Cognition 239 (2023) 105540

Contents lists available at ScienceDirect

Cognition

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

# Three's a crowd: Fast ensemble perception of first impressions of trustworthiness

Fiammetta Marini [a,*], Clare A.M. Sutherland [a,b], Bārbala Ostrovska [a], Mauro Manassi [a]

[a] School of Psychology, University of Aberdeen, King's College, Aberdeen, UK
[b] School of Psychological Science, University of Western Australia, Crawley, Western Australia, Australia

ARTICLE INFO

*Keywords:*
Trustworthiness
Ensemble perception
Summary statistics

ABSTRACT

Trustworthiness impressions are fundamental social judgements with far-reaching consequences in many aspects of society, including criminal justice, leadership selection and partner preferences. Thus far, most research has focused on facial characteristics that make a face individually appear more or less trustworthy. However, in everyday life, faces are not always perceived in isolation but are often encountered in crowds. It has been proposed that we deal with the large amount of facial information in a group by extracting summary statistics of the crowd, a phenomenon called ensemble perception. Prior research showed that ensemble perception occurs for various facial features, such as emotional expression, facial identity, and attractiveness. Here, we investigated whether observers can integrate the level of trustworthiness from multiple faces to extract an average impression of the crowd. Across four studies, participants were presented with crowds of faces and were asked to report their average level of trustworthiness with an adjustment (Experiment 1) and a rating task (Experiments 2 and 3). Participants were able to extract an ensemble perception of trustworthiness impressions from multiple faces. Moreover, observers were able to form a summary statistic of trustworthiness impressions from a group of faces as quickly as 250 ms (Experiment 4). Taken together, these results demonstrate that ensemble perception can occur at the level of impressions of trustworthiness. Thus, these critical social judgements not only occur for individual faces but are also integrated into a unique ensemble impression of crowds. Our findings contribute to the development of a more ecological approach to the study of trust impressions, since they provide an understanding of trustworthiness judgements not only on an individual level, but on a much broader social group level. Furthermore, our results drive forward new theory because they demonstrate for the first time that ensemble representations cover a broad range of phenomena than previously recognized, including complex high-level facial trait judgements such as trustworthiness impressions.

## 1. Introduction

Trustworthiness impressions are critical social judgements at the core of facial evaluation (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Twele & Mondloch, 2022). Judgements of trust from faces are rapidly formed; 33 ms exposure is more than sufficient to form a trustworthiness impression (Todorov et al., 2009). Although there is little evidence that these impressions are veridical, people often agree on their impressions (Foo et al., 2021; Korva et al., 2013; Todorov et al., 2015; Wilson & Rule, 2015), and they spontaneously and strongly guide our social behaviour with deep repercussions in many aspects of society (Olivola et al., 2014). For example, facial features that convey high trustworthiness are associated with judgements of low criminality and a higher likelihood of receiving lower sentencing outcomes in the courts (Flowe, 2012; Wilson & Rule, 2015), with achieving leadership positions and receiving promotions (Linke et al., 2016; Rule & Ambady, 2008), and with higher romantic attraction (Olivola et al., 2014; South Palomares & Young, 2017; Valentine et al., 2020).

So far, research in social cognition has mainly focused on facial characteristics that make an individual face appear more or less trustworthy (for reviews, see Olivola et al., 2014; Oosterhof & Todorov, 2008; Sutherland & Young, 2022; Todorov et al., 2015). The vast majority of previous studies on trustworthiness impressions have used isolated and de-contextualized face images. However, in our daily experience, faces are rarely perceived in isolation. Instead, we often interact with groups of people, and faces are constantly immersed within

* Corresponding author.
*E-mail address:* f.marini.21@abdn.ac.uk (F. Marini).

https://doi.org/10.1016/j.cognition.2023.105540
Received 24 October 2022; Received in revised form 7 May 2023; Accepted 27 June 2023
Available online 19 July 2023

spatial, situational, and social contexts. At present, whether the context in which faces are embedded influences trustworthiness impressions remains an open theoretical question. A deeper understanding of this aspect represents an important step towards a more ecological approach to the study of trustworthiness judgements.

In contrast to this work on impression formation, many recent studies in vision science have investigated group perception, and these studies demonstrate that humans extract important social information from crowds of faces (Haberman and Whitney, 2007). According to this line of research, when we encounter groups of faces, our visual system is exposed to large amounts of facial information in a short period of time, but our limited working memory abilities (Luck & Vogel, 1997) prevent us from analysing each individual face in the scene. For this reason, it has been proposed that the visual system extracts summary statistics of similar objects, such as the average emotional expression in a group of people (Bai et al., 2015; Haberman and Whitney, 2007; Haberman & Whitney, 2009; Li et al., 2016). This phenomenon is referred to as ensemble perception (Ariely, 2001; Whitney & Yamanashi Leib, 2018). Essentially, instead of analysing each face singularly, the redundancies of the scene are taken advantage of, and multiple faces are integrated into a summary statistical representation (Alvarez, 2011; Ariely, 2001). For example, when we enter a social gathering, whether in person or in an online meeting, searching through the crowd to analyse the individual expressions of each specific face would be relatively slow. Instead, we extract a summary statistic of the emotional expression of the crowd, and thus very easily and with no effort we understand in very few seconds the mood of the group and we act accordingly. Ensemble perception has been proposed to occur immediately when we encounter groups of objects (Whitney & Yamanashi Leib, 2018). In line with this view, it has been shown that very brief exposure to a set of stimuli is enough to form summary statistics, much less than the time required to analyse each object of the scene singularly (Chong & Treisman, 2005), and that our visual system is able to form an ensemble representation without discriminating and recognizing every single item of a set (Fischer & Whitney, 2011; Haberman et al., 2009; Haberman & Whitney, 2011; Manassi et al., 2017; Sweeny et al., 2015; Yamanashi Leib et al., 2020).

Thus far, many studies found that observers can accurately extract summary statistics of facial features from groups of faces, such as the average emotional expression (Haberman et al., 2009; Haberman & Whitney, 2007; Li et al., 2016), sex ratio (Alt et al., 2019; Goodale et al., 2018; Haberman & Whitney, 2007), facial identity (Bai et al., 2015; de Fockert & Wolfenstein, 2009; Leib et al., 2014), and attractiveness (Carragher et al., 2021; Luo & Zhou, 2018). Nevertheless, to the best of our knowledge, no study has focused on summary statistics of trustworthiness impressions. This research gap is surprising, given that these kinds of impressions have been shown to be at the core of social face evaluation (Lin et al., 2021; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Twele & Mondloch, 2022) and to mediate our social interactions on a daily basis (Olivola et al., 2014). Presumably, the effect of trustworthiness impressions is important not only when we interact with single individuals, but also with groups of people. Moreover, given that ensemble perception is found for other facial judgements, it may seem likely to also apply to impression formation. However, trustworthiness impressions are particularly highly complex face judgements. Indeed, they are not only extracted from a multitude of bottom-up facial features such as eyebrow height and mouth curvature (Oosterhof & Todorov, 2008; Vernon et al., 2014) but are also modulated by our personal exposure to faces across life (Sutherland et al., 2020) and by our personal conceptual beliefs regarding how certain personality traits convey trustworthiness (Stolier et al., 2018). Moreover, trustworthiness impressions, although related to other facial judgements such as identity, emotion, and attractiveness, are also distinct (Oosterhof & Todorov, 2008, 2009). Thus, it remains an open question whether ensemble representations can be formed for such complex and high-level face social judgements.

Here, we propose that ensemble perception, an important contextual effect demonstrated to be present in various aspects of face perception, can occur for trustworthiness impressions. Forming an ensemble perception of trustworthiness impressions from multiple faces could have important implications. Most obviously, computing the average level of trustworthiness of a group of people could influence how we perceive a group and consequently how we act. In crowd navigation, for example, behavioural decisions are suggested to be influenced by summary statistical information of the group, since understanding the average gaze direction and mean head rotation of the crowd allows one to quickly determine where to walk next to, as well as the speed of walking (Sweeny & Whitney, 2014). Similarly, trustworthiness impressions of a group might guide our social behaviour. For instance, if we approach a group of people on the street where the majority of the members appear to have a very trustworthy-looking face, this crowd may appear less threatening (and thus approached) compared to approaching a group of people where the majority of members are untrustworthy-looking (and thus one may wish to change path). The ability to form rapid summary statistics from a group of faces is particularly useful given that it permits to quickly extract social information (e.g., the threat or approachability of the group as a whole), allowing for more efficient and faster social decisions than focusing on each individual member of the group separately.

Computing the average level of trustworthiness of a crowd of people could also influence our judgement of singular members of the group, in turn influencing our behaviour and social interactions towards that individual. In this regard, recent findings suggest that the social context in which faces are embedded can play an important role in how we perceive certain individual facial characteristics (Carragher et al., 2021). A phenomenon related to this aspect is the “Cheerleader effect”: although the attractiveness of an individual is signalled by morphological facial features, such as averageness, symmetry, and sexual dimorphism (Baudouin & Tiberghien, 2004; Little et al., 2011), a face seen in group results to be perceived as more attractive compared to when seen alone (Carragher et al., 2021; Walker & Vul, 2014; Ying et al., 2019). The cheerleader effect has been suggested to be a consequence of ensemble coding (Walker & Vul, 2014). Indeed, when observers encounter a group of faces, they might form an ensemble average of the group which results to be highly attractive given that average faces are perceived as particularly attractive (Baudouin & Tiberghien, 2004). Then, individual faces from the group are remembered as more similar to the highly attractive ensemble average face and consequently their attractiveness is enhanced (Ying et al., 2019). In this case, the social context plays an important role in how we perceive facial characteristics. Similarly, the ensemble perception of trustworthiness impressions might influence how much we perceive a member of a group as more or less trustworthy (Carragher et al., 2021). In police line-ups, for example, where multiple suspects' faces are presented to individuate the criminal, the perceived level of trustworthiness of a single face may be affected by the presence of other surrounding faces with different trustworthiness appearances, and this might have an effect in eyewitness identifications accuracy. To this regard, Carragher et al., 2021 investigated whether the cheerleader effect occurs also for judgements of facial trustworthiness. Their findings revealed that untrustworthy individual faces appear more trustworthy when seen in groups of both trustworthy or untrustworthy faces, whereas there was no change in trust appearance for singular faces that were already highly trustworthy individually. However, it is still not clear whether ensemble perception, a phenomenon proposed as an underlying mechanism involved in the cheerleader effect, can occur for trustworthiness impressions. For this reason, the investigation of ensemble perception of trustworthiness impressions is extremely timely in order to more deeply understand the effects of the social context on trust judgements of individuals within a group.

The current study therefore aimed to investigate whether observers extract an ensemble percept of trustworthiness impressions from multiple faces. Throughout four experiments, we investigated ensemble

coding through an experimental paradigm typical of ensemble perception studies (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Yamanashi Leib et al., 2020). By presenting groups of faces varying in their level of trustworthiness in different set size conditions, we measured how much participants integrated trustworthiness information from the crowd in an ensemble percept. Specifically, we ensured that participants were correctly integrating the trustworthiness of the whole set of faces in an ensemble percept instead of randomly reporting the trustworthiness of one face from the crowd (i.e., subsampling). We were able to rule out subsampling because the different set size conditions in the experimental paradigm allowed us to simulate participants' performance in the case in which they had randomly subsampled just a random face from the crowd, instead of integrating the whole set to respond.

Our results demonstrated that ensemble perception occurs at the level of trustworthiness impressions, bridging the gap between social cognition and visual science literatures which have hitherto remained distinct. These findings provide a novel understanding of the role that the social context plays in trustworthiness impressions and extend the theories of trustworthiness impression formation, previously confined to an individual level, to a broader and more ecological group level.

## 2. Experiment 1

The general goal across all experiments was to investigate whether participants were able to extract an ensemble percept of trustworthiness impressions from multiple faces. For this purpose, in Experiment 1 we displayed sets of multiple computer-generated faces varying in their level of trustworthiness and we utilized an adjustment task to measure participants' ability to compute the mean trustworthiness of a previously seen group of faces. In order to understand whether observers subsampled just one random item from the set of faces to perform the task, we inserted different set size control conditions, to simulate participants' responses if they used subsampling strategies.

