**Abstract**

Trustworthiness is a fundamental dimension underlying trait impressions of individual faces, and these impressions predict real-world social consequences. Building on ensemble coding research from the vision sciences, we explored to what extent statistical information about trustworthiness is gleaned from rapid exposure to crowds of faces. We showed that with half-second exposures to sets of eight faces perceivers are sensitive to the set's average level of trustworthiness (Study 1). Moreover, this group-level sensitivity biases individual group member evaluations (Study 2), as well as downstream social behavior related to those evaluations (Study 3), towards the mean of the group. Together, the findings add to a growing body of "people perception" research and show that even high-level social characteristics like personality traits may be spontaneously gleaned from rapid exposure to crowds of faces.

*Keywords:* person perception, face impressions, ensemble coding, trustworthiness, social cognition

## Trustworthiness of Crowds Is Gleaned in Half a Second

The classic maxim states that individuals should not judge a book by its cover, yet social perceivers form immediate, consequential impressions from faces across many social dimensions including gender, race, and age. Beyond these visually evident social dimensions, perceivers also form rapid and reliable impressions of personality traits within 100 ms of visual exposure to a face (Hehman et al., 2017; Todorov et al., 2009; Willis & Todorov, 2006) and often outside conscious awareness (Freeman et al., 2014). Trustworthiness is thought to be a fundamental and functionally adaptive dimension on which people evaluate others (Oosterhof & Todorov, 2008). It accounts for the bulk of variance in face impressions across world regions and portends a variety of downstream social consequences, ranging from criminal sentencing, to electoral success, to career attainment (Jones et al., 2021; Blair et al., 2004; Todorov et al., 2005; Rule & Ambady, 2008).

Vision sciences research on ensemble coding has established that statistical information from an ensemble of visual stimuli, such as average motion or orientation of a group of dots, can be reliably gleaned with brief exposures (Dakin & Watt, 1997; Miller & Sheldon, 1969; Watamaniuk & McKee, 1998). More recently, research has investigated whether perceivers extract similar statistical information from ensembles of faces, including judgments of face ensembles' identity, emotion, gender, race, and eye gaze (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007; Jung et al., 2017; T. D. Sweeny & Whitney, 2014; Goodale et al., 2018). Such perceptual abilities are consequential, impacting perceptions of threat and social belonging (Alt et al., 2019; Goodale et al., 2018). Further investigation has revealed how perceivers extract complex trait information from groups of faces, finding that perceivers are sensitive to group-level attractiveness and dominance (Luo & Zhou, 2018; Phillips et al., 2018).

However, research has yet to investigate whether such a perceptual ability exists for trustworthiness, a core dimension of facial impressions.

Gleaning mean trustworthiness would be valuable in the real-world perception, as trait-related facial appearances of groups often cluster together. Elective social groups, such as fraternities/sororities, sports teams, or Facebook friend groups, exhibit homophily not only in terms of shared interests and beliefs (McPherson et al., 2001), but also at the level of facial appearance (Hehman et al., 2018). For example, members of a fraternity could have faces that overall appear dominant or trustworthy. Thus, extracting a summary trait estimate from groups of faces would provide social information to guide decision-making and downstream behavior towards groups and the individuals within them.

We also explored to what extent the rapid ensemble perception of trustworthiness may have biasing effects on individual faces. The visual context surrounding a face, such as a scene, impacts perception of multiple social dimensions (e.g., race, emotion, trustworthiness), whereby perceptions become more congruent with contextual information (Freeman et al., 2011; Barrett & Kensinger, 2010; Brambilla et al., 2018; Masuda et al., 2008). Recent work has shown that the perceived emotion of a single face in an ensemble is biased towards the mean emotion of the ensemble (Alwis & Haberman, 2020; Corbin et al., 2018). Other related work has found that impressions of attractiveness are biased by the group mean (Walker & Vul, 2014; Carragher et al., 2018, 2020; Ying et al., 2019). Here, we examine to what extent mean trustworthiness perceived across an ensemble impacts perception of individual faces within the ensemble.

Across three studies, we first establish perceivers' sensitivity to mean trustworthiness of groups of faces at brief exposures (Study 1). We then explore how the perceived trustworthiness of a group of faces exerts a contextual impact on individual faces, biasing perceptions (Study 2)

and downstream behavior (Study 3) toward the mean. All data and analysis scripts are available on OSF (https://osf.io/nfgx9/?view_only=90fc68dbcd9a41fc8fb383430eaf4e72). All stimuli are available either on OSF or as permitted by third-party usage agreements.

