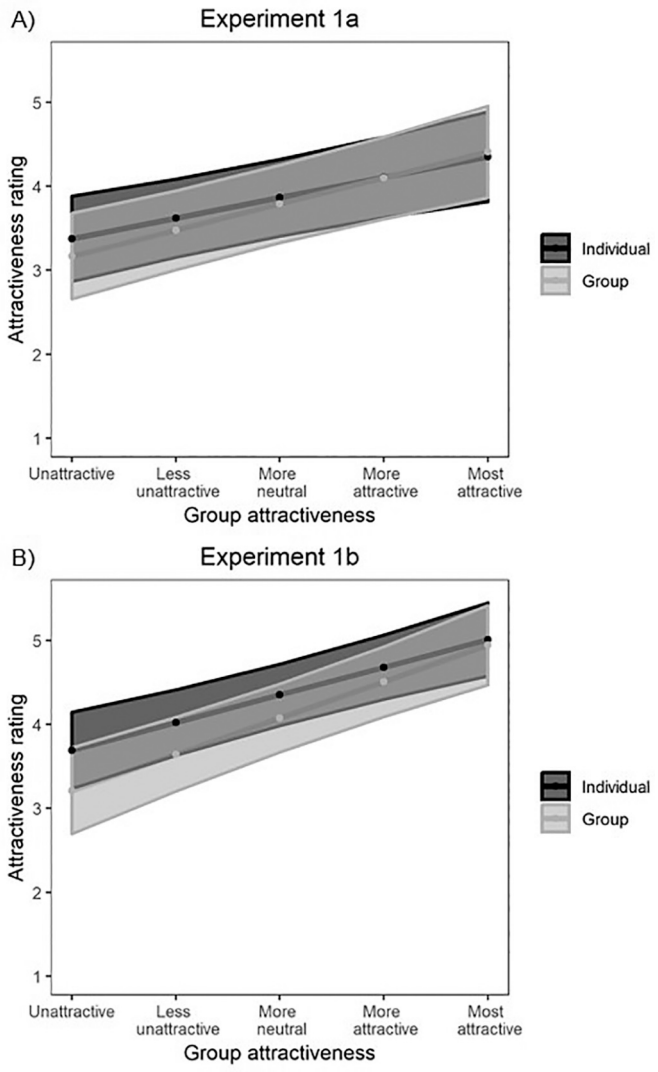

Fig. 1. Influence of rating type on attractiveness ratings for groups of faces with varying attractiveness in Experiments 1a and 1b. Predicted attractiveness ratings based on groups of faces (Group) or on the mean of the ratings of the individual members of the group (Individual). The only difference between Experiments 1a and 1b was the response scale: participants in Experiment 1a used an integer-based scale to make their ratings, while participants in Experiment 1b used a continuous slider. Error bars represent 95% confidence intervals of the regression slope.

Overall, group ratings were more negative ($M = 4.08$, $SE = 0.20$, 95% CI [3.67, 4.48]) than the ratings based on the average of the ratings of individual group members ($M = 4.35$, $SE = 0.18$, 95% CI [3.99, 4.72]; difference = −0.27, $SE = 0.10$, 95% CI [−.49, −.06], $d = -0.16$, $SE = 0.06$, 95% CI [−.29, −.04]). Once again, as the attractiveness of the groups increased, both the group ratings ($B = 0.43$, $SE = 0.07$, 95% CI [0.29, 0.58], $t(35.48) = 6.19$, $p < .001$), and the ratings based on the average of the ratings of the individual members of the groups increased ($B = 0.33$, $SE = 0.06$, 95% CI [0.20, 0.46], $t(26.68) = 5.17$, $p < .001$; Fig. 1b). However, as reflected by the significant interaction between attractiveness and rating type, the influence of attractiveness was again greater for the group ratings than for ratings based on the average ratings of the group members (difference $B = 0.10$, $SE = 0.03$, 95% CI [0.04, 0.16], $d = 0.06$, $SE = 0.02$, 95% CI [0.02, 0.10]).

For unattractive groups of faces, group ratings were significantly

more negative than the average ratings of the individual group members (difference = −0.48, *SE* = 0.11, CI [−.79, –.17], *t*(39.00) = 4.19, *p* < .001, *d* = −0.29, *SE* = 0.07, 95% CI [−.43, –.15]). As the attractiveness of the groups increased, the difference between the group ratings and the average of the ratings of the individual members once again decreased (less unattractive: difference = −0.38, *SE* = 0.11, CI [−.66, –.09], *t*(39.00) = 3.58, *p* = .002, *d* = −0.23, *SE* = 0.06, 95% CI [−.35, –.10]; more neutral: difference = −0.27, *SE* = 0.10, CI [−0.56, 0.01], *t* (39.00) = 2.63, *p* = .02, *d* = −0.16, *SE* = 0.06, 95% CI [−.29, –.04]; more attractive: difference = −0.17, *SE* = 0.11, CI [−0.48, 0.13], *t* (39.00) = 1.53, *p* = .17, *d* = −0.10, *SE* = 0.07, 95% CI [−0.24, 0.03]). However, for even the most attractive groups of faces, group ratings were still more negative than the average ratings of the individual group members, although this difference was not significant (difference = −0.07, *SE* = 0.13, CI [−0.41, 0.27], *t*(39.00) = 0.54, *p* = .59, *d* = −0.04, *SE* = 0.08, 95% CI [−0.19, 0.11]).