### 2.1. Methods

#### 2.1.1. Participants

A total of thirty-five student volunteers were recruited for this experiment (http://www.sona-systems.com/) at the School of Psychology of University of Aberdeen (UK). This sample size was in line with previous studies that investigated ensemble perception (Epstein & Emmanouil, 2021; Khayat & Hochstein, 2018). Participants were undergraduate students at the University of Aberdeen (UK) and received course credits for their participation. All the research methods and procedures were approved by the Ethics Committee of the School of Psychology of the University of Aberdeen (UK) and participants gave informed consent at the beginning of the study.

As exclusion criteria, we removed from the analysis any participants that declared that their data should be excluded for experienced technical problems (one participant excluded), external distractions during the course of the experiment, or misunderstanding of the instructions (three participants excluded). This information was gathered with a questionnaire at the end of the experiment (see below). Thirty-one participants (22 females, 9 males, M = 21.1 years, s.d. = 2.4 years) were included in the study.

Since our study was run online and we could not control for participants' attention during the completion of the experiment, we decided to additionally analyse participants' data across the four experiments with stricter exclusion criteria in order to exclude from the analysis observers that did not pay attention, or who were less able to complete the task (see Supplementary Materials). More specifically, these stricter exclusion criteria consisted in (1) removing from the analysis participants that had a chance level performance in at least one of the 6 blocks of trials, and (2) excluding participants with an error rate in one block that exceeded more than twice the average performance in the other blocks, thus indicating an unstable performance. Overall, the results obtained with these stricter exclusion criteria were consistent with the results obtained with the present exclusion criteria (see Supplementary Materials).

#### 2.1.2. Stimuli

The experiment was created using PsychoPy (https://psychopy.org/, Peirce et al., 2019), and participants' data were collected online through the Pavlovia platform (http://www.pavlovia.org). Participants were asked to conduct the experiment only on computers (tablets were excluded) and to use the Google Chrome browser.

The stimuli were computer-generated neutral faces varying in their level of trustworthiness (Fig. 1A). The images were obtained from a dataset of forward-facing computer-generated FaceGen images created to have both a trustworthy and an untrustworthy looking appearance (Todorov et al., 2013). The face stimuli selected from the dataset consisted of three facial identities with a corresponding trustworthy and untrustworthy version. In order to control for facial low-level features, faces were converted to greyscale, and their luminance and contrast were adjusted to the average of all images by using the SHINE toolbox (Willenbockel et al., 2010) in MATLAB R2017b (The MathWorks, USA). We chose only male-looking faces, as previous research has shown that gender interacts with trustworthiness (Sutherland et al., 2015). Furthermore, given that a very trustworthy-looking appearance make male faces look more feminine, the trustworthy-looking versions of each identity had a level of +1 s.d. on the trustworthiness dimension, whereas the untrustworthy-looking morphed faces versions had a level of −3 s.d. on the trustworthiness dimension. We decided to apply this asymmetry for the concern that the more trustworthy looking were the faces, the more they might start to convey a gender transformation and might appear androgynous and female (Oliveira et al., 2020).

For each identity, a face morph continuum of 51 grayscale images was created between the trustworthy and untrustworthy looking initial faces using PsychoMorph software (https://users.aber.ac.uk/bpt/jpsychomorph/; for a guide to morphing procedure, see Sutherland, 2015; Sutherland, Rhodes and Young, 2017;Tiddeman et al., 2001). The least trustworthy morphed face was labeled as 0, and the most trustworthy morphed face was labeled as 50 (Fig. 1A).

Since the experiment was run on participants' home computers and thus viewing distance could not be controlled, we defined the size of stimuli in PsychoPy "Height" units (each face image had a 0.33 (horizontal) and a 0.33 (vertical) PsychoPy "height" unit size). These units indicate the size of the stimulus in relation to the window's aspect ratio. The horizontal index for this kind of unit ranges from −0.5 (bottom of the screen) to +0.5 (top of the screen). The vertical index is computed based on the ratio between the height of the screen window and its width. For instance, the coordinates of a screen with a 16:10 aspect ratio go from the bottom left (−0.8, −0.5) to the top right (+0.8, +0.5).

In each trial, a set of five faces was presented in a X-shaped format on the screen. The minimum distance between any two faces was 0.36 PsychoPy "height" unit coordinates. In order to avoid face adaptation across trials (Afraz & Cavanagh, 2008), a random number ranging from −0.04 to 0.04 was added to the x and y PsychoPy "height" unit coordinates of each face location. As a result, the position of the faces on the screen slightly varied from trial to trial, even across subsequent trials with the same set size condition. Stimuli were presented against a homogenous grey background.

#### 2.1.3. Procedure

On each trial, a group of five faces varying in the level of trustworthiness was displayed for 2000 ms. The procedure to select which morph faces were displayed was as follows: at the beginning of each trial, a position on the morph continuum of one of the three identities was pseudo-randomly selected; this initial position corresponded to a face on the morph continuum, which represented the mean morphed face (not displayed to participants) of the group of five faces shown on the screen.

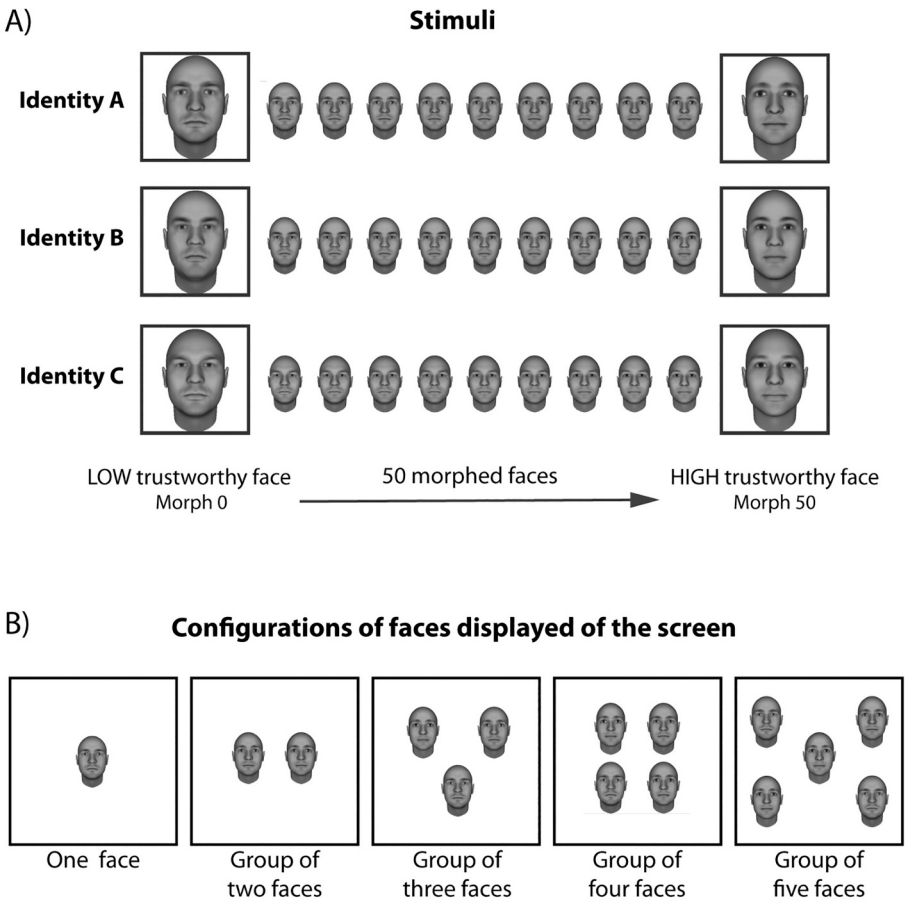



**Fig. 1.** Stimuli. (A) Three pairs of computer-generated FaceGen images with extreme trustworthiness levels (Todorov et al., 2013) were used as the extremes to create a morph continuum of 51 faces. The morph continuum ranged from a very untrustworthy looking face (morph 0) to a very trustworthy looking face (morph 50). (B) Configurations of the different sets of faces that were displayed on the screen.

This mean corresponded to the arithmetic mean of all the morph values of the five faces in the set, and it could range from 11 to 40 morph units. Next, five unique morphed faces surrounding the mean were selected with a standard deviation of 15. To avoid repeated average values across trials that could result from a true random selection, we constrained the mean value of the set to vary between 11 and 40.

Across trials, different numbers of faces could be displayed from this initially selected set of five faces: in the set size five condition, the whole set of five initially selected faces was displayed, whereas within the other set size conditions, one, two, three, or four morphed faces selected from the original set of five faces were displayed (Fig. 1B). We constrained the average of items displayed on the screen to be equally likely to vary from 11 to 40 independently from the set size condition. Thus, the mean trustworthiness of larger sets of faces was just as probably to vary from 11 to 40 as the mean trustworthiness of smaller sets. Across set size conditions, stimuli could be shown at the centre of the screen (set size 1), on the left and right of the centre of the screen (set size 2), in a triangle-shaped format (set size 3), in an invisible 2 × 2 matrix (set size 4), and in an X-shaped format (set size 5) (Fig. 1B). The location of each morphed face from the initial face set was randomly assigned across the possible screen locations for a certain set size condition. The identity displayed was randomly chosen from three possible identities and was always held the same within each trial.

After a brief inter-stimulus-interval (ISI) of 250 ms, a randomly selected (test) face from the same morph continuum was then shown in the middle of the screen. Participants were asked to adjust the appearance of the test face by scrolling through a morphed continuum with the left/right arrow keyboard keys. The task was to match the average level of trustworthiness of the previously seen group of faces. We chose an adjustment task since it permits to directly test participants' perception of the average trustworthiness of the group of faces and it allowed to avoid edge effects in responses or demand bias, which might have been present in a rating task. After selecting the face, they confirmed their response by pressing the spacebar. In the set size one condition trials (where only one face was shown), participants were asked to adjust the trustworthy appearance of the test face to match the previously seen face. Participants were given unlimited time to respond. Finally, following a 500 ms Inter-Trial-Interval (ITI), the next trial started. The trial sequence is illustrated in Fig. 3A.

Six blocks of 40 trials each were presented, for a total of 240 trials. On each block, in eight trials the set of five faces was displayed. In the remaining 32 trials, sets of four, three, two, or one faces from the group of initially selected five faces were displayed in equal proportion, resulting in a total of 48 trials for each set size condition across the whole experiment (8 trials * 6 blocks). During the experiment, trials with different sets of faces were randomly interleaved. We inserted different set size conditions to estimate the number of faces participants were able to incorporate into an ensemble percept of trustworthiness impressions (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Yamanashi Leib et al., 2020; see Data Analysis section for further details on this integration analysis). At the end of each block, participants received their block average error as feedback on their performance and could take a break.

In order to familiarise themselves with the stimuli and the adjustment task, before starting the experiment, each participant was instructed on how to scroll through left/right arrow keys over the 50-

step morph continuum for each identity. In addition, participants completed a training block of 15 example trials. The first five trials had a longer presentation (4000 ms) of the set of faces to give participants time to get accustomed to the stimuli. The following 10 trials had the standard duration (2000 ms), which was then held constant across the entire experiment.

The experiment lasted approximately 45 min. In the final part, participants were asked whether their data should be excluded for any reason (i.e., experienced technical problems, misunderstanding of the instructions) and they could type their response.

### 2.2. Data analysis

Across the four experiments, we conducted the following analyses:

#### 2.2.1. Performance analysis

For each participant, we calculated the adjustment response, or average response error, relative to the average of the face items presented on the screen in each trial for each set size condition. In order to determine the chance level of performance for the task, we shuffled the association between responses and the mean of faces displayed on the screen 10,000 times. On each iteration, we calculated the mean of the response errors (response minus the average of the morphed faces displayed on the screen). This procedure allowed us to estimate the chance level of performance by calculating the mean of the bootstrapped average error distribution. Based on this analysis, we concluded that average response errors above 14.2 morph units (95% CI [9.9, 18.83]) indicated a random performance. Critically, the response error analysis does indicate whether observers were integrating information from the set of faces. Different trends from that analysis (improvement or deterioration in performance) do not have any theoretical implication for the purpose of our research question. Instead, the response error analysis indicates whether participants are indeed doing the task for all set sizes (this is not a trivial issue, as participants could respond randomly with five faces, for example).