## Study 1

We first aimed to demonstrate that ensemble perception of facial trustworthiness is cognitively possible. To do this, we borrowed a paradigm from recent work examining ensemble perception of faces (Haberman & Whitney, 2007; Haberman et al., 2009). Participants were presented with ensembles of 8 faces, with the ensemble's average trustworthiness level varying widely. After a 500 ms exposure, participants were asked to judge whether a new individual face (the probe) had a higher or lower level of trustworthiness than the ensemble's mean. To maximize precision and control, we used computer-generated faces that were systematically manipulated on trustworthiness.

### Method

***Participants.*** We recruited 202 participants from Prolific to complete our study (age: $M$ = 38.81, $SD$ = 14.68; gender: 53.63% male, 44.69% female, 1.68% other; race: 72.63% White, 9.50% Black, 8.38% Asian, 2.23% American Indian and 7.26% other).[1] One participant was removed from our dataset based on attention check performance (described below).

***Face ensembles.*** Ensembles were generated from individual faces created in facial morphing software FaceGen (Blanz & Vetter, 1999). The faces were manipulated along the trustworthiness trait dimension (Oosterhof & Todorov, 2008) to create faces that varied continuously on trustworthiness. For each identity, seven levels of trustworthiness were created (-3 SD to +3 SD). Faces were cropped such that each image was centered and focused on each

[1] Due to server error, demographics did not save for 22 participants in Study 1, 1 participant in Study 2, and 2 participants in Study 3.

face. A total of 25 unique identities were created. We randomly selected 8 identities at a time to create our ensembles. Using these 8 identities, we then randomly selected one of the 7 trustworthiness variants for each identity resulting in one ensemble. We repeated this process 100 times, randomly sampling from the total space of possible ensembles. This yielded ensembles that varied on their mean trustworthiness ($M$ = 3.99, $SD$ = 0.78), as well as the variance within ensembles ($M_{SD}$ = 1.895). Faces were arranged into a 2 x 4 grid of faces. All participants saw all 100 ensemble stimuli. In order to control for low-level visual properties of the stimuli, all images were passed through the SHINEToolbox to normalize low-level visual features (Willenbockel et al., 2010).

To validate our trustworthiness manipulation, we gathered trustworthiness ratings on the individual face stimuli from 49 independent raters via Prolific ($M_{age}$=38.92, SD=16.79; 65.31% female, 34.69% male; 63.27% White, 16.33% Black, 12.24% Asian, 8.16% other). Raters completed a simple, untimed task rating the trustworthiness of each face (1 = Not at all Trustworthy, 7 = Very trustworthy). Raters also completed attention checks ("Press *x*"), and raters that failed more than 30% of attention checks were excluded. No raters met this threshold. Expectedly, interrater agreement was high (ICC=0.920), and trustworthiness ratings were strongly correlated with the morph levels of our stimuli, r(173)=0.769, 95% CI [0.7, 0.82], p<0.001. Given the known correlation between trustworthiness and attractiveness judgements (Oosterhof & Todorov, 2008), we also collected attractiveness judgements from another set of independent raters to control for attractiveness and isolate the effect of trustworthiness. A total of 53 raters from Prolific participated ($M_{age}$=32.44, SD=12.49; 59.62% male, 38.46% female, 1.92% other; 67.31% White, 11.54% Black, 9.62% Asian, 11.54% other) to complete an analogous untimed ratings task for attractiveness. On each trial, participants were shown a single

face and asked to rate the attractiveness of the face on a scale of 1 ("Not at all attractive") to 7 ("Very attractive"). One rater failed more than 30% of attention checks and was excluded. Raters demonstrated high agreement in attractiveness ratings, ICC=0.843. As expected, trustworthiness and attractiveness ratings were highly correlated, $r(173)=0.717$, 95% CI [0.64, 0.78], $p<0.001$.

***Probe task.*** On each trial, participants were shown a fixation cross for 2000 ms, followed by an ensemble of 8 faces for 500 ms. After viewing the ensemble, participants were shown a randomly selected probe face. Probe faces were drawn randomly from the full set of computer-generated face stimuli, although they could not be of any identity present in an ensemble and could not equal the mean of the ensemble. Thus, the probe would be randomly higher or lower than the mean trustworthiness of the depicted ensemble and vary on its distance from the mean. Participants were given an unlimited amount of time to decide whether the probe face was more or less trustworthy than the mean of the ensemble. Participants completed attention checks ("Please press higher/lower.") that were randomly interspersed with experimental trials. Participants who failed more than 30% of attention checks were excluded.