1.3. Discussion

Across Experiments 1a and 1b, we observed results consistent with a group extremity effect, where the influence of attractiveness on group ratings was greater for ratings based on the group than for ratings based on the average ratings of the individual group members. For unattractive groups of faces, ratings of the group were more negative than the average ratings of the group members. For the most attractive groups of faces, ratings of the group were slightly more positive than the average ratings of the group members in Experiment 1a, but not in Experiment 1b. More importantly, group ratings relative to the average of individual ratings changed with attractiveness in the predicted manner across both experiments.

Experiments 1a and 1b were very similar in design, differing only in the method that participants used to enter their responses. While participants in Experiment 1a used an integer-based response scale, participants in Experiment 1b used a continuous slider scale. Despite these different response methods, we found evidence in support of a group extremity effect across both experiments, suggesting that the observed results were not an artifact of the response scale. In fact, if anything, rather than diminishing the size of the group extremity effect, the use of a more fine-grained scale appeared to enhance it.

Since previous studies have documented group extremity effects at either end of the attractiveness spectrum (Miller & Felicio, 1990; van Osch et al., 2015; Willis, 1960), we suspect that the lack of significant differences for the more attractive groups had to do with the levels of attractiveness of the groups. While we selected the more attractive faces based on their attractiveness ratings, many of the ratings for the individual faces still fell below the midpoint of the attractiveness scale, meaning the groups were only attractive in relative terms. Moreover, the group size of four used in these experiments was smaller than the average group size previously used to demonstrate the group attractiveness effect (van Osch et al., 2015).

2. Experiment 2

The primary goal of Experiment 2 was to reproduce the results of Experiment 1, while allowing detection of group extremity effects at both ends of the attractiveness spectrum. To accomplish this, several changes were made to the experimental design. First, in order to increase the attractiveness of the attractive groups of faces, we used a different set of faces. At the same time, to make the neutral groups of faces more truly neutral, the composition of these groups was changed so that they were made up of faces with ratings that fell in between the ratings of the unattractive and attractive groups, instead of consisting of a mixture of unattractive and attractive faces. In addition, because the group attractiveness effect has been shown to be larger for groups of faces with more members (van Osch et al., 2015), the number of faces included in each group was increased from four to eight. Finally, because these other changes increased the number of individual faces in the experiment, the slightly unattractive and slightly attractive groups were dropped to reduce the length of the experiment. Along with these changes to the experimental design, we also increased the sample size. The hypothesis and methods for Experiment 2 were pre-registered on the Open Science Framework website https://osf.io/nv6qt.

A secondary goal of Experiment 2 was to explore whether selective attention to extreme group members drives the group extremity effect. To do this, we examined whether variability in group member ratings predicts the size of the group extremity effect. If selective attention underlies the group extremity effect, then we would expect groups with greater variability in the ratings of their group members, which offer selective attention more room to operate, to yield larger group extremity effects. Conversely, if selective attention is not responsible for the group extremity effect, then we would not expect to observe this relationship.

Finally, in order to begin investigating the role of time in the group extremity effect, we conducted an exploratory analysis examining whether reaction times during group ratings would predict the group extremity effect. If selective attention is driving the group extremity effect, then we might expect that longer reaction times would afford participants more time to attend to extreme group members, leading to more biased group ratings. However, at the same time, since a study from the ensemble perception literature shows that people are able to judge the attractiveness of groups after only short exposure times to the groups (Luo & Zhou, 2018), we also considered the possibility that additional time might not influence group ratings.

2.1. Method

2.1.1. Participants

Participants (*N* = 181; 82 female, 98 male, 1 other; M_{age} = 43.14, SD_{age} = 11.64; primary racial/ethnic background: 8% Asian/Pacific Islander, 8% Black or African American, 2% Hispanic or Latino, 1% Native American or American Indian, 2% Other, 80% White) were recruited from Amazon.com's Turk Prime website and each participant was compensated $1.20 for their time. We recruited only Master level workers with a 95% or higher approval rating who were located in the United States. In order to achieve a final sample size of *N* = 175, we aimed to collect data from 180 participants. However, one extra participant completed the task, resulting in a sample of 181 participants.