#### 2.2.2. Integration analysis

Although we explicitly asked participants to report the average of all displayed faces, it is possible that they might have randomly reported the trustworthiness of one face from the crowd, instead of integrating the trustworthiness of the whole set of faces - a necessary requirement for ensemble coding (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Whitney & Yamanashi Leib, 2018). We therefore presented different set size (one to five) conditions to ascertain whether participants were able to extract an ensemble percept of trustworthiness impressions and critically to rule out subsampling. The different set size conditions allowed us to simulate participants' performance in the case in which they had randomly subsampled just a subset of faces from the crowd, instead of integrating the whole set to respond (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Yamanashi Leib et al., 2020). Thus, in order to understand whether participants were integrating or not, we took into account the initially selected mean, which was the randomly selected face from the morph continuum at the beginning of each trial (not displayed to participants). This initially selected mean did not correspond to the mean of the morph displayed on the screen but can be considered as a "theoretical" value for each trial that we used for running the integration analysis. On each trial, five unique morphed faces surrounding this mean were selected, and depending on the trial set size condition, a different number of faces from the set of five were displayed (one, two, three, four, or five). Here we measured integration by calculating how much participants' responses deviated from the initially selected mean (adjustment response minus the mean of the five initially selected faces). Across the paper, we called this measure of integration "response deviation from the initially selected mean relative to five items". A high value indicated that participants' responses were highly different from the initially selected mean, whereas a low value indicated that participants' responses were closer to the initially selected mean. On the one hand, if participants integrated multiple faces into an ensemble percept, their response deviation relative to the set size five mean should decrease as a function of set size, as more information is given with the increased number of items displayed from the initially selected set of five faces (Fig. 2A). On the other hand, if participants only randomly sampled one face from the set on each trial, their response deviation relative to the set size five mean would be high and remain stable across set size conditions (Fig. 2B), because people are subsampling even when information relative to the number of faces from the set of five is increasingly available across set sizes.

### 2.3. Results and discussion

First, we investigated whether participants were able to extract an ensemble percept of trustworthiness impressions from multiple faces. For this purpose, we calculated how much participants were able to adjust the test face to match the average of the morphed faces previously displayed on the screen (performance analysis). We computed the average response error relative to the items presented on the screen in each trial for each set size condition (adjustment response minus the average of the morphed faces displayed on the screen). Participants' average response time, which was measured from when the stimuli disappeared and they could respond, was 2500 ms (s.e. = 150). As shown in Fig. 3B, participants' average errors were far below the chance performance in all the set size conditions. The average adjustment errors were higher as a function of set size (average adjustment error for 1 face = 9.52, ± s.e: 0.50; 2 faces = 11.02 ± 0.50; 3 faces = 11.32 ± 0.50; 4 faces = 11.45 ± 0.50; 5 faces 11.83 ± 0.52). A repeated measures ANOVA showed a significant main effect of set size on adjustment error ($F(4, 120) = 12.38$, $p < .001$, $\eta^2 = 0.29$). Next, sequential pairwise dependent sample *t*-tests were run to investigate whether participants' performance increased or decreased across set size conditions. Specifically, we compared participants' average adjustment errors of set size 1 and 2 ($t(30) = -3.42$, $p < .05$), set size 2 and 3 ($t(30) = -0.96$, $p = .34$), set size 3 and 4 ($t(30) = -0.36$, $p = .72$), set size 4 and 5 ($t(30) = -1.24$, $p = .22$). These t-tests suggested that participants' performance deteriorated as set size increased. In addition, we fitted for each participant a linear regression slope of the five average errors in each set size condition, and we ran a one-sample t-test against zero on the linear regression slopes. Participants' performance decreased as a function of set size ($t(30) = 6.14$, $p < .001$, $d = 1.10$). This decrease in performance might be explained by considering that the adjustment task may have been easier when one face was presented on the screen and participants successively adjusted the appearance of one test face. Contrary, when multiple faces were displayed, the adjustment task might have become more difficult. Overall, these results confirmed that observers were able to extract the average trustworthiness of the set of faces. However, although this analysis was informative relative to participants' performance based on the number of items displayed on screen, it did not assess whether observers integrated the faces displayed into an ensemble percept.

In addition, we considered participants' average adjustment errors across blocks to investigate a possible decrease in error rates during the experiment (see Supplementary Materials). We did not find any indication of error rates decreasing across blocks, suggesting that no form of learning occurred across the duration of the experiment.

Second, and more importantly, we investigated whether participants integrated the trustworthiness of the whole set of faces displayed on the screen to report their response - a necessary requirement for ensemble coding (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Whitney & Yamanashi Leib, 2018) - or whether they instead randomly reported the trustworthiness of a subset of faces from the crowd. For this purpose, we calculated the response deviation from the initially selected mean relative to five items (adjustment response minus mean of the five initially selected morphed face) for each set size

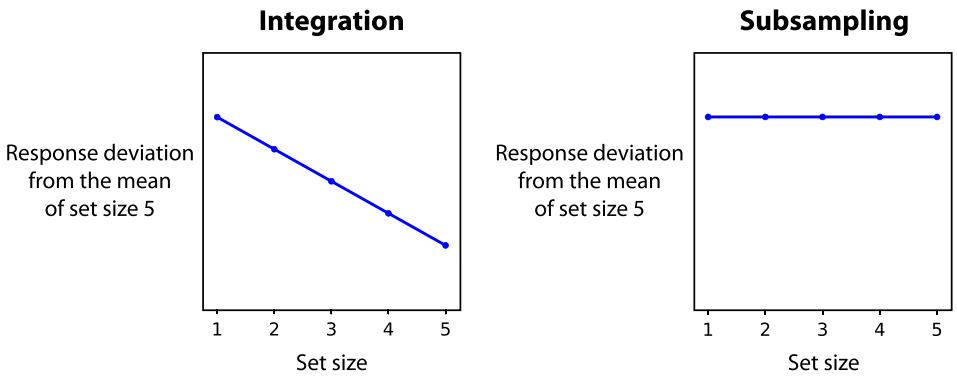



**Fig. 2.** Integration analysis in ensemble perception experiments. (A) If participants integrated multiple faces into an ensemble percept, their response should become more and more similar to the initially extracted set size five mean, as more information is given across set sizes with the increase in the number of faces displayed from the set. Subsequently, their response deviation from the mean of set size five should decrease as a function of set size. (B) If participants only randomly sampled one face from the set on each trial, their response would remain distant from the initially selected set size five mean, even if information on the set size five mean is increasingly available with the increase in the number of faces displayed from the set of five across set size conditions.

condition (integration analysis). Fig. 3C illustrates the participants' average response deviation from set size five mean for each set size condition. As we revealed more information to the participants by showing more items from the set of five faces, their adjustment responses deviated less from the initially selected set size five mean (average response deviation for 1 face = 14.07 ± s.e: 0.34; 2 faces = 12.31 ± 0.50; 3 faces = 11.85 ± 0.50; 4 faces = 11.65 ± 0.48; 5 faces = 11.83 ± 0.52). A repeated measures ANOVA showed a significant main effect of set size ($F(4, 120) = 13.52$, $p < .001$, $\eta^2 = 0.31$). Sequential pairwise dependent sample *t*-tests were run to investigate the direction of the effect. Specifically, we compared participants' average adjustment responses deviation from the initially selected set size five mean of set sizes 1 and 2 ($t(30) = 4.13$, $p < .001$), set sizes 2 and 3 ($t(30) = 1.20$, $p = .24$), set sizes 3 and 4 ($t(30) = 0.57$, $p = .57$), set sizes 4 and 5 ($t(30) = -0.55$, $p = .59$). These t-tests showed that participants' average deviations from the set size five decreased from set size 1 to set size 2 condition. In addition, for each participant, we fitted a linear regression slope of the average deviation in each set size condition from the set size five mean and we ran a one-sample t-test against zero on the linear regression slopes to confirm whether participants shared this pattern on average. This individual-level analysis demonstrated a significant decrease in participants' response deviation from the set size five mean as a function of set size condition ($t(30) = -5.54$, $p < .001$, $d = 0.99$). Taken together, these results show that participants extracted an ensemble percept of trustworthiness impressions from multiple faces displayed on the screen, instead of randomly subsampling one item from the crowd. Interestingly, the integration slope was particularly steep for response deviations from set size 1 to set size 2, whereas it was far less steep from set size 3 to 5, thus participants appeared to be able to integrate around two faces from the group displayed on average. To further investigate the possibility that the decrease in participants' average deviations from the set size five as a function of set size was largely driven by the data for "set size 1", suggesting the integration of two faces in an ensemble percept, we conducted an additional analysis. We fit linear regression lines for each participant's data from set size 2 to 5 (without considering set size 1 data), and we run a one-sample t-test against zero on the linear regression slopes. We did not find a significant effect of set size condition on participants' response deviation from the set size five mean average ($t(30) = -1.46$, $p = .15$, $d = -0.26$). Integrating two or more items is sufficient evidence for an ensemble representation (Whitney & Yamanashi Leib, 2018), but this result may indicate that there could be a limit in the capacity to integrate trustworthiness information. Interestingly, in Experiments 2 and 3 observers were able to integrate three faces from the group displayed on average. We will return to this point in the General Discussion.

Overall, Experiment 1 results revealed that ensemble perception can occur for trustworthiness impressions from a crowd of faces.

## 3. Experiment 2

In Experiment 1, we found that participants were able to extract an ensemble percept of trustworthiness impressions from a group of faces in an adjustment task. The advantage of using an adjustment task is that it directly tests participants' perception of the average trustworthiness of the group while avoiding edge effects in responses or demand bias. However, a natural limitation of the adjustment task used in Experiment 1 was that it increased in difficulty with increasing set size, as shown by participants' overall performance. Indeed, it is naturally easier to adjust the test face appearance when one face was presented on the screen compared to when multiple faces were displayed. Moreover, the downside of using an adjustment task is that it may be based just on the adjustment of the general facial appearance of the faces, and it might not rely on explicit facial trustworthiness judgements per se. To rule out this possibility and to test whether the judgement of average trustworthiness really involves explicit facial trustworthiness judgements, we ran Experiment 2, in which we used a rating scale task instead of an adjustment task. Our aim was to ensure that the extracted summary statistical information was not related to low-level visual cues in the face but regarded the holistic properties of faces that make a face appear more or less trustworthy. If participants in Experiment 1 were just adjusting the test face without integrating together trustworthiness judgements from the crowd, we would no longer expect to find an ensemble effect in Experiment 2. Conversely, if participants were integrating the level of trustworthiness of multiple faces to compute the average level of trust of the crowd, then in Experiment 2 we would expect to find an integration size effect similar to the one found in Experiment 1. Thus, using a rating scale task allowed us to check that the ensemble percept of trustworthiness impressions is not (just) based on the overall appearance of the faces, but it involves a judgement of trustworthiness.

### 3.1. Methods

#### 3.1.1. Participants

A total of thirty-eight students were recruited for this experiment (http://www.sona-systems.com/). Participants were undergraduate students at the University of Aberdeen (UK) and received course credits for their participation. We used the same exclusion criteria as Experiment 1. One participant was excluded from the analysis since they declared to having experienced external distractions during the experiment which affected the data accuracy. Thirty-seven participants (28 females, 9 males, M = 24.2 years, s.d. = 6.1 years) were therefore included in the study.