**Results and Discussion**

Given that the randomly selected probe's trustworthiness level had a 50/50 chance of being higher or lower than the ensemble mean, we initially tested whether participants were more accurate in inferring the ensemble mean than chance. Calculating the proportion of correct responses for each participant (e.g., choosing "higher" when the probe was higher than the ensemble mean) revealed a mean accuracy of 67.9% (SD = 9.8%). A one-sample *t*-test confirmed this was significantly better than chance (50%), $t(200)=25.919$, 95% CI [66.6%, 69.3%], $p<0.001$. To control for potential response bias, we provided converging evidence using a signal detection analysis (Green & Swets, 1966). Arbitrarily assigning a "higher" probe-

ensemble relationship as signal, we calculated the number of hits, false alarms, misses, and correct rejections for each participant. From these counts, we calculated *d′* scores, providing an estimate of discriminability corrected for response bias. Participants showed strong discriminability (*d′*) ($M$=1.002, $SD$= 0.573) that was significantly higher than zero, one-sample $t(200)=24.814$, 95% CI [0.923, 1.082], $p<0.001$.

We further analyzed the data using a multi-level regression model to demonstrate that the effects cannot be explained by attractiveness, which tends to co-vary with trustworthiness. For each trial, we calculated the absolute difference in trustworthiness between the ensemble and the probe face, as well as the absolute difference in attractiveness between the ensemble and the probe face. If trustworthiness ensemble perception is genuinely driving correct responses, then we would expect accuracy to increase as the trustworthiness between the ensemble and probe become more different from one another (as the trial is therefore easier to discern), even though the mean of the group is never presented and must be extracted by the perceiver across faces. Furthermore, this effect should occur above and beyond the analogous difference in attractiveness between the ensemble and probe. We ran a generalized linear mixed model, predicting the likelihood of a correct response (0=*incorrect,* 1=*correct*) from the absolute difference between the ensemble and probe for trustworthiness, as well as the analogous difference for attractiveness (formula: correct ~ 1 + *absolute trustworthiness difference* + *absolute attractiveness difference*). We used the *lme4* package in R for this model and all subsequent mixed-effects models, using the *lmer* function and *glmer* function with a binominal link function for continuous and binary outcomes respectively (Bates et al., 2015). All predictors were centered prior to analysis for all mixed-effects models in Studies 1-3. To account for random variability in the specific ensemble stimuli used (all participants viewed the same set of ensembles), we also included a random effect for specific

ensemble stimuli. A model with random slopes for each participant and ensemble stimulus failed to converge; the reported model includes only random intercepts for participants and ensemble stimuli. Indeed, as the distance in trustworthiness increased, the likelihood of a correct response strongly increased as well, even when statistically accounting for attractiveness distance (log-odds=0.348, *SE*=0.015, 95% CI [0.319, 0.376], *z*=23.847, p<0.00001). Attractiveness distance was also a significant predictor of a correct response, but with a considerably smaller effect size (log-odds=0.323, *SE*=0.061, 95% CI [0.204, 0.442], *z*=5.307, p<0.001). This is not surprising given that these traits co-vary and multiple traits may be utilized for judgment.

To provide evidence in support of genuine ensemble perception of mean trustworthiness, as opposed to merely attending to one face at random, we conducted two additional simulations that modeled what the trustworthiness distance effect would be if it were to have arisen by participants selecting a face at random. Both analyses strongly suggest that the trustworthiness distance effect observed arose due to extracting the mean across the ensemble rather than choosing a single face at random to infer trustworthiness. See Supplementary Material for details.

Taken together, Study 1 shows that participants genuinely extract trustworthiness from ensembles of faces. Participants show perceptual sensitivity to the mean of the group, even though the mean is never presented. The effects could not be explained by participants randomly attending to one face at a time, or by co-varying traits such as attractiveness.

## Study 2

Having demonstrated that perceivers are sensitive to group trustworthiness, we now investigate how ensemble encoding impacts the perception of individual constituent faces. Prior studies have shown that the perceived emotional expression of a single face in a group is biased towards the mean group expression (Alwis & Haberman, 2020; Corbin, et al., 2018; Masuda et

al., 2008) We tested whether a group's level of mean trustworthiness exerts a contextual impact on the perceived trustworthiness of individual faces. We also aimed to strengthen our previous findings of ensemble perception of facial trustworthiness. By comparing ratings of single faces with those of groups of faces, here we provide converging evidence about whether participants are genuinely extracting information from the group of faces rather than a single constituent face, as well as generalizing this phenomenon to real faces.