A sample size of *N* = 175 would have resulted in a 95% power to detect an effect with a standardized mean difference effect size of d_z = 0.25 at an alpha level of 5%. While the a priori power analysis for the previous experiment used a standardized mean difference effect size of d_z = 0.6, we chose to use a smaller effect size here in order to increase our chances of finding significant pairwise comparisons. We decided to use an effect size of d_z = 0.25 instead of an even smaller effect size of d_z = 0.2 because an effect size of d_z = 0.2 would have required a much larger sample size of *N* = 272 to achieve the same level of power.

As described in our pre-registered analysis plan, we initially intended to exclude participants who entered ratings equivalent to the starting value of the slider (the midpoint of the scale) or the end values of the slider on more than 10% of trials, under the assumption that only a few participants would meet this criterion. However, contrary to our assumption, nearly a third of participants met this criterion. To achieve a sample size close to what we had intended, we decided to report results based on the full sample. Importantly, removing the sixty participants who met the old exclusion criterion did not change which results reached significance.

2.1.2. Stimuli

Forty-eight faces with neutral facial expressions were selected from the Chicago Face Database based on attractiveness ratings made on a 1–7 Likert scale (Ma, Correll, & Wittenbrink, 2015). Because the faces receiving the lowest attractiveness ratings had a very different gender distribution than the faces receiving the highest attractiveness ratings,

only female faces were included in the experiment. Participant gender did not predict the attractiveness ratings of the faces on its own ($F(2, 180) = 1.05$, $p = .35$) or when it was included as an additional main effect in the main analysis ($F(2, 181) = 0.49$, $p = .61$). In addition, to help ensure a focus on attractiveness instead of on other visual features such as race/ethnicity, only faces with White appearance, which represents the majority racial/ethnic background where we conducted the study, were selected.

The sixteen least attractive faces were selected as the unattractive faces and the sixteen most attractive faces were selected as attractive faces. The mean for the unattractive faces fell below the midpoint of the scale ($M = 2.15$, $SD = 0.30$; scale midpoint = 4), and the mean for the attractive faces fell above the midpoint ($M = 4.76$, $SD = 0.24$). Sixteen faces with ratings that fell in between the means of the unattractive and attractive face sets were selected as neutral faces. Although the mean for the neutral set fell below the midpoint of the scale ($M = 3.46$, $SD = 0.13$), we refer to it as "neutral" both because of the way the faces were chosen and because the mean for this set is the closest to the midpoint of the scale out of the three sets.

Unattractive, neutral, and attractive groups of faces were created by combining the sixteen faces in each category into sets of eight. To increase the number of groups, each face was shown in two separate groups. Half of the faces that appeared together in one group appeared together in a second group. Since there were 48 faces, this meant that there were 12 groups total.

2.1.3. Procedure

The procedure was the same as that used in Experiment 1. Since the continuous slider response scale in Experiment 1b was more fine-grained and yielded stronger results than the integer-based scale used in Experiment 1a, we used a slider scale in the current experiment as well. Unlike in Experiment 1b, reaction times were recorded. Because the groups of faces were made up of eight faces instead of four faces, each group of faces was arranged in the shape of a rectangle, with two rows and four columns, instead of square. In addition, the presentation size for the faces was slightly smaller. Assuming a viewing distance of between 40 and 60 cm, each individual face subtended between $3.6^\circ \times 3.6^\circ$ and $5.4^\circ \times 5.4^\circ$ of visual angle.

2.1.4. Analysis

2.1.4.1. Main analysis. The only difference in the main analysis for this experiment versus the previous experiment was in the coding of group attractiveness. In the current experiment, group attractiveness values ranged from negative eight (eight unattractive faces) to zero (eight neutral faces) to eight (eight attractive faces). Once again, group attractiveness was treated as a continuous variable. This analysis and the following linear mixed-effects analyses used the same software packages as in Experiment 1.

2.1.4.2. Group variability. To test whether groups with greater variability in the ratings of their members yielded larger group extremity effects, we performed a linear mixed-effects analysis predicting the group extremity effect from the standard deviation of the ratings of the individual group members. Because we did not expect to observe a group extremity effect for neutral groups of faces, this analysis was based on data from only the unattractive and attractive groups of faces. The group extremity effect was quantified as the difference between group ratings and the average ratings based on individual group members. The difference was calculated in opposite directions for the unattractive and attractive groups of faces so that larger values always indicated larger group extremity effects.

We ran two models: a simple model without covariates, and a full model. In the simple model, group variability was treated as a continuous predictor, and participants were modeled as random effects with varying intercepts, with varying slopes for group variability. Face group was also modeled as a random effect with varying intercepts. In the full model, we predicted the group extremity effect from group variability, the attractiveness of the group, rating order, and the interactions between these factors. In this model, group variability was treated as a continuous predictor, and group attractiveness and order were treated as categorical predictors. Participants were modeled as random effects with varying intercepts, with varying slopes for the interaction between group variability and group attractiveness. Face group was once again modeled as a random effect with varying intercepts.