#### 3.1.2. Stimuli and procedure

The stimuli and procedure in Experiment 2 were the same as described in Experiment 1, except for a rating task instead of an adjustment task (Fig. 4A). In Experiment 2, participants saw a rating

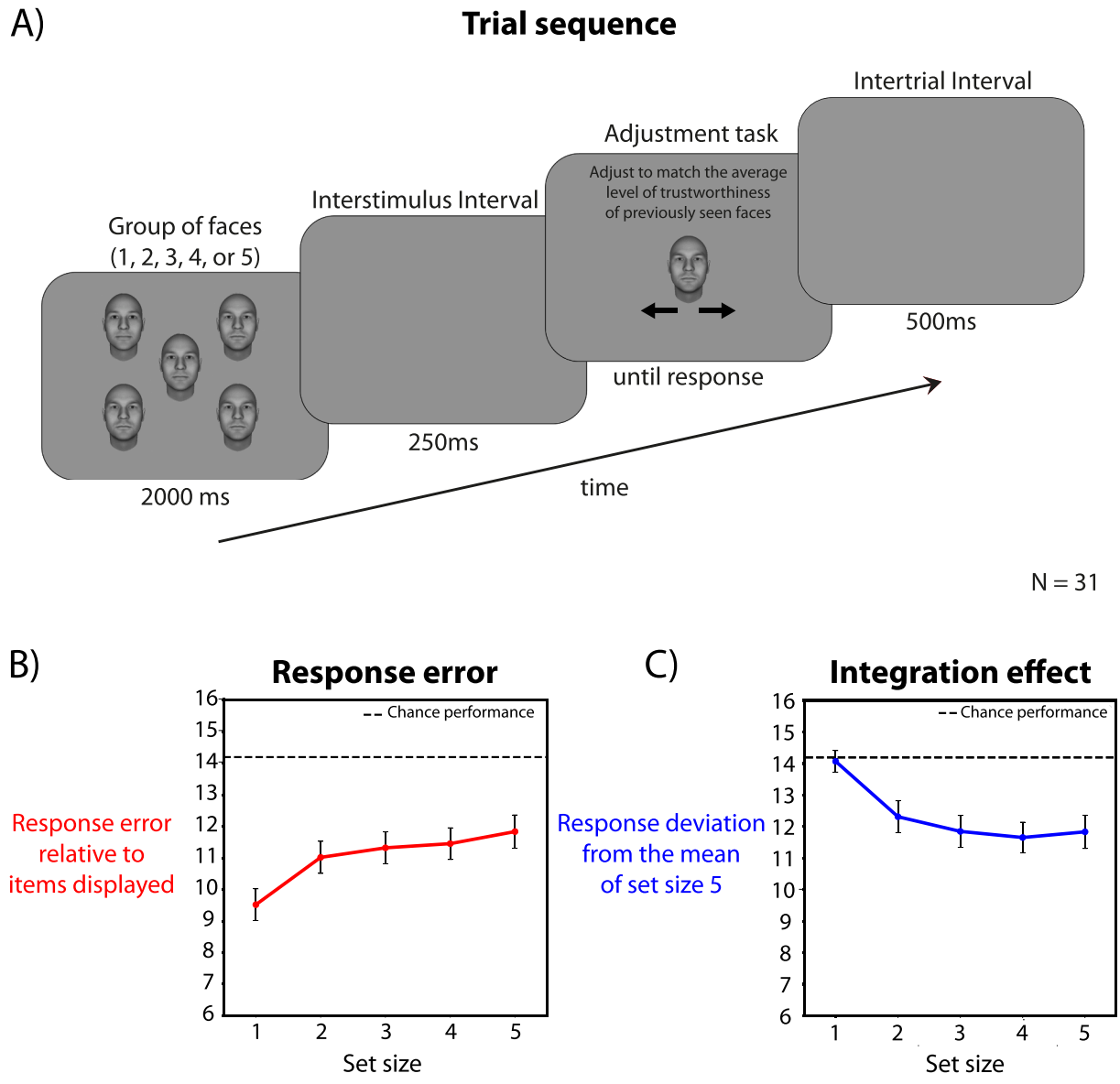



**Fig. 3.** Trial sequence and results of Experiment 1. (A) In each trial participants were shown a group of computer-generated faces of the same identity varying mainly in their level of trustworthiness (Todorov et al., 2013). After an ISI of 250 ms, they were asked to adjust a randomly selected test face to match the average level of trustworthiness of the previously presented faces. After an ITI of 500 ms the next trial started. (B) Performance analysis: the participants' average error relative to items presented on the screen in morph units is plotted on the y-axis (response minus the average of the morphed faces displayed on the screen). The set size conditions are plotted on the x-axis; a different number of faces from the set of five initially selected faces were displayed (one, two, three, four, or five). Error bars represent the standard error of the mean. (C) Integration analysis: the participants' average error relative to the initially selected set size five mean in morph units is plotted on the y-axis (response minus the mean of the five initially selected faces); the set size conditions are plotted on the x-axis. Error bars represent the standard error of the mean. Here the average error relative to the initially selected set size five mean decreases as a function of set size; this tendency suggests that participants are increasingly integrating multiple faces from the set of faces displayed.

scale from 0 to 50 in steps of 1 unit appearing in the middle of the screen. They were asked to click on the rating scale to indicate the average level of trustworthiness of the previously seen group of faces. In set size one condition trials, when only one face was shown, participants were asked to indicate on the rating scale their judgement on the previously seen face's level of trustworthiness. We also had an initial familiarisation and practice trials, as in Experiment 1.

### 3.2. Results and discussion

First, we investigated whether participants were able to extract an ensemble percept of trustworthiness impressions from multiple faces. For this purpose, we calculated the average response error relative to the items presented on the screen in each trial for each set size condition (performance analysis). To measure group performance, we averaged the mean response error of all the participants in each set size condition. The average response time on each trial across the participants was 4340 ms (s.e. = 173 ms). As shown in Fig. 4B, participants' average errors were far below the chance performance (14.2 morph units) in all the set size conditions. The average rating errors were slightly lower as a function of set size (average rating error for 1 face = 9.75, s.e. ± 0.43; 2 faces = 9.50 ± 0.50; 3 faces = 9.24 ± 0.44; 4 faces = 9.12 ± 0.41; 5 faces 9.12 ± 0.46). A repeated measures ANOVA showed a significant main effect of set size ($F(4, 144) = 1.497$, $p = .020$, $\eta^2 = 0.04$). To investigate whether participants' performance increased or decreased across set size conditions, pairwise dependent sample *t*-tests were run. Specifically, we compared participants' average adjustment errors of set size 1 and 2 ($t(36) = 0.74$, $p = .46$), set size 2 and 3 ($t(36) = 0.85$, $p =$

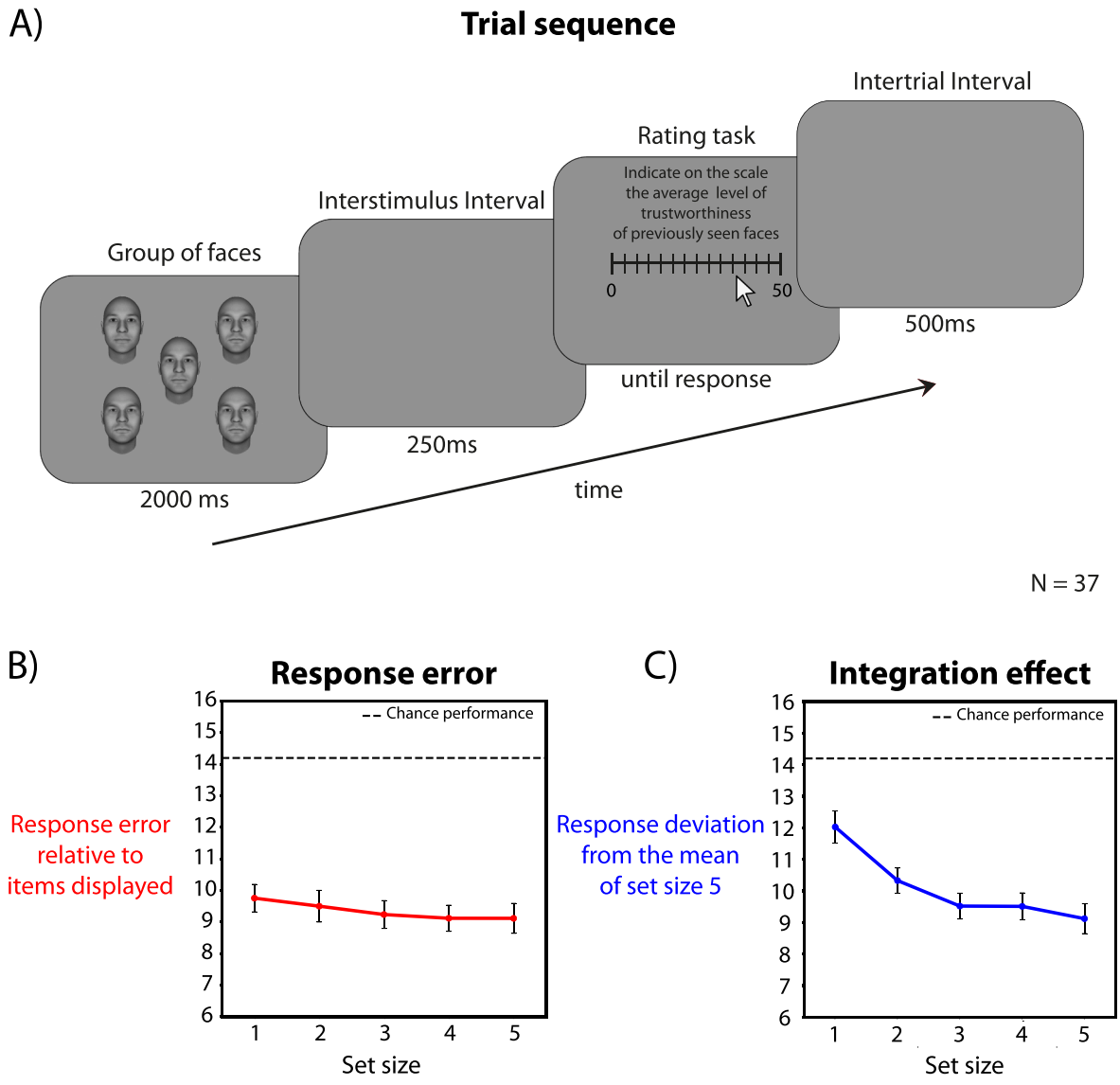



**Fig. 4.** Trial sequence and results of Experiment 2. (A) Experiment 2 paradigm was the same as Experiment 1, with a rating scale task instead of an adjustment task. (B) Performance analysis: the participants' average error relative to items presented on the screen in morph units is plotted on the y-axis (response minus the average of the morphed faces displayed on the screen). The set size conditions are plotted on the x-axis; a different number of faces from the set of five initially selected faces were displayed (one, two, three, four, or five). Error bars represent the standard error of the mean. (C) Integration analysis: the participants' average error relative to the initially selected set size five mean in morph units is plotted on the y-axis (response minus the mean of the five initially selected faces); the set size conditions are plotted on the x-axis. Error bars represent the standard error of the mean. Here the average error relative to the initially selected set size five mean decreases as a function of set size; this tendency suggests that participants are increasingly integrating multiple faces from the set of faces displayed.

.40), set size 3 and 4 ($t(36) = 0.45$, $p = .64$), set size 4 and 5 ($t(36) = 0.009$, $p = .99$). These t-tests suggested that participants' performance slightly improved when sets of five faces were presented. Moreover, we ran a one-sample t-test against zero on the linear regression slopes of each participant fitted on the average errors of each set size condition. However, this analysis indicated no significant improvement in performance as a function of set size conditions ($t(36) = -1.78$, $p = .08$, $d = -0.29$), suggesting that participants' maintained a similar performance far below chance across all the set size conditions. Overall, the results confirmed that observers were able to extract the average trustworthiness of the set of faces.

Second, and more importantly, we investigated whether participants integrated the trustworthiness of the whole set of faces displayed on the screen to report their response (instead of randomly reporting the trustworthiness of few faces from the crowd). For this purpose, we calculated the response deviation from the initially selected mean relative to five items (integration analysis). Fig. 4C illustrates the participants' average response deviation from set size five mean for each set size condition. As we revealed more information on the initially selected mean by displaying higher subsets of faces, the participants' adjustment responses deviated less from the initially selected set size five mean (average response deviation for 1 face = 12.02, s.e. ± 0.51; 2 faces = 10.32 ± 0.40; 3 faces = 9.51 ± 0.40; 4 faces = 9.50 ± 0.42; 5 faces = 9.11 ± 0.47). A repeated measures ANOVA confirmed a significant main effect of set size ($F(4, 144) = 20.91$, $p < .001$, $\eta^2 = 0.37$). Pairwise dependent sample t-tests were run to investigate the direction of the effect. Specifically, we compared participants' average adjustment responses deviation from the initially selected set size five mean of set size 1 and 2 ($t(36) = 3.80$, $p < .001$), set size 2 and 3 ($t(36) = 2.90$, $p < .05$), set size 3 and 4 ($t(36) = 0.03$, $p = .97$), set size 4 and 5 ($t(36) = 1.64$, $p = .11$). These t-tests showed that participants' average deviations from the set size five decreased as a function of set size. In addition, for each participant, we fitted a linear regression slope of the average deviations from the set size five mean in each set size condition and we ran

a one-sample t-test against zero on the linear regression slopes ($t(36) = -5.95$, $p < .001$, $d = -0.97$). Results indicated a decrease in deviations from the set size five mean as a function of set size condition. Moreover, to investigate whether the integration effect was mainly driven by the average data for "set size 1" we fit linear regression lines from each participant's data from set size 2 to 5 and then run a one-sample t-test against zero on the linear regression slopes.