**Method**

***Participants.*** We recruited 200 participants from Mechanical Turk. After removing participants who failed more than 30% of attention checks, 195 participants remained (age: *M*=42.68, *SD*=13.82; gender: 55.15% female, 43.81% male, 1.03% other; race: 81.96% White, 8.25% Asian, 6.70% Black, 3.11% Other)[1].

***Face Ensembles.*** Ensembles comprised 8 faces drawn from the combined set of all White male faces in the Chicago and Radboud Face Databases, resulting in 233 faces (Langner et al., 2010; Ma et al., 2015). White male faces were used to avoid attentional confounds related to target gender and race. Faces from the two databases were cropped to have similar portraiture and normalized on luminance and contrast using SHINE Toolbox (Willenbockel et al., 2010). Independent raters (N=51, $M_{age}$=34.67, SD=12.46; 70.59% female, 29.41% male; 66.67% White, 17.65% Black, 1.96% Asian, and 12.72% other) were recruited from Prolific to provide untimed trustworthiness judgements of all faces from the combined stimulus set on a scale from 1 ("Not at all trustworthy") to 7 ("Very trustworthy"). Raters also completed attention checks. No raters failed more than 30% of attention checks. Raters demonstrated high agreement, ICC=0.846. In each ensemble, 7/8 of the faces had a consistent level of very low trustworthiness or very high trustworthiness, and the remaining face (the target) was always the opposite extreme. For

instance, a given ensemble might contain 7 highly untrustworthy faces and the target: a highly trustworthy face (or vice-versa). We created 25 ensembles with a low trustworthiness mean and a high trustworthiness target face and 25 ensembles with a high trustworthiness mean and a low trustworthiness target face, resulting in a total of 50 unique ensemble stimuli. The location of the target face was randomized. Faces were again passed through the SHINEToolbox to normalize low-level visual properties (Willenbockel et al., 2010).

***Ratings Tasks.*** Participants completed three tasks in randomized order: rating individual target faces in isolation (*single ratings*), rating ensembles (*ensemble ratings*), and rating individual target faces highlighted with a border within their ensembles (*highlighted ratings*). When rating ensembles, as in the previous studies, participants were instructed to rate the trustworthiness of the group as a whole. When rating highlighted targets, participants were instructed to rate the trustworthiness of the highlighted face only and to ignore the other faces. For single ratings, only the faces that served as targets in the ensemble stimuli were rated. As in the previous studies, for all tasks stimuli were presented for 500 ms followed by a 200 ms backward mask, after which participants made a rating using a Likert scale of 1 (“Not at all trustworthy”) to 7 (“Very trustworthy”). Participants judged all 50 ensemble stimuli in the single ratings, ensemble ratings, and highlighted ratings tasks, for a total of 150 trials. All participants saw the same ensembles.

**Results and Discussion**

Because ensemble perception revolves around integrating information across multiple targets presented simultaneously, greater precision can be gained by averaging over multiple datapoints (i.e., faces). Thus, if perceivers are indeed encoding average information from the ensemble, they should demonstrate greater sensitivity to trustworthiness in groups of faces

compared to individual faces (Elias et al., 2017; Haberman et al., 2009; Sweeny et al., 2013; Sweeny & Whitney, 2014). To provide converging evidence that ensemble perception is occurring, for each subject, an error score was calculated for each single face as the difference between their rating and the rating provided by independent raters. For each ensemble, error scores were defined as the difference between participant responses and the numerical average of trustworthiness of constituent individual faces as provided by independent raters. This yielded two distributions of error scores for each subject. We then calculated the SD of each distribution separately for each subject as a measure of perceptual sensitivity. Since ensemble perception results in greater accuracy via averaging, the SD of ensemble error scores should be smaller on average than that of single faces if ensemble perception is occurring. If perceivers are simply attending to a single face when looking at an ensemble, then there should be no meaningful difference in sensitivity between single face and ensemble trials. Alternatively, if perceivers are indeed incorporating information from multiple faces, evaluations of ensemble trials should be more precise due to the efficiency of ensemble perception, as perceivers average over multiple faces that are individually noisy signals of the mean (Elias et al., 2017; Haberman et al., 2009; Sweeny et al., 2013; Sweeny & Whitney, 2014).

A paired-samples t-test found that SDs were on average smaller for ensemble trials than for single face trials ($M_{diff}$=0.100, $SE$=0.017, 95% CI [0.066, 0.134], $t(194)$=5.867, $p$<0.0001), indicating that perceivers were more sensitive to trustworthiness in groups of faces than individual faces (Figure 1).