In addition to the group variability analysis described above, we also conducted a second, exploratory, test of the selective attention hypothesis (see Supplementary Materials). In this analysis, which was not meant to directly demonstrate the group extremity effect, we examined whether group members with more extreme ratings have a larger influence on group ratings than group members with less extreme ratings. As predicted by the selective attention hypothesis, we found that in both the unattractive and attractive conditions, the group ratings tended to be closest to the most unattractive or attractive, respectively, faces in the group.

2.1.4.3. Reaction time. To examine the role of reaction time in the group extremity effect, we performed a linear mixed-effects analysis predicting the group extremity effect from reaction times in the group rating condition. As with the previous analysis, this analysis was based on data from only the unattractive and attractive groups of faces. We ran two models: a simple model without covariates, and a more complex model. In the simple model, reaction time was treated as a continuous predictor, and participants were modeled as random effects with varying intercepts, and with varying slopes for reaction time. Face group was also modeled as a random effect with varying intercepts. For the more complex model, the full model failed to converge, and therefore we ran a reduced model predicting the group extremity effect from reaction time, group attractiveness, and order. In this model, reaction time was treated as a continuous predictor, and group attractiveness and order were treated as categorical predictors. Participants were modeled as random effects with varying intercepts, and with varying slopes for reaction time and group attractiveness. Face group was once again modeled as a random effect with varying intercepts.

2.2. Results

2.2.1. Findings

The attractiveness of the group, $F(1, 38.40) = 832.38$, $p < .001$, the two-way interaction between attractiveness and rating type, $F(1, 181.00) = 110.41$, $p < .001$, and the three-way interaction between attractiveness, rating type, and rating order, $F(1, 181.00) = 19.48$, $p < .001$, all had significant effects on group ratings. None of the other effects reached significance (rating type: $F(1, 181.00) = 1.37$, $p = .24$; order: $F(1, 180.88) = 0.002$, $p = .97$; attractiveness X order: $F(1, 180.79) = 3.12$, $p = .08$; rating type X order: $F(1, 181.00) = 1.69$, $p = .20$).

As can be seen in Fig. 2, the two-way interaction between attractiveness and rating type reflected a stronger influence of attractiveness for group ratings than for ratings based on the average ratings of the group members. This was the case both when participants rated the individual faces first (group rating: $B = 0.28$, $SE = 0.01$, 95% CI [0.26, 0.31]; average of individual ratings: $B = 0.22$, $SE = 0.01$, 95% CI [0.20, 0.23]; difference $B = 0.07$, $SE = 0.01$, 95% CI [0.06, 0.08], $t(181.00) = 10.89$, $p < .001$, $d = 0.06$, $SE = 0.01$, 95% CI [0.05, 0.07]) and when they rated the groups first (group rating: $B = 0.24$, $SE = 0.01$, 95% CI [0.22, 0.27]); average of individual ratings: $B = 0.21$, $SE = 0.01$, 95% CI [0.19, 0.23]; difference $B = 0.03$, $SE = 0.01$, 95% CI [0.01, 0.04], $t(181.00) = 4.18$, $p < .001$, $d = 0.03$, $SE = 0.01$, 95% CI [0.01, 0.04]). However, as indicated by the three-way interaction, the two-way