Results indicated a significant decrease in deviations from the set size five mean as a function of set size condition ($t(36) = -4.60$, $p < .001$, $d = -0.76$).

These results showed that participants extracted an ensemble percept of trustworthiness impressions from multiple faces displayed on the screen, instead of randomly subsampling one item from the crowd. In this experiment, the integration slope was particularly steep for response deviations from set size 1 to set size 3, and from set size 4 to 5 it was less dramatically steep. Again, this pattern might indicate that participants are on average able to integrate around three faces from the group displayed. Overall, Experiment 2 revealed that ensemble perception occurs for trustworthiness impressions, now directly measured by judgements on a crowd of faces.

## 4. Experiment 3

In Experiments 1 and 2, we found that participants were able to extract an ensemble perception of trustworthiness impressions from a group of faces both in an adjustment task (Experiment 1) and a rating task (Experiment 2). In the first two experiments we presented faces with the same identity within each trial. We used this procedure in order to completely control for identity (given that ensemble perception of face identity is already very well-known, e.g., de Fockert & Wolfenstein, 2009; Leib et al., 2014). However, to investigate in a more ecological way whether ensemble perception of trustworthiness really involves facial trustworthiness judgements, we carried out Experiment 3 such that there were now multiple identities within a single trial. The advantage of using multiple identities within the same trial is that they have higher ecological validity, as judging the overall trustworthiness of a crowd based on multiple identities instead of just one represents how we experience groups of faces every day. Here, within each trial we displayed to participants groups of faces varying in their level of trustworthiness with different identities and asked them to indicate the average level of trust of the crowd in a rating scale task. On the one hand, if participants' responses were based just on the facial identity similarity of the crowd, in Experiment 3 we would expect observers to not be able to indicate the average trustworthiness level of the displayed faces. On the other hand, if participants' responses were based on trustworthiness judgements of faces, in Experiment 3 we would expect observers to extract an ensemble percept of trustworthiness impressions from multiple faces, as found in Experiments 1 and 2.

### 4.1. Methods

#### 4.1.1. Participants

Thirty-six student volunteers were recruited for this experiment (http://www.sona-systems.com/). Participants were undergraduate students at the University of Aberdeen (UK) and received course credits for their participation. The exclusion criteria were the same as Experiment 1. One participant was excluded from the analysis since they reported to having experienced external distractions during the course of the experiment that affected their data quality. Thirty-five participants (20 females, 14 males, 1 non binary, $M = 26.3$ years, s.d. $= 9$ years) were thus included in the study.

#### 4.1.2. Stimuli and procedure

We used the same stimuli and procedures described in Experiment 2, except for the fact that in Experiment 3 we displayed different identities within the same trial (Fig. 5A). In each trial, the identity of each face displayed on the screen was randomly chosen from the three possible identities (Fig. 1A).

### 4.2. Results and discussion

First, we investigated whether participants were able to extract an ensemble percept of trustworthiness impressions from multiple faces. The average response time of all the participants was 6200 ms (s.e. = 285 ms). We calculated the average response error relative to the items presented on the screen in each trial for each set size condition (performance analysis). To measure group performance, we averaged the mean response error of all the participants in each set size condition. As is shown in Fig. 5B, participants' average errors were below the chance performance (14.2 morph units) in all the set size conditions. The average rating errors were lower as a function of set size (average rating error for 1 face = 9.30, s.e. ± 0.44; 2 faces = 8.48 ± 0.41; 3 faces = 8.39 ± 0.41; 4 faces = 7.99 ± 0.32; 5 faces = 8.13 ± 0.36). A repeated measures ANOVA showed a significant main effect of set size ($F(4, 136) = 5.66$, $p < .001$, $\eta^2 = 0.14$). To investigate whether participants' performance increased or decreased across set size conditions, pairwise dependent sample *t*-tests were run. Specifically, we compared participants' average adjustment errors of set size 1 and 2 ($t(34) = 2.34$, $p < .05$), set size 2 and 3 ($t(34) = 0.42$, $p = .67$), set size 3 and 4 ($t(34) = 1.58$, $p = .12$), set size 4 and 5 ($t(34) = -0.49$, $p = .62$). These t-tests suggested that participants had a better performance as set size increased, contrary to Experiment 1 results. In addition, we carried out a one-sample t-test against zero on the linear regression slopes of the average errors in each set size condition of each participant. This analysis indicated a significant improvement in performance as a function of set size conditions ($t(34) = -3.63$, $p < .001$, $d = -0.61$). Overall, the results confirmed that observers were able to extract the average trustworthiness of the set of faces. However, although this analysis was informative relative to participants' performance based on the number of items displayed on screen, it did not assess whether observers integrated the faces displayed into an ensemble perception.

Second, and more importantly, we therefore investigated whether participants integrated the trustworthiness of the whole set of faces displayed on the screen. To this end, we calculated the response deviation from the initially selected mean relative to five items (integration analysis). Fig. 5C illustrates the participants' average response deviation from set size five mean, for each set size condition separately. As we revealed more information to the participants, their adjustment responses deviated less from the initially selected set size five mean (average response deviation for 1 face = 11.82, s.e. ± 0.42; 2 faces = 9.19 ± 0.29; 3 faces = 8.52 ± 0.37; 4 faces = 8.30 ± 0.33; 5 faces = 8.13 ± 0.36). A repeated measures ANOVA confirmed a significant main effect of set size ($F(4, 136) = 42.9$, $p < .001$, $\eta^2 = 0.55$). Pairwise dependent sample t-tests were run to investigate the direction of the effect. Specifically, we compared participants' average adjustment responses deviation from the initially selected set size five mean of set size 1 and 2 ($t(34) = 7.65$, $p < .001$), set size 2 and 3 ($t(34) = 3.19$, $p < .05$), set size 3 and 4 ($t(34) = 1.01$, $p = .31$), set size 4 and 5 ($t(34) = 0.60$, $p = .54$). These t-tests showed that participants' average deviations from the set size five decreased as a function of set size. In addition, for each participant, we fitted a linear regression slope of the average deviations from the set size five mean in each set size condition and we run a one-sample t-test against zero on the linear regression slopes. This analysis showed a significant decrease in deviations from the set size five mean as a function of the set size condition ($t(34) = -7.57$, $p < .001$, $d = -1.28$). Additionally, we fit linear regression lines from each participant's data from set size 2 to 5 and then run a one-sample t-test against zero on the linear regression slopes to investigate whether the integration effect was driven by the average data for "set size 1". These results suggest that the decrease in deviations from the set size five mean as a function of set size condition was not solely driven by set size 1 data ($t(34) = -3.44$, $p < .01$, $d = -0.58$). These results demonstrated that participants extracted an ensemble percept of

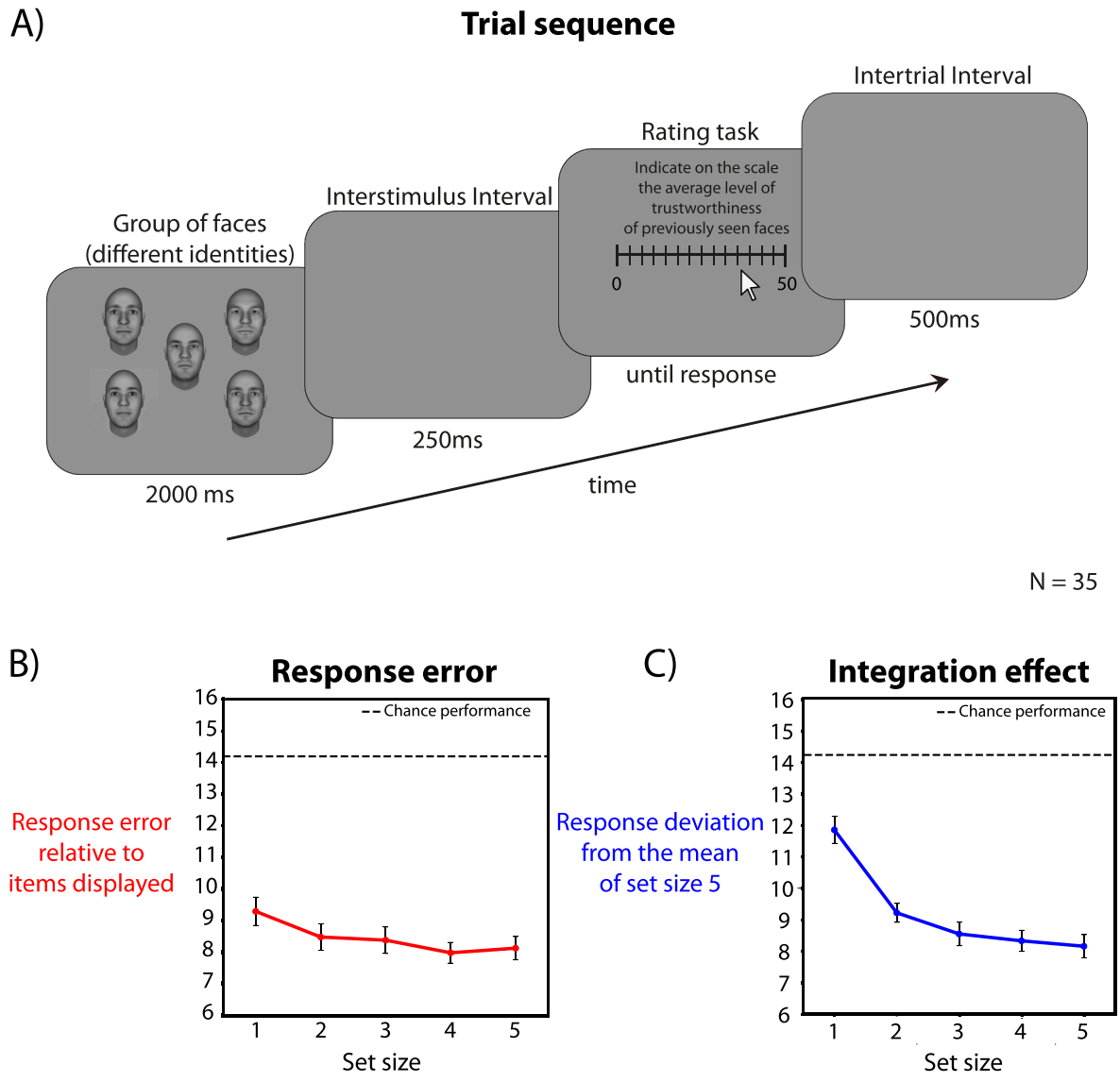



**Fig. 5.** Trial sequence and results of Experiment 3. (A) Experiment 3 experimental procedure was the same as in Experiments 2, except that different identities were presented within the same group of faces. (B) Performance analysis: the participants' average error relative to items presented on the screen in morph units is plotted on the y-axis (response minus the average of the morphed faces displayed on the screen). The set size conditions are plotted on the x-axis; a different number of faces from the set of five initially selected faces were displayed (one, two, three, four, or five). Error bars represent the standard error of the mean. (C) Integration analysis: the participants' average error relative to the initially selected set size five mean in morph units is plotted on the y-axis (response minus the mean of the five initially selected faces); the set size conditions are plotted on the x-axis. Error bars represent the standard error of the mean. The average error relative to the initially selected set size five mean decreased as a function of set size; this tendency suggests that participants were increasingly integrating multiple faces from the set of faces displayed.

trustworthiness impressions from multiple faces displayed on the screen, instead of randomly subsampling one item from the crowd.