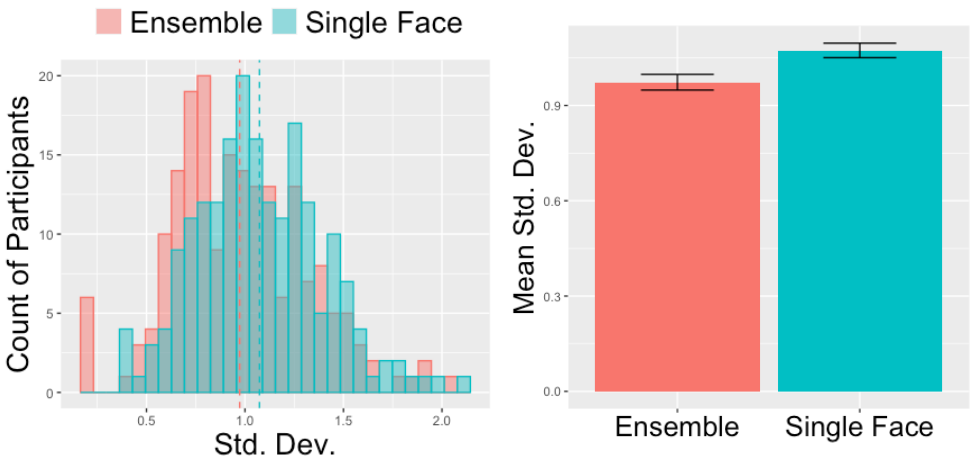


**Figure 1. Error score analysis in Study 2.** Distribution of SDs of error scores for ensemble and single face trials with the mean plotted as a dotted line (left). Average SD of error scores for ensemble and single face trials, plotted with bars indicating the standard error of the mean.

To assess the impact of ensemble context on the perceived trustworthiness of individual faces, we fitted a linear mixed-effects model to predict highlighted ratings from single ratings, ensemble ratings, and their interaction (formula: *highlighted rating ~ single ratings + ensemble ratings + single ratings : ensemble ratings*) with random effects for participant and ensemble stimulus. The model initially failed to converge; the reported model includes random slopes and intercepts for participants and random intercepts for ensemble stimuli. As expected, there was a significant effect of single rating, such that targets judged to be more trustworthy when presented in isolation were also judged to be more trustworthy when highlighted within an ensemble, $B$=0.278, $SE$=0.015, 95% CI [0.249, 0.307], $t$(211.772)=18.678, $p$<0.0001. This shows that participants were indeed sensitive to a target's facial features when judging the trustworthiness of highlighted targets within ensembles. More critically, there was a significant effect of ensemble trustworthiness, such that the perceived trustworthiness of a highlighted target increased when the ensemble's average trustworthiness was higher, $B$=0.048, $SE$ =0.012, 95%

CI [0.023, 0.072], $t(196.319)=3.814$, $p<0.001$ (Figure 2). The interaction did not reach significance, $B=0.009$, $SE=0.007$, 95% CI [-0.003, 0.023], $t(3430.208)=1.482$, $p=0.138$.

These results indicate that trustworthiness impressions of single faces are biased towards the mean of its group, suggesting that perceivers are not only sensitive to group-level trustworthiness when explicitly asked to holistically evaluate a group, but also when asked to evaluate an individual group member.

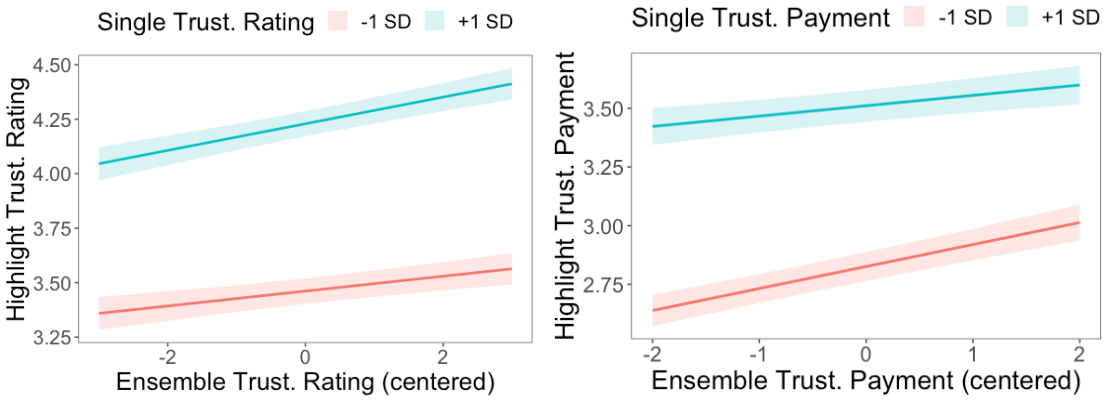


**Figure 2. Results of Studies 2 and 3.** Model predicted values of highlighted trustworthiness ratings (Study 2, left) or trust payments (Study 3, right) are plotted as a function of average ensemble trust ratings or trust payments. Shaded region indicates standard error of the model fit. Highlighted trust ratings and payments increase with the average rating or payment to the group, both for faces that receive high and low trust ratings and payments in isolation.