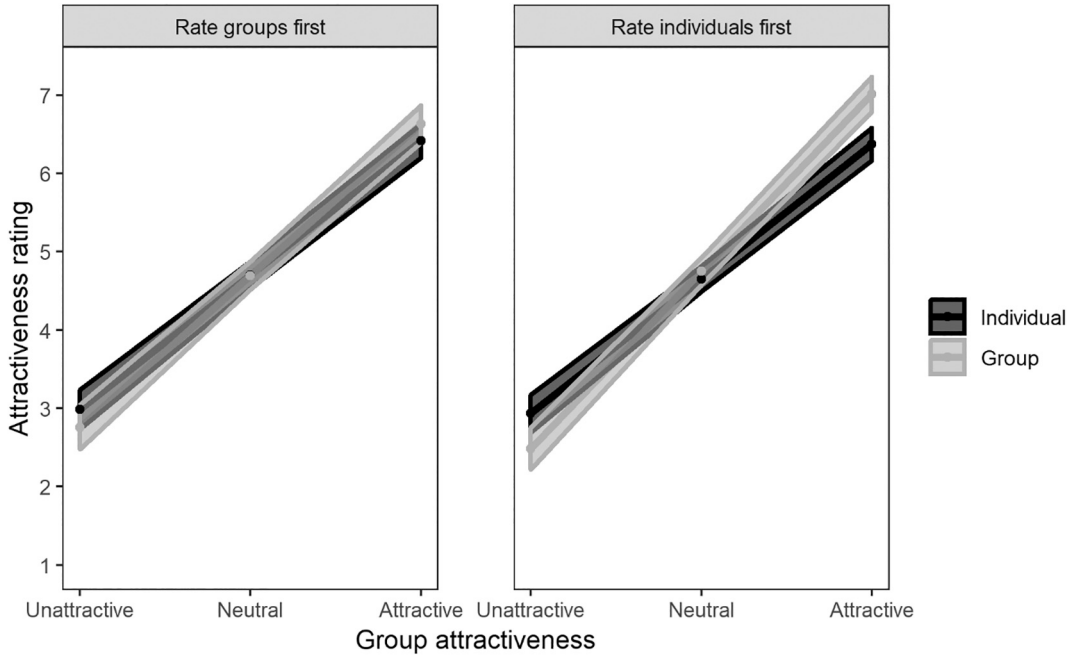

Fig. 2. Influence of rating type on attractiveness ratings for groups of faces with varying attractiveness by rating order in Experiment 2. Predicted attractiveness ratings based on groups of faces (Group) or on the mean of the ratings of the individual members of the group (Individual) for participants who rated the groups of faces (left panel) versus the individual faces first (right panel). Error bars represent 95% confidence intervals of the regression slope.

interaction between attractiveness and rating type was stronger when participants rated the individual faces before the groups of faces (difference $B = 0.04$, SE = 0.009, 95% CI [0.02, 0.06], $d = 0.03$, $SE = 0.01$, 95% CI [0.01, 0.04]). Examining this three-way interaction more closely, the slopes for the ratings based on the average ratings of individual members were similar across order conditions (difference $B = 0.001$, $SE = 0.01$, 95% CI [−0.02, 0.02], $t(180.73) = 0.06$, $p = .96$, $d = 0.001$, $SE = 0.01$, 95% CI [−.02, 0.02]). In contrast, there was a stronger influence of attractiveness for the slopes based on group ratings when participants rated the individual faces versus the groups of faces first (difference $B = 0.04$, $SE = 0.01$, 95% CI [0.01, 0.07], $t(180.89) = 2.92$, $p = .004$, $d = 0.04$, $SE = 0.01$, 95% CI [0.01, 0.06]).

Despite the three-way interaction, the pairwise comparisons at either end of the attractiveness spectrum were significant regardless of order. For unattractive groups of faces, group ratings were significantly more negative than the average ratings of group members (individual ratings first: difference = −0.45, $SE = 0.07$, 95% CI [−.64, –.26], $t(181.00) = 6.43$, $p < .001$, $d = −0.41$, $SE = 0.06$, 95% CI [−.54, –.28]; group ratings first: difference = −0.23, $SE = 0.07$, 95% CI [−.43, –.03], $t(181.00) = 3.06$, $p = .005$, $d = −0.21$, $SE = 0.07$, 95% CI = [−.34 –.07]). In contrast, for attractive groups of faces, group ratings were significantly more positive than the average ratings of group members (individual ratings first: difference = 0.64, $SE = 0.07$, 95% CI [0.44, 0.84], $t(181.00) = 8.54$, $p < .001$, $d = 0.58$, $SE = 0.07$, 95% CI [0.45, 0.72]; group ratings first: difference = 0.22, $SE = 0.08$, 95% CI [0.01, 0.43], $t(181.00) = 2.73$, $p = .01$, $d = 0.20$, $SE = 0.07$, 95% CI [0.05, 0.34]). Finally, for neutral groups of faces, group ratings did not differ from the average ratings of the individual group members (individual ratings first: difference = 0.09, $SE = 0.05$, 95% CI [−0.05, 0.23], $t(181.00) = 1.80$, $p = .09$, $d = 0.09$, $SE = 0.05$, 95% CI [−0.01, 0.18]; group ratings first: difference = −0.01, $SE = 0.06$, CI [−0.15, 0.14], $t(181.00) = 0.09$, $p = .93$, $d = −0.005$, $SE = 0.07$, 95% CI [−0.10, 0.10]).