Taken together, Experiment 3 showed that ensemble perception occurs for trustworthiness impressions from a group of faces even among different identities, thus suggesting that the ensemble percept is based on trustworthiness judgements which can be separated from identity.

## 5. Experiment 4

In Experiments 1–3, we provided evidence that participants were able to integrate the social context in which faces are embedded in an ensemble perception of trustworthiness impressions when groups of faces were presented for 2000 ms. Importantly, these results suggest that such relevant social judgements occur on a much broader group impressions level. However, an open question is how quickly it is possible to form a summary statistic of trust information from a group. In Experiment 4, we aimed to test the temporal integration limit of ensemble perception of trustworthiness impressions. To investigate this, in Experiment 4, now we presented sets of faces for different time duration conditions (2000 ms, 1000 ms, 500 ms, 250 ms, 100 ms and 50 ms). This experiment also allowed us to test whether the previous results were solely dependent on the 2000 ms stimuli exposure duration. Each participant was presented with blocks of trials with different stimulus durations (in random block order), and participants were asked to indicate on a rating scale the average level of trust of the crowd of faces displayed. In Experiment 4, at the end of each block no feedback on performance was provided.

### 5.1. Methods

#### 5.1.1. Participants

Forty-seven student volunteers were recruited for this experiment

(http://www.sona-systems.com/). Participants were undergraduate students at the University of Aberdeen (UK) and received course credits for their participation. The exclusion criteria were the same as Experiment 1. One participant was excluded from the analysis since they reported to having experienced external distractions during the experiment that affected their data quality, and another participant was excluded because they did not complete the experiment. Forty-five participants (34 females, 9 males, 2 non binary, M = 26.1 years, s.d. = 11.3 years) were thus included in the study.

*5.1.2. Stimuli and procedure*

We used the same stimuli and procedures described in Experiment 3, except for the fact that in Experiment 4 different display durations were included. Specifically, in Experiment 4 the set of faces were shown for 2000 ms (comparable to Experiments 1–3), as well as 1000 ms, 500 ms, 250 ms, 100 ms, and 50 ms.

Six blocks of 60 trials each were presented, for a total of 360 trials (please note that the number of trials per set size condition per block was 12, thus different compared to the Experiments 1, 2 and 3 which had 8 trials of each set size condition for each block). Each display duration

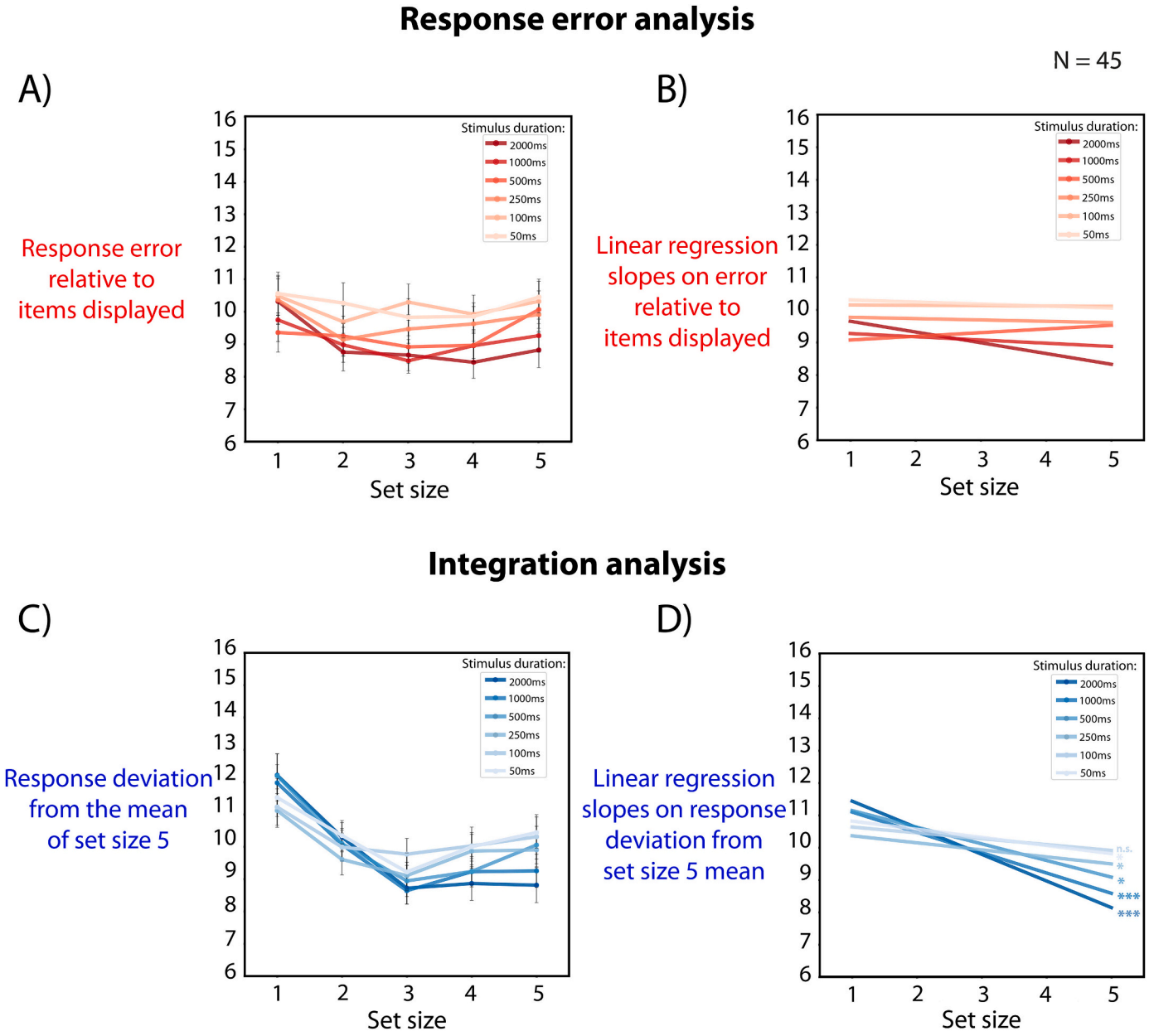



**Fig. 6.** Results of Experiment 4. (A) Performance analysis: the participants' average error relative to items presented on the screen in morph units is plotted on the y-axis (response minus the average of the morphed faces displayed on the screen) for each display duration condition. The set size conditions are plotted on the x-axis; a different number of faces from the set of five initially selected faces were displayed (one, two, three, four, or five). Error bars represent the standard error of the mean. (B) Linear regression slopes of performance: on the y-axis the linear regression slopes of participants' average error relative to items presented on the screen in morph units are plotted for each display duration condition. (C) Integration analysis: the participants' average error relative to the initially selected set size five mean in morph units is plotted on the y-axis (response minus the mean of the five initially selected faces) for each display duration condition; the set size conditions are plotted on the x-axis. Error bars represent the standard error of the mean. The average error relative to the initially selected set size five mean decreased as a function of set size across display time conditions; this tendency suggests that participants were increasingly integrating multiple faces from the set of faces displayed. (D) Linear regression slopes in Integration analysis: on the y-axis the linear regression slopes on participants' average error relative to the initially selected set size five mean in morph units are plotted for each display duration condition.

condition was randomly assigned to one of the six blocks. On each block, in twelve trials a set of five faces was displayed. In the remaining 48 trials, sets of four, three, two, or one faces from the group of initially selected five faces were displayed in equal proportion.

### 5.2. Results and discussion

First, we investigated whether participants were able to extract an ensemble percept of trustworthiness impressions from multiple faces across the display duration conditions. The average response time of all the participants was 1470 ms (s.e. = 140 ms). We calculated the average response error relative to the items presented on the screen in each trial for each set size condition (performance analysis). To measure group performance, we averaged the mean response error of all the participants in each set size condition. As is shown in Fig. 6A, participants' average errors were below the chance performance (14.2 morph units) in all the set size conditions.

The average rating errors were lower as a function of set size, as shown in Table 1. A repeated measures ANOVA showed a significant main effect of set size in the duration condition 2000 ms ($F(4, 176) = 4.68$, $p < .001$, $\eta^2 = 0.10$), but not 1000 ms ($F(4, 176) = 1.92$, $p = .10$, $\eta^2 = 0.05$), 500 ms ($F(4, 176) = 1.98$, $p = .10$, $\eta^2 = 0.043$), 250 ms ($F(4, 176) = 1.62$, $p = .17$, $\eta^2 = 0.036$), 100 ms ($F(4, 176) = 0.76$, $p = .54$, $\eta^2 = 0.017$), and 50 ms ($F(4, 176) = 0.87$, $p = .47$, $\eta^2 = 0.020$). To further investigate whether participants' performance increased or decreased across set size conditions in the duration condition 2000 ms pairwise dependent sample *t*-tests were run. Specifically, we compared participants' average adjustment errors in 2000 ms duration condition for set size 1 and 2 ($t(44) = 2.87$, $p < .05$), set size 2 and 3 ($t(44) = 0.18$, $p = .86$), set size 3 and 4 ($t(44) = 0.46$, $p = .64$), set size 4 and 5 ($t(44) = -1.10$, $p = .27$). These t-tests suggested that participants in the 2000 ms duration condition had a better performance as set size increased, but not in the other exposure duration condition. In addition, we carried out a one-sample t-test against zero on the linear regression slopes of the average errors in each set size condition of each participant for each duration condition (Fig. 6B). This analysis indicated an improvement in performance as a function of set size conditions for 2000 ms ($t(44) = -3.27$, $p < .05$), but not for 1000 ms ($t(44) = -0.77$, $p = .43$), 500 ms ($t(44) = 1.08$, $p = .28$), 250 ms ($t(44) = -0.30$, $p = .76$), 100 ms ($t(44) = -0.09$, $p = .92$), and 50 ms ($t(44) = -0.49$, $p = .62$).

Overall, the results confirmed that observers were able to extract the average trustworthiness of the set of faces. However, although this analysis was informative relative to participants' performance based on the number of items displayed on screen, it did not assess whether observers integrated the faces displayed into an ensemble perception.

Second, and more importantly, we investigated whether participants integrated the trustworthiness of the whole set of faces displayed on the screen. Indeed, even if we explicitly asked participants to report the average of all displayed faces, it is possible that they might have randomly reported the trustworthiness of one face from the group, instead of integrating the trustworthiness of the whole set of faces - a necessary requirement for ensemble coding (Ariely, 2001; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Whitney & Yamanashi Leib, 2018). To this end, we calculated the response deviation from the initially selected mean relative to five items (integration analysis). Fig. 6C illustrates the participants' average response deviation from set size five mean, for each set size condition separately and each exposure duration condition. As we revealed more information to the participants, their adjustment responses deviated less from the initially selected set size five mean for the 2000 ms duration condition (see Table 2).

A repeated measures ANOVA confirmed a significant main effect of set size in the duration conditions 2000 ms ($F(4, 176) = 17.29$, $p < .001$, $\eta^2 = 0.28$), 1000 ms ($F(4, 176) = 16.43$, $p < .001$, $\eta^2 = 0.27$), 500 ms ($F(4, 176) = 14.79$, $p < .001$, $\eta^2 = 0.25$), 250 ms ($F(4, 176) = 5.65$, $p < .001$, $\eta^2 = 0.11$), and 50 ms ($F(4, 176) = 5.32$, $p < .001$, $\eta^2 = 0.10$), but not for 100 ms ($F(4, 176) = 2.20$, $p = .07$, $\eta^2 = 0.04$). Pairwise dependent sample *t*-tests were run to investigate the direction of the effect. Specifically, we compared participants' average adjustment responses deviation from the initially selected set size five mean of set size 1 and 2, set size 2 and 3, set size 3 and 4 and 5 for each duration condition. These t-tests showed that participants' average deviations from the set size five decreased as a function of set size for 2000 ms, 1000 ms, 500 ms, 250 ms,100 ms and 50 ms duration conditions, as shown in Table 3.