## Study 3

In Study 2, we established that impressions of individual faces are impacted by average trait-related information of an ensemble. Here we assess to what extent this group-level biasing affects downstream trust-related behavior.

### Method

***Participants.*** We recruited 207 participants from Mechanical Turk. After removing participants who failed more than 30% of attention checks or exited the study early, 163

participants remained (age: *M*=37.45, *SD*=10.37; gender: 50.31% male, 48.45% female, 0.62% decline; 0.62% other; race: 78.26% White, 8.07% Black, 6.21% Asian, 7.45% Other)[1].

***Stimuli.*** Ensembles were identical to those used in Study 2.

***Trust Games.*** We adapted a trust game paradigm used in previous studies to capture a participant's trust interactions with target groups (van't Wout & Sanfey, 2008; Berg et al., 1995). The procedure followed the identical three-task structure as Study 2, except participants engaged in trust payment decisions for ensembles as well as for individual targets, rather than making trustworthiness evaluations. Perceivers completed three trust games in randomized order. Participants were instructed that they had been randomly assigned to be an investor. Participants were told that they would sometimes be investing in either individual business associates (*single trust decisions*), groups of business associates (*ensemble trust decisions*), or an individual associate surrounded by other associates (*highlighted trust decisions*). These tasks parallel the three ratings tasks of Study 2. Participants were instructed that they would be given $1.00 on each trial. Participants were told they could choose any amount between $0.00 and $1.00 in increments of $0.25 to invest. The investment by the participant would be tripled, and the group of business associates, individual associates, or individual associates surrounded by other associates would decide how much of the money to return. Participants were not given trial-by-trial feedback on the amount of money returned. Participants were presented with the ensembles, individual faces, or individual faces surrounded by the other members of the ensemble one at a time in randomized order for 500 ms. Presentations were followed by a 200 ms backward mask to prevent afterimage processing. Following presentation, participants were given an unlimited amount of time to decide how much money they wished to invest. Participants completed 50

trials in the single ratings, ensemble ratings, and highlighted ratings tasks, for a total of 150 trials. All participants saw the same ensembles.

**Results and Discussion**

In order to provide evidence for genuine ensemble perception, we conducted the same error score analysis described in Study 2. However, we first mapped trustworthiness evaluations into trust payment space by rescaling them to scores between 1 and 5. A paired-samples t-test again found that SDs were on average smaller for ensemble trials than for single face trials, $M_{diff}$=0.129, *SE*=0.021, 95% CI [0.089, 0.169], *t*(162)=6.291, *p*<0.0001.

Similar to Study 2, we fitted a linear mixed-effects model to predict highlighted trust decisions from single trust decisions, ensemble trust decisions, and their interaction (formula: *highlighted trust decisions ~ single trust decisions + ensemble trust decisions + single trust decisions : ensemble trust decisions*) with random effects for participant and a random intercept for ensemble stimuli. The pattern of results replicated that of Study 2. Trust payments to highlighted targets increased for those targets entrusted with more money when presented in isolation, *B*=0.255, *SE*=0.021, 95% CI [0.213, 0.296], *t*(138.695)=12.000, *p*<0.0001. More critically, we again found that trust payments to a highlighted associate increased when participants entrusted more money to the group as a whole, *B*=0.069, *SE*=0.015, 95% CI [0.039, 0.099], *t*(128.448)=4.470, *p*<0.0001 (Figure 2). The interaction was marginally significant, *B*=-0.018, *SE*=0.009, 95% CI [-0.037, 0.0001], *t*(1912.289)=-1.943, *p*=0.052. We decomposed this interaction at ±1SD of single trust decisions. The effect of ensemble trust payments on trust payments to highlighted targets held the same pattern and was positive and significant at both levels of trust payments in isolation, but it was relatively weaker at higher levels (*b*=0.044, *SE*=0.021, 95% CI [0.002, 0.086], Z=2.062, p=0.039) and relatively stronger at lower levels

($b$=0.094, $SE$=0.019, 95% CI [0.058, 0.130], $Z$=5.071, $p<0.0001$). Thus, perceivers are not only sensitive to a group-level estimate of trustworthiness even when asked to attend to individual group members, but this sensitivity biases trust-related behavior as well.