2.2.1.1. Group variability. Because the unattractive and attractive groups of faces consisted of faces with similar levels of attractiveness, mean group variability was low ($M = 0.96$, $SD = 0.44$). Nevertheless, as group variability increased, the size of the group extremity effect also increased ($F(1, 122.97) = 9.78$, $p = .002$; $B = 0.25$, $SE = 0.08$, 95% CI [0.09, 0.41]). When the attractiveness of the group, rating order, and the interactions between group variability and these factors were added to the model as predictors, the influence of group variability on the group extremity effect only decreased slightly ($F(1, 133.17) = 7.56$, $p = .007$; $B = 0.21$, $SE = 0.08$, 95% CI [0.05, 0.36]). The effect of order was also significant in this model ($F(1, 129.15) = 5.49$, $p = .02$), such that participants who evaluated the individual faces first ($M = 0.53$, $SE = 0.06$, 95% CI [0.42, 0.65]) showed a larger group extremity effect compared to those who evaluated the groups of faces first ($M = 0.22$, $SE = 0.06$, 95% CI [0.10, 0.34]; difference = 0.32, $SE = 0.07$, 95% CI [0.17, 0.46], $d = 0.23$, $SE = 0.05$, 95% CI [0.12, 0.34]). None of the other effects reached significance (group attractiveness: $F(1, 98.26) = 1.29$, $p = .26$; group variability X group attractiveness: $F(1, 153.57) = 0.02$, $p = .89$; group variability X order: $F(1, 126.75) = 0.05$, $p = .82$; group attractiveness X order: $F(1, 159.13) = 0.12$, $p = .73$; group variability X group attractiveness X order: $F(1, 145.06) = 0.02$, $p = .88$).

2.2.1.2. Reaction time. The mean reaction time for rating the groups of faces was just under 5.5 s ($M = 5.41$, $SD = 3.89$). As reaction times increased, the size of the group extremity effect also increased ($F(1, 10.27) = 7.08$, $p = .02$; $B = 0.02$, $SE = 0.01$, CI [0.004, 0.04]). When group attractiveness and order were added to the model as predictors, the influence of reaction time on the group extremity effect did not substantially change ($F(1, 353.71)$,=7.65, $p = .006$; $B = 0.02$, $SE = 0.01$, 95% CI [0.005, 0.03]). The effect of order was also significant in this model ($F(1, 180.94) = 19.31$, $p < .001$), such that the group extremity effect was larger for participants who rated the individual faces first ($M = 0.54$, $SE = 0.06$, 95% CI [0.42, 0.66]) versus the groups of faces first ($M = 0.23$, $SE = 0.06$, 95% CI [0.10, 0.35]; difference = 0.32, $SE = 0.07$, 95% CI [0.17, 0.46], $d = 0.26$, $SE = 0.06$, 95% CI [0.14, 0.37]). The effect of group attractiveness was not significant ($F(1, 21.28) = 2.06$, $p = .17$).

2.3. Discussion

In Experiment 2, we observed a group extremity effect, where groups at either end of the attractiveness spectrum were rated in more extreme terms than would be expected based on the average ratings of the individual group members. After changes to the experimental design and an increase in sample size, pairwise comparisons between the group ratings and the ratings based on the average ratings of individual members were significant at both ends of the attractiveness spectrum. In support of the selective attention hypothesis, the group extremity effect was found to be larger for groups of faces with greater variability in the ratings of their group members. The group extremity effect was also found to be larger for trials where participants took longer amounts of time to evaluate the groups of faces.

We unexpectedly found an influence of rating order on the size of the group extremity effect. Although the group extremity effect was significant regardless of rating order, it was larger for participants who rated the individual faces before the groups of faces. In contrast, van Osch et al. (2015) found the opposite effect of order in one of their experiments, such that the group attractiveness effect was larger for participants who evaluated a group of faces before the individual group members. Looking to other experiments, rating order is not relevant for studies of ensemble perception because these studies tend to use a different type of experimental design, where the mean of the group is compared to a probe stimulus.

It is not clear how rating individual faces before rating a group influences attention to group members. In line with our observed results, rating the individual faces first may allow participants to spend more time focusing on the extreme faces when viewing the group, leading to more biased ratings. Alternatively, as proposed by van Osch et al. (2015), rating the individual faces first could encourage participants to focus their attention more equally among group members. Since it is possible to argue that order effects in either direction support a mechanism of selective attention, this finding should be interpreted with caution.

3. General discussion

Across three experiments, we found evidence for a group extremity effect, where group ratings of negatively and positively evaluated groups were more extreme than the average ratings of individual group members. While Experiments 1a and 1b found a pattern of results consistent with the group extremity effect, Experiment 2 additionally demonstrated significant pairwise comparisons at both ends of the attractiveness spectrum. For unattractive groups of faces, group ratings were significantly more negative than the average ratings of the individual group members. In contrast, for attractive groups of faces, group ratings were significantly more positive than the average ratings of the individual group members.

Although Experiment 2 focused on only unattractive, neutral, and attractive groups of faces, we expect that the group extremity effect would emerge to a lesser degree for less extreme groups of faces. Indeed, the systematic change in group ratings versus the average of the individual ratings with increasing attractiveness in Experiment 1 supports this view. This finding also fits with a meta-analysis of the experiments in the group attractiveness effect paper, which found larger group attractiveness effects for more attractive groups (van Osch et al., 2015). Thus, while the exact level of attractiveness where pairwise comparisons become significant will depend on both the stimuli and the sample size, these data suggest that the group extremity effect is likely to be more pronounced for groups with more extreme ratings.