In addition, for each participant, we fitted a linear regression slope of the average deviations from the set size five mean in each set size and duration condition and we run a one-sample t-test against zero on the linear regression slopes (Fig. 6D). This analysis showed a significant decrease in deviations from the set size five mean as a function of the set size condition at 2000 ms ($t(44) = -5.83$, $p < .001$), 1000 ms ($t(44) = -5.91$, $p < .001$), 500 ms ($t(44) = -4.12$, $p < .001$), 250 ms ($t(44) = -2.18$, $p < .05$), and 50 ms ($t(44) = -2.05$, $p = .04$), but not at 100 ms ($t(44) = -1.34$, $p = .18$). Next, to investigate whether the integration effect was driven by the average data for "set size 1", we fit linear regression lines from each participant's data from set size 2 to 5 and then run a one-sample t-test against zero on the linear regression slopes. Results suggested a decrease in deviations from the set size five mean as a function of set size condition at 2000 ms ($t(44) = -2.68$, $p < .05$), but not at 1000 ms ($t(44) = -1.45$, $p = .15$), 500 ms ($t(44) = 0.11$, $p = .90$), 250 ms ($t(44) = 1.17$, $p = .24$), 100 ms ($t(44) = 0.79$, $p = .43$), and 50 ms ($t(44) = 0.85$, $p = .39$). These results suggest that for shorter presentation duration two faces from the set might have have been integrated. Integrating two or more items is sufficient evidence for an ensemble representation (Whitney & Yamanashi Leib, 2018), however this result may indicate that there could be a limit in the capacity to integrate trustworthiness information for short exposure.

These results demonstrated that participants extracted an ensemble

**Table 1**
Average rating errors as a function of set size across the exposure conditions of Experiment 4.

| Duration condition | Average rating error | | | | |
|---|---|---|---|---|---|
| | Set size 1 | Set size 2 | Set size 3 | Set size 4 | Set size 5 |
| 2000 ms | 10.30 ± 0.69 | 8.74 ± 0.57 | 8.65 ± 0.47 | 8.43 ± 0.49 | 8.81 ± 0.54 |
| 1000 ms | 9.73 ± 0.66 | 8.97 ± 0.53 | 8.48 ± 0.38 | 8.94 ± 0.48 | 9.25 ± 0.51 |
| 500 ms | 9.34 ± 0.60 | 9.24 ± 0.61 | 8.90 ± 0.50 | 8.95 ± 0.40 | 10.05 ± 0.57 |
| 250 ms | 10.35 ± 0.75 | 9.13 ± 0.49 | 9.45 ± 0.44 | 9.61 ± 0.53 | 9.90 ± 0.52 |
| 100 ms | 10.47 ± 0.59 | 9.68 ± 0.49 | 10.28 ± 0.55 | 9.91 ± 0.58 | 10.30 ± 0.68 |
| 50 ms | 10.54 ± 0.66 | 10.25 ± 0.62 | 9.81 ± 0.57 | 9.84 ± 0.41 | 10.44 ± 0.48 |

**Table 2**
Average response deviation from set size 5 mean as a function of set size across the exposure conditions of Experiment 4.

| Duration condition | Average response deviation from set size 5 mean | | | | |
|---|---|---|---|---|---|
| | Set size 1 | Set size 2 | Set size 3 | Set size 4 | Set size 5 |
| 2000 ms | 12.21 ± 0.65 | 10.29 ± 0.45 | 8.71 ± 0.49 | 8.86 ± 0.52 | 8.81 ± 0.54 |
| 1000 ms | 11.97 ± 0.56 | 10.06 ± 0.45 | 8.64 ± 0.41 | 9.22 ± 0.46 | 9.25 ± 0.51 |
| 500 ms | 12.19 ± 0.67 | 10.09 ± 0.47 | 8.94 ± 0.48 | 9.23 ± 0.39 | 10.05 ± 0.57 |
| 250 ms | 11.12 ± 0.52 | 9.60 ± 0.48 | 9.10 ± 0.41 | 9.86 ± 0.52 | 9.90 ± 0.52 |
| 100 ms | 11.23 ± 0.56 | 9.99 ± 0.43 | 9.77 ± 0.48 | 10.02 ± 0.58 | 10.30 ± 0.68 |
| 50 ms | 11.52 ± 0.58 | 10.34 ± 0.46 | 9.22 ± 0.58 | 10.00 ± 0.44 | 10.44 ± 0.48 |

**Table 3**
Pairwise dependent sample t-tests on participants' average adjustment responses deviation from the initially selected set size five mean across the exposure conditions of Experiment 4.

| Duration conditions | Set size conditions compared in the pairwise dependent sample t-tests | | | |
|---|---|---|---|---|
| | Set size 1 and 2 | Set size 2 and 3 | Set size 3 and 4 | Set size 4 and 5 |
| 2000 ms | $t(44) = 3.60$, $p < .001$ | $t(44) = 4.55$, $p < .001$ | $t(44) = -0.30$, $p = .76$ | $t(44) = 0.13$, $p = .89$ |
| 1000 ms | $t(44) = 4.31$, $p < .001$ | $t(44) = 2.87$, $p < .05$ | $t(44) = -1.62$, $p = .11$ | $t(44) = -0.05$, $p = .95$ |
| 500 ms | $t(44) = 4.42$, $p < .001$ | $t(44) = 3.33$, $p < .05$ | $t(44) = -0.76$, $p = .45$ | $t(44) = -1.70$, $p = .09$ |
| 250 ms | $t(44) = 3.70$, $p < .001$ | $t(44) = 1.23$, $p = .22$ | $t(44) = -2.15$ $p < .05$ | $t(44) = -0.07$, $p = .93$ |
| 100 ms | $t(44) = 2.72$, $p < .05$ | $t(44) = 0.42$, $p = .67$ | $t(44) = -0.55$, $p = .58$ | $t(44) = -0.49$, $p = .62$ |
| 50 ms | $t(44) = 1.98$, $p = .05$ | $t(44) = 2.77$, $p < .05$ | $t(44) = -1.58$, $p = .12$ | $t(44) = -0.96$, $p = .33$ |

percept of trustworthiness impressions from multiple faces displayed on the screen, instead of randomly subsampling one item from the crowd, even when the set of faces was displayed for very short exposure time such as 250 ms. More anecdotally, we also found a small significant effect of set size ($p$-value $<.04$) for a duration of 50 ms.

## 6. General discussion

The present study investigated whether ensemble perception occurs for trustworthiness impressions from a crowd. For this purpose, we asked participants to report the average level of trustworthiness from sets of faces. In Experiment 1, we found evidence of ensemble perception in an adjustment task. In Experiments 2 and 3, by using a rating task, we found that the ensemble percept of trustworthiness impressions is not only based on the overall appearance of the faces, but also involves explicit trustworthiness judgements (Experiment 2), even across different identities (Experiment 3). In Experiment 4, we investigated how quickly observers were able to form a summary statistic of trustworthiness and found that it is possible to integrate facial trust information from a group within 250 ms.

Overall, the present study demonstrates that visual mechanisms of ensemble coding are involved in representing the social properties of a crowd and shed light on the possibility of the social context in which faces are embedded to influence trustworthiness impressions. Importantly, these findings provide an understanding of trustworthiness judgements not only on an individual level, but on a much broader group impressions level. Moreover, our results make a significant contribution to multiple areas of research, from visual perception to social cognition, by showing that the social context in which faces are embedded can be integrated in an ensemble percept of trustworthiness impressions.

Our results are in line with a series of studies that reported ensemble perception for facial high-level features, such as emotional expression (Haberman & Whitney, 2007; Haberman & Whitney, 2009; Li et al., 2016), sex ratio (Alt et al., 2019; Haberman & Whitney, 2007), facial identity (Alt et al., 2019; Bai et al., 2015; de Fockert & Wolfenstein, 2009; Roberts et al., 2019), and attractiveness (Carragher et al., 2021; Luo & Zhou, 2018). Previous work that investigated whether ensemble coding in faces is merely based on low-level physical attributes of stimuli supports the high-level nature of ensemble perception of groups of faces. From Haberman & Whitney (2007, 2009) studies on ensemble perception of faces investigated whether summary statistics from faces are high-level representations; in these studies, participants' ensemble coding effect appeared to be weaker for inverted or scrambled faces. In addition, Leib et al. (2014) found that summary statistical perception of facial identity operated across faces with different viewpoints. Han et al. (2021) further tested this aspect by presenting to participants groups of Mooney faces (Mooney, 1957). These face stimuli are two-tone, shadow-defined images that cannot be recognized in a low-level manner. Thus, to see the image of a person in a Mooney face it must be necessarily processed holistically. The results of this study show that participants were able to compute average emotional valence from crowds of Mooney faces. In addition, observers were not as sensitive to crowds of inverted or scrambled faces compared to canonically oriented ones. These findings suggest that ensemble perception can operate selectively on holistic representations of human faces and constitute high-level representations (Han et al., 2021). Our findings reinforce these ideas and further suggest that not only summary statistics of faces can reflect high-level representations, but that ensemble perception cover a broad range of phenomenon than previously recognized, including complex high-level facial trait judgements such as trustworthiness impressions.

Across the four experiments, individuals integrated multiple faces from the group into an ensemble percept. Integration is an important property of ensemble representations, and previous studies on facial features such as identity or emotional expression found that observers can integrate between 4 and 8 faces from a crowd (Haberman & Whitney, 2010; Leib et al., 2014). We used the integration control conditions to ensure that observers were integrating multiple faces from the group, instead of randomly subsampling subsets of stimuli. This method involves incorporating subset conditions in the experimental paradigm and comparing observers' performance as a function of set size and it has been extensively used in previous studies that investigated ensemble perception (Ariely, 2001; Sweeny & Whitney, 2014; Yamanashi Leib et al., 2020). Moreover, by using a rating task we also ensured that the summary statistic was not based just on the adjustment of the general facial appearance of the faces, but on facial trustworthiness judgements per se dependent. Interestingly, across the four experiments, we found that the integration slope, which illustrated how much participants' responses deviated from the initially selected set size five mean, was particularly steep for response deviations from set size 1 to set size 3, whereas from set size 4 to 5 it appeared to be far less steep, especially in Experiments 2–4. We interpreted this result as an index of the fact that participants appeared to be able to integrate around three faces from the group displayed on average. These results are not against the ability to create an ensemble representation since integrating two or more items is sufficient evidence for an ensemble representation (Whitney & Yamanashi Leib, 2018). However, they indicate that there could be a limit in the capacity to integrate trustworthiness information from a certain number of faces in a summary statistic. Given that trustworthiness impressions are particularly high-level social judgements, this aspect perhaps makes their corresponding ensemble representations a good candidate for a limited capacity process compared to the summary statistics of other less complex facial features, such as facial emotional expressions (Haberman et al., 2009; Haberman & Whitney, 2009; Leib et al., 2014). Currently, whether ensemble perception is characterized by a limited or unlimited capacity of integration of the items of a group presented is still a matter of discussion (Whitney & Yamanashi Leib, 2018). To this regard, recently Ji et al. (2018) studied the processing capacity of ensemble representations of facial emotional expressions in groups of faces, finding evidence of a limited-capacity process. Although multiple factors and individual differences are probably involved in how much information observers can integrate in a single ensemble representation, a general rule of integration capacity has been proposed (Whitney & Yamanashi Leib, 2018). Accordingly, observers are suggested to be able to integrate approximately the square root of the number of the individual stimuli of a group in an ensemble perception (Whitney & Yamanashi Leib, 2018). Intriguingly, our study seems to be in line with this integration capacity rule, since we showed participants a maximum of five faces; and on average they seemed to be able to

integrate around three faces from the crowd of faces presented ($\sqrt{5} = 2.2360$). In this respect, an interesting question for future research would be to precisely determine the capacity limit of integration of the number of faces in ensemble perception of trustworthiness, as well as to compare to other kinds of ensemble coding. For instance, the presentation of particularly noisy groups of faces might decrease integration of trustworthiness still further. Moreover, in Experiment 4 we tested both relatively long exposure time durations (2000 ms) and extremely short ones (50 ms) and found that the integration slopes were steep for response deviations from set size 1 to set size 3 up to 250 ms of face set exposure, thus indicating integration, and were far less steep when stimuli were presented for shorter exposures. These results could be interpreted as evidence of the fact that ensemble perception of trust information has a temporal limit, under which forming a summary statistic of the group becomes challenging.