## General Discussion

Across three studies, we document the perceptual ability to perceive a group-level estimate of trustworthiness that impacts behavior and biases impressions of individual faces. Specifically, we showed that perceivers were sensitive to group-level trustworthiness at half-second exposures (Study 1). We further demonstrated that impressions of the trustworthiness of an individual face embedded in a group of faces and the corresponding trust-related behavior that followed it were biased by the group's average trustworthiness (Studies 2-3).

Together, this work builds on a growing body of "people perception" research investigating how perceivers visually construe groups of individuals. This is the first work, to our knowledge, to examine the perception of trustworthiness in groups of faces. As discussed earlier, trustworthiness is arguably the most important trait dimension and accounts for the bulk of variance in face impressions (Oosterhof & Todorov, 2008). Trait impressions of individual faces follow a well-known correlation structure, with trustworthiness and dominance argued to be fundamental dimensions through which all other dimensions (e.g., competence, extraversion) are inferred (Oosterhof & Todorov, 2008). More recent research suggests that this structure may be more variable and depend on experience and learning and the social group memberships to which targets and perceivers belong (Hehman et al., 2017; Xie et al., 2019). It is an open question whether a two-dimensional structure or more dynamic trait structure emerges in trait evaluations of groups of faces. Ensemble encoding may result in convergence or disparities between the dimensional structure of "people" and "person" perception.

Our finding that group-level encoding biases individual face impressions adds to a growing list of biases in social ensemble perception (Goldenberg et al., 2021; Ying et al., 2019; Alwis & Haberman, 2020; Corbin et al., 2018) and is consistent with top-down contextual effects on face perception (Brambilla et al., 2018; Masuda et al., 2008; Freeman et al., 2011; Stolier et al., 2018). Extant theoretical models of person perception incorporate bottom up (e.g., facial features) and top-down (e.g., social-conceptual knowledge, goals, affective state) factors to determine facial impressions (Freeman et al., 2020). Yet these models have not incorporated how ensemble encoding may impact individual perceptions and interact with top-down information. Recent accounts of people perception have begun to explore such interactions, although more research is needed to better understand the role social-conceptual knowledge plays in people perception (Phillips, et al., 2014; Alt & Phillips, 2022).

Representations generated via ensemble coding efficiently abstract across multiple individual faces (Cohen et al., 2016; Whitney & Yamanashi Leib, 2018). Our results support prior work suggesting that sensitivity to a group-level characteristic is dependent on some averaged percept rather than sampling across multiple constituent faces, consistent with work showing perceivers are more likely to report having seen a morphed facial composite of a set of faces (which in fact was never presented) than any actual set member (de Fockert & Wolfenstein, 2009; Neumann et al., 2013). Co-occurring social cues (e.g., race, gender, emotion) readily interact to shape perceptions of single faces (Freeman et al., 2020; Hugenberg & Bodenhausen, 2004; Johnson et al., 2012), and emerging research has begun to demonstrate how social cues distributed across an ensemble aggregate to form an averaged percept and influence group evaluations (Lamer et al., 2018). Future work should investigate more directly how information

is pooled across multiple targets and how social-conceptual knowledge might impact such pooling.

The present work has several limitations. Our stimuli were exclusively White male faces in order to avoid potential confounds related to individual differences in gender and racial bias. For instance, target race and gender influence facial trustworthiness evaluations, although significant variance in these impressions are still linked to features that cue trustworthiness independent of race or gender (Hehman et al., 2019; Xie et al., 2019). Future work should establish the generality of the effects and test potential interactions with race and gender. Finally, perceivers in the world at large do not interact with grids of faces on sanitized backgrounds; using naturally occurring groups of faces should be a priority in work moving forwards.

In summary, our research illustrates how social perceivers extract a group-level estimate of trustworthiness from a group of faces that empowers evaluations and behaviors. Such group-level estimates can even bias evaluations of individual faces. These findings add to a growing body of "people perception" research and show that the core dimension of trait impressions, trustworthiness, is gleaned from rapid exposure to a crowd of faces.