A potential mechanism underlying the group extremity effect is selective attention to extreme group members. In support of this mechanism, we found that unattractive and attractive groups of faces with more variability in the ratings of their individual group members yielded larger group extremity effects. Similarly, for the group attractiveness effect, individual participants with more variability in their ratings of group members have been found to show larger group attractiveness effects (van Osch et al., 2015). At the same time, studies demonstrating that ensemble judgments are biased toward extreme group members provide converging evidence for the selective attention hypothesis (Goldenberg et al. 2021; Kanaya et al., 2018).

In the current experiments, participants were given unlimited time to evaluate both the individual faces and the groups of faces. In Experiment 2, where reaction time was measured, participants took 5.5 s on average to evaluate the groups of faces. Under these conditions, the group extremity effect was larger when participants responded more slowly, suggesting that longer amounts of time may allow people more time to attend to extreme faces. Unlike in the current work, the two studies demonstrating that ensemble judgments are biased toward extreme group members both displayed stimuli for limited amounts of time. For judgments of circle size, Kanaya et al. (2018) found that size judgments for sets of circles are biased after presentation times of 3 s. Meanwhile, Goldenberg et al. (2021) found that emotion judgments for sets of faces are biased after presentation times of 1 s, and that longer presentation times of 1.8 s yield larger effects. Together, these data suggest that the group extremity effect would also be present after limited presentation times, but that it is likely to be larger after unlimited presentation times.

Although the data support the hypothesis that selective attention underlies the group extremity effect, there are still a number of open questions. For example, the two conditions that were shown to increase the group extremity effect, namely increased variability and longer response times, can both be seen as leading to decreased ensemble perception. With group variability, increased variability means that the average is less representative of the set as a whole. As variability increases, there may come a point when people cease to rely on ensemble perception and instead simply sample individual items from the set. With time, longer amounts of time mean that people are better able to attend to individual items, which may make them less reliant on ensemble representations. In support of this idea, accuracy in identifying the average of a set of faces has been found to decrease with presentation times of 3200 ms and longer (Neumann, Ng, Rhodes, & Palermo, 2018). In the current work, it is not clear where to draw the line between people weighing individual items more heavily in their group judgments versus basing their judgments on single items. On a related note, there is an active debate in the ensemble perception literature about what happens to items in a set that are not attended. One possibility is that selective attention means that participants see only a subset of items in a set (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013), while another possibility is that they see all of the items in the set, but simply attend to some over others (Baek & Chong, 2020).

A limitation of the current experiments is that the groups were relatively homogenous in appearance. For example, the groups in Experiment 2 were made up of young female faces with White appearance. Although the groups were designed this way in order to help ensure a focus on physical attractiveness, it is not clear what would happen with more diverse groups. For example, because participants in the current experiments varied in their own race/ethnicity, this meant that participants differed in whether they were viewing groups of same- or other-race/ethnicity faces. Future work should investigate how race/ethnicity influence group impressions by explicitly including racial/ethnic variation among both the participants and stimuli. A related question is what would happen with groups of faces where the faces within the group differ from each other along dimensions other than physical attractiveness. When extracting the mean expression from a set of emotional faces, the visual system seems to unintentionally discount emotional outliers, at least after short presentation times (Haberman & Whitney, 2010). Future work should examine whether the visual system does something similar for outliers along other dimensions, and for stimuli presented for longer times.

In the current experiments, the groups of neutral faces were constructed in two different ways. In Experiment 1, the neutral groups were

made up of two unattractive faces and two more attractive faces. In Experiment 2, the neutral groups were created to be more truly neutral and they were made up of faces that themselves had ratings that were close to the neutral group mean. In both experiments, the neutral groups were meant to serve as placeholders between the less attractive and more attractive groups of faces, and, as expected, they had ratings that fell between the other groups of faces. However, groups made up of unattractive and attractive faces, like those in Experiment 1, could be used to address questions about how selective attention operates, such as whether people base their ratings on both unattractive and attractive faces, whether they are biased toward one type of face or the other, and whether this changes with longer presentation times. For example, with short presentation times, it seems likely that people will base their group ratings on multiple group members, yielding relatively neutral group ratings. If group ratings became less neutral with longer presentation times, this would suggest that people were paying attention to fewer group members or shifting their attention to group members with one type of rating. A related question, which has begun to be addressed in the context of judgments of group emotion (Goldenberg et al. (2021)), is whether people's group ratings are biased because they are drawn to attend to particular faces or because they have trouble disengaging from other faces.