Interestingly, we found a difference in participants' performance as a function of set size across Experiments. In Experiment 1, participants' adjustment errors relative to the items presented on the screen increased as a function of set size; this showed that participants were more accurate in adjusting the test face to match smaller groups of faces instead of the whole set of faces. In contrast, in Experiments 2 and 3, participants' adjustment errors relative to the items presented on the screen decreased as a function of set size; thus, participants were more accurate in rating the average level of trustworthiness of a more numerous group of faces compared to a smaller ones. This difference might be due to the nature of the tasks used. On the one hand, in the adjustment task in Experiment 1, participants were asked to directly adjust the appearance of one single faces, and this might have been easier when just one face was presented compared to multiple faces. On the other hand, in the rating task in Experiments 2–3 participants did not adjust the appearance of a face but were asked to compute the average and respond on the rating scale, thus there were no facilitation for set size 1. Interestingly, it has been proposed that computing the average of a group of items might be a mechanism that helps cancel uncorrelated noise that derives from individual items (Whitney & Yamanashi Leib, 2018). This idea might be in line with our results, given that we found an increased ability to compute the correct average with the increase in set size. Although this analysis was informative relative to participants' performance based on the number of items displayed on screen, the most important analysis across the four studies was the integration analysis, which showed that observers integrated multiple faces displayed into an ensemble perception.

An ongoing debate in the ensemble perception field (Whitney & Yamanashi Leib, 2018) focuses on whether ensemble representations occur in parallel without attention or whether they are serial processes that require attention (Haberman & Whitney, 2011; Whitney & Yamanashi Leib, 2018; Furtak, Mudrik, & Bola, 2022). Two alternatives to the formation mechanisms of the average representations have been proposed. On the one hand, it has been proposed that the visual system first allocates attentional resources to all the stimuli of the set, and only after creating representations of each item an average representation is computed by averaging all the items of the set (de Fockert & Wolfenstein, 2009; Myczek & Simons, 2008). In this view, individual objects in a scene are firstly accessed, and in a later stage the global percept of the scene emerges, based on the integration of all the elements. On the other hand, it has been proposed that the visual system processes groups of stimuli in parallel and creates an ensemble percept as an average stimulus integrated from the items of a set of stimuli, without acquiring accurate representations of individual stimuli (Ariely, 2001; Hochstein & Ahissar, 2002; Hochstein et al., 2015). This view suggests that the summary process happens automatically, and in parallel, immediately when we encounter groups of objects (Chong & Treisman, 2005). Hochstein et al. (2015) proposed that we first form global statistics instead of individual item representations because they might serve as a mechanism that guides attention and visual-motor behaviour. Attention can be guided for example in identifying deviance in a group of similar objects or singular salient items (Furtak, Mudrik, & Bola, 2022; Haberman & Whitney, 2012). In addition, ensemble perception might be helpful in providing stability in a rapidly changing environment (Manassi et al., 2017) as a mechanism of noise reduction within individual objects representations (Sun & Chong, 2020), thus promoting that sense of having a complete and stable picture of the world even if a representation of individual items in the scene is actually missing. Our experiments were not designed to test whether ensemble perception of trustworthiness impressions happens automatically and in parallel, or as a serial process that requires attention. Either or both mechanisms could reasonably explain our current results, and indeed it would be interesting in future to further probe whether the mechanisms underlying trustworthiness ensemble perception are similar or different to those debated to underly other aspects of face perception.

Even if in everyday life faces are often not perceived in isolation, impression formation and visual perception of groups remain so far two separate and parallel areas of study, and the intersection between them is still unexplored. This gap is surprising, given that group perception is at the basis of many social psychological phenomena, and ensemble perception of groups of faces might be very useful for high-level social processing, as they convey social information that is impossible to extract by analysing singularly each individual in the visual scene. For example, the sex composition of a crowd of faces has been demonstrated to influence the social attitude towards a group and its perceived level of threat (Alt et al., 2019; Goodale et al., 2018), suggesting that the perception of crowd approachability might be influenced by summary statistical information. In addition, visuosocial summary statistical information is not only important for recognition and awareness but is also an important cue to guide action. Understanding quickly whether a group is approachable or not is of vital importance in determining behaviour towards that group. In this view, global impressions of a group can be helpful in simplifying a complex world. If first impressions are functionally relevant because they may reduce cognitive load in decision making, in effect, a visual stereotype (Siddique et al., 2022), then forming a rapid impression of a group of faces is further relevant in reducing cognitive load. Indeed, in contrast to when we encounter a single individual, in a group there is much more facial information to process all together, which would require far more computation and might slow dramatically our decision making. In crowd navigation, for example, behavioural decisions have been proposed to be influenced by summary statistical information, since understanding the average gaze direction and mean head rotation of the crowd allows one to quickly determine where to walk next to or the speed of walking (Sweeny & Whitney, 2014). Thus, there is the possibility that integrated information of facial trustworthiness from the social context may be efficient in guiding social behaviour and permit quick social inferences and decisions important for social interactions.

However, even if the extraction of the mean level of trustworthiness from a group of faces may be useful in guiding social behaviour in some circumstances, trustworthiness face-based inferences are neither accurate nor reliable. The relationship between face morphology and people's trustworthiness traits is far from perfect (Todorov et al., 2015) such that trustworthiness impressions are not accurate enough to be a useful basis for real-world decision-making (Foo et al., 2021). For example, trustworthiness impressions were shown to be detrimental to legal decisions, with a higher likelihood for untrustworthy-looking defendants to be sentenced severely compared to trustworthy-looking ones (Korva et al., 2013), even though innocent (Wilson & Rule, 2015). Moreover, even if there are links between facial morphology and underlying personality traits, this does not imply a direct biological link between face cues and trustworthiness (Foo et al., 2021). Indeed, other people's opinions and expectations may lead a person to behave in ways that end up confirming the expectations – a phenomenon called self-fulfilling prophecy (Merton, 1948). Thus, trustworthiness impressions might be considered as “visual heuristics” that our visual system exploits to form sophisticated – but largely inaccurate - expectations about others' behaviour and intentions quickly (Siddique et al., 2022). In this view,

future research must aim to reduce facial biases in decision-making (Jaeger et al., 2020). In the same way, ensemble perception of trustworthiness impressions might also influence how much we perceive a member of a group as more or less trustworthy, in a biased way, given the effect that the social context can have on the perception of individuals (Carragher et al., 2021). The effect of the social context could be important for example in police line-ups, in which if the perceived trustworthiness of a suspect face is affected by the other surrounding faces, the eyewitness identification might be affected and a bias in decision making might manifest.

### 6.1. Future research directions and limitations

The neural mechanisms underlying ensemble perception impressions remain still to be explored. It has been proposed that, given that summary statistics occur at many levels of visual processing, it is unlikely that they consist of a unified mechanism, but rather, there might be distinct neural bases for ensemble perception of different visual characteristics (Whitney & Yamanashi Leib, 2018). Indeed, ensemble perception of social judgements, such as trustworthiness impressions, suggests that ensemble perception neural basis may also include higher-level visual areas. So far, previous research investigated the neural basis of summary representations of other facial features related to trustworthiness, such as emotional expressions (Im et al., 2017). Recently, an fMRI study that studied ensemble perception of emotional expressions found a different involvement of the ventral and dorsal visual streams in the perception of respectively individual faces or crowd facial expressions (Im et al., 2017), suggesting completely different neural path for the perception of emotional expressions for singular and multiple faces. A similar investigation has been also carried out for ensemble perception of face identity. To this regard, an EEG study found that the face-sensitive event-related potential (ERP) component N170, involved in the processing of singular faces, increased in magnitude when a group of faces was displayed compared when just a single face was presented (Puce et al., 2013). In line with this study, Roberts et al. (2019) examined summary representations of facial identity with EEG method and found a sensitivity to N170 ERP and P1 and P2 to multiple faces presented, even if the N170 for individual faces and groups showed different temporal dynamics. These results suggest that ERPs related to individual face processing are also involved in the perception of a group of faces for identity perception. Taken together, these studies gave rise to the first investigations of the neural basis of ensemble perception in trait impressions. Nevertheless, there are still open questions remaining. One interesting question is whether the summary statistics of trait impressions are represented in a single area or whether different stages of processing are involved. Another uncertain aspect is whether the mechanisms involved in the processing of traits from a singular face are also involved in the perception of multiple faces. In addition, we do not yet know whether the underlying mechanisms of trait ensemble percepts are shared with other facial trait impressions. Overall, our studies here open the door to this further research which could identify precisely where or how ensemble coding of trustworthiness impressions occurs at the neural level.

Second, investigating whether social cognition differences across participants influence their ability to form ensemble percepts of trustworthiness impressions with an individual difference approach would be highly interesting in future (especially given the growing interest in the field, see Sutherland et al., 2020). For example, variations in the social and cultural background might affect ensemble coding performance, together with differences in gender, and for typical and atypical cognitive development (Whitney & Yamanashi Leib, 2018).

In addition, our study used computer generated faces, in line with the majority of ensemble perception studies (Bai et al., 2015; de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2009; Sweeny & Whitney, 2014), but future work may wish to use real (and ideally, naturalistic) faces. In the future, moving beyond the set of computer-generated faces utilized here would allow to introduce more real-world noise in the ensemble perception task, resulting in more ecological validity. Furthermore, it would be possible to model some of the real-world noise in ensemble perception tasks by investigating the effects of presenting atypical or outliers faces in a group and varying the inter-item variance of the set of faces on the formation of summary statistics of trust. Using such controlled stimuli minimizes facial cues that observers exploit to form first impressions; for example, in this study, we choose only male-looking faces to control for gender cues that influence impressions of trustworthiness (e.g. Sutherland et al., 2015). This asymmetry was applied to avoid that the more trustworthy looking were the faces, the more they might start to convey a gender transformation and might appear particularly androgynous and female (Oliveira et al., 2020). Given that investigating gender was not the focus of the current study, we wanted to avoid this clear confounding variable. Future studies may also wish to see if ensemble perception effects interact with each other (e.g. gender and trustworthiness). On the other hand, computer-generated faces represent an ideal methodological option for a first approach to the investigation of a phenomenon not studied before, given that obtaining real faces that vary only in the level of trustworthiness is particularly difficult (and indeed perhaps logically impossible, given the nature of these judgements). In fact, the facial cues that drive trustworthiness impressions represent a combination and multitude of different facial attributes, from head tilt to eyebrow height (Vernon et al., 2014). Moreover, impressions themselves are inherently highly intercorrelated with other social judgements (Keles et al., 2021; Siddique et al., 2022). Here, using controlled stimuli allowed us to better target trustworthiness impressions, the area of interest of our research, while minimizing confounding variables. To this regard, the face stimuli used in the present study have been previously validated to lie on a vector that maximizes trust and minimizes other potential confounding factors such as gender and emotional expressions (Todorov et al., 2013). In addition, a recent study by Swe et al. (2020) found significant differences in the electrophysiological cortical responses to trustworthy and untrustworthy faces by using the same FaceGen stimuli presented in our study. Importantly, their study used EEG recording with the FPVS (Fast Periodic Visual Stimulation) paradigm, a procedure that does not require explicit judgements from participants, suggesting that these computer-generated stimuli were implicitly differentiated by observers for their facial trustworthiness cues. Similarly, we found that observers were able to compute the average level of trustworthiness from a group of faces both in an adjustment task and a rating task experiment, indicating that participants were actually able to process appearance cues of trustworthiness from our stimuli, which they efficiently translated into explicit judgements of trust.

## 7. Conclusion

In conclusion, a large literature has focused on trustworthiness impressions from individual faces, but far less effort has been devoted to investigating the trustworthiness impressions that observers form from groups of faces. Our findings filled this research gap, demonstrating that observers are able to extract an ensemble perception of trustworthiness impressions from a crowd of faces. These results provide insight into our ability to integrate facial information of trustworthiness from groups and contribute to the growing body of research on the important role of social context in trait impressions.

## Credit author statement

**Fiammetta Marini:** Conceptualization, Data curation, Software, Formal analysis, Visualization, Writing - original draft. **Clare A.M. Sutherland:** Conceptualization, Supervision, Methodology, Writing - review & editing. **Bārbala Ostrovska:** Data curation, Formal analysis. **Mauro Manassi:** Conceptualization, Methodology, Supervision, Writing - review & editing.

All authors contributed to the article and approved the submitted version.

## Data availability

Data have been uploaded on OSF at the link https://osf.io/ha295/.

## Acknowledgements