## References

Alt, N. P., & Phillips, L. T. (2022). Person Perception, Meet People Perception: Exploring the Social Vision of Groups. *Perspectives on Psychological Science, 17*(3), 768-787. https://doi.org/10.1177/17456916211017858

Alt, N. P., Goodale, B., Lick, D. J., & Johnson, K. L. (2019). Threat in the Company of Men: Ensemble Perception and Threat Evaluations of Groups Varying in Sex Ratio. *Social Psychological and Personality Science*, *10*(2), 152–159. https://doi.org/10.1177/1948550617731498

Alwis, Y., & Haberman, J. M. (2020). Emotional judgments of scenes are influenced by unintentional averaging. Cognitive Research: Principles and Implications, 5(1), 28.

Barrett, L. F., & Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. Psychological Science, 21(4), 595–599.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. Games and Economic Behavior, 10(1), 122–142.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques.

Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology*, *78*, 34–42.

Carragher, D. J., Lawrence, B. J., Thomas, N., & Nicholls, M. E. R. (2018). Visuospatial

asymmetries do not modulate the cheerleader effect. Scientific Reports, 8, Article 2548. https:// doi.org/10.1038/s41598-018-20784-5

Carragher, D. J., Thomas, N. A., Gwinn, S. O., & Nicholls, M. E. R. (2020). The cheerleader effect is robust to experimental manipulations of presentation time. Journal of Cognitive Psychology, 32(5–6), 553–561. https://doi.org/10.1080/2 0445911.2020.1776718

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, *20*(5), 324–335.

Corbin, J. C., Crawford, L. E., Zwaan, R., & O'Connor, A. (2018). Biased by the group: Memory for an emotional expression biases towards the ensemble. Collabra: Psychology, 4(1). https://online.ucpress.edu/collabra/article-abstract/4/1/33/112995

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192.

de Fockert, J., & Wolfenstein, C. (2009). Short article: Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, *62*(9), 1716–1722.

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble Perception of Dynamic Emotional Groups. Psychological Science, 28(2), 193–203.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the Part: Social Status Cues Shape Race Perception. *PLoS ONE*, *6* (9), e25107.

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 61, pp. 237–287). Academic Press.

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. The Journal of

Neuroscience: The Official Journal of the Society for Neuroscience, 34(32), 10573–10581.

Goldenberg, A., Weisz, E., Sweeny, T. D., Cikara, M., & Gross, J. J. (2021). The Crowd-Emotion-Amplification Effect. *Psychological Science*, *32*(3), 437–450.

Goodale, B. M., Alt, N. P., Lick, D. J., & Johnson, K. L. (2018). Groups at a glance: Perceivers infer social belonging in a group based on perceptual summaries of sex ratio. *Journal of Experimental Psychology. General*, *147*(11), 1660–1676.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969-2012). New York: Wiley.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology: CB*, *17*(17), R751–R753.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), 1.1–13.

Hehman, E., Flake, J. K., & Freeman, J. B. (2018). The Faces of Group Members Share Physical Resemblance. *Personality & Social Psychology Bulletin*, *44*(1), 3–15.

Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass*, *13*(2), e12431.

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342–345.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: how covarying

phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Adil Saribay, S., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169.

Jung, W., Bülthoff, I., & Armann, R. G. M. (2017). The contribution of foveal and peripheral visual information to ensemble representation of face race. *Journal of Vision*, *17*(13), 11.

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.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, *24*(8), 1377–1388.

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision*, *18*(8), 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*(4), 1122–1135.

Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., & Van de Veerdonk, E. (2008). Placing the face in context: cultural differences in the perception of facial emotion. *Journal of personality and social psychology*, *94*(3), 365.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social

networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Miller, A. L., & Sheldon, R. (1969). Magnitude estimation of average length and average inclination. *Journal of Experimental Psychology*, *81*(1), 16–21.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, *128*(1), 56–63.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092.

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.

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.

Rule, N. O., & Ambady, N. (2008). The face of success: inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*(2), 109–111.

Stolier, R. M., Hehman, E. A., & Freeman, J. B. (2018).A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22* (3), 197-200.

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–1913. https://doi.org/10.1177/0956797614544510

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, 329–337. https://doi.org/10.1037/a0028712

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition*, *27*(6), 813–833.

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. Cognition, 108(3), 796–803.

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

Watamaniuk, S. N., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics*, *60*(2), 191–200.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105–129.

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

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, *17*(7), 592-598.

Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology*, *117*(2), 364–385.

Ying, H., Burns, E., Lin, X., & Xu, H. (2019). Ensemble statistics shape face adaptation and the

cheerleader effect. *Journal of Experimental Psychology. General*, *148*(3), 421–436.