Although the current experiments focused on judgments of physical attractiveness, attractiveness judgments are highly correlated with other trait judgments based on appearance (Oosterhof & Todorov, 2008). Since judgments of stimuli as different from each other as the mean size of circles (Kanaya et al., 2018) and the mean emotional expression of faces (Goldenberg et al. 2021) have both been found to be biased by extreme group members, we think it is likely that the group extremity effect will extend from attractiveness to other trait judgments, including judgments that are less highly correlated with attractiveness.

Judgments about attractiveness and other trait judgments reflect not only the physical properties of the face, but also knowledge about social categories and specific people (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). This suggests that prior learning is likely to play a role in the group extremity effect. Consistent with this observation, our findings can be seen as fitting with a recent study demonstrating that groups can exacerbate racial stereotyping (Cooley & Payne, 2019). In this study, groups were rated as more aggressive than individuals overall, but the effect was especially pronounced for groups of Black versus White faces. While Cooley and Payne (2019) focused on racial stereotypes, our results suggest that other types of learning that influence the evaluation of individual faces would likely have a similar effect on group judgments.

In summary, we found a group extremity effect, where group ratings of negatively and positively evaluated groups of faces are more extreme than average ratings of individual group members. The group extremity effect demonstrates how groups can amplify judgments based on facial appearance, with important implications for social interactions.

4. Open practices

The hypothesis and methods for Experiment 1a (https://osf.io/bkm6d) and Experiment 2 (https://osf.io/nv6qt) were pre-registered on the Open Science Framework website. The data and scripts for the analyses are also available on the Open Science Framework website (https://osf.io/swtqj/).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104161.

References

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

Alwis, Y., & Haberman, J. M. (2020). Emotional judgments of scenes are influenced by unintentional averaging. *Cognitive Research: Principles and Implications, 5*. Article 28.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*, 394–400.

Anderson, N. H., Lindner, R., & Lopes, L. L. (1973). Integration theory applied to judgments of group attractiveness. *Journal of Personality and Social Psychology, 26*, 400–408.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Carragher, D. J., Thomas, N. A., Gwinn, O. S., & Nicholls, M. E. R. (2019). Limited evidence of hierarchical encoding in the cheerleader effect. *Scientific Reports, 9*. Article 9329.

Cooley, E., & Payne, B. K. (2019). A group is more than the average of its parts: Why existing stereotypes are applied more to the same individuals when viewed in groups than when viewed alone. *Group Processes & Intergroup Relations, 22*, 673–687.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: the stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

Goldenberg, A., Weisz, E., Sweeny, T., Cikara, M., & Gross, J. (2021). The crowd emotion amplification effect. *Psychological Science, 32*, 437–450.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*, 751–753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble encoding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 718–734.

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

Hamilton, D. L. (2007). Understanding the complexities of group perception: broadening the domain. *European Journal of Social Psychology, 37*, 1077–1101.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review, 103*, 336–355.

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, 285*, 20172770.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*, 115–121.

Lenth, R. (2020). *Emmeans package: estimated marginal means, aka least-squares means.* https://CRAN.R-project.org/package=emmeans.

Levy, L. H., & Richter, M. L. (1963). Impressions of groups as a function of the stimulus values of their individual members. *Journal of Abnormal and Social Psychology, 67*, 349–354.

Lundqvist, D., Flykt, A., & Ohman, A. (1998). *The Karolinska directed emotional faces (KDEF)*. Stockholm: Department of Neurosciences Karolinska Hospital.

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision, 18*, 7.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*, 1122–1135.

Maner, J. K., Kenrick, D. T., Becker, D. V., Delton, A. W., Hofer, B., Wilbur, C. J., & Neuberg, S. L. (2003). Sexually selective cognition: Beauty captures the mind of the beholder. *Journal of Personality and Social Psychology, 85*, 1107–1120.

Mattavelli, G., Andrews, T. J., Asghar, A. U. R., Towler, J. R., & Young, A. W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia, 50*, 2205–2211.

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience, 8*, 285–299.

Miller, C. T., & Felicio, D. M. (1990). Person-positivity bias: Are individuals liked better than groups? *Journal of Experimental Social Psychology, 26*, 408–420.

Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2018). Ensemble coding of face identity is not independent of coding of individual identity. *Quarterly Journal of Experimental Psychology, 71*, 1357–1366.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. In , *105. Proceedings of the National Academy of Sciences of the USA* (pp. 11087–11092).

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

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*, 766–785.

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

Core Team, R. (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna: Austria. URL. https://www.R-project.org/.

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia, 48*, 3596–3605.
Sui, J., & Liu, C. H. (2009). Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin Review, 16*, 276–281.
Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545.
Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science, 25*, 230–235.
Whitney, D., & Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105–129.
Wiese, H., Altmann, C. S., & Schweinberger, S. R. (2014). Effects of attractiveness on face memory separated from distinctiveness: Evidence from event-related brain potentials. *Neuropsychologia, 56*, 26–36.
Willis, R. H. (1960). Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology, 60*, 365–373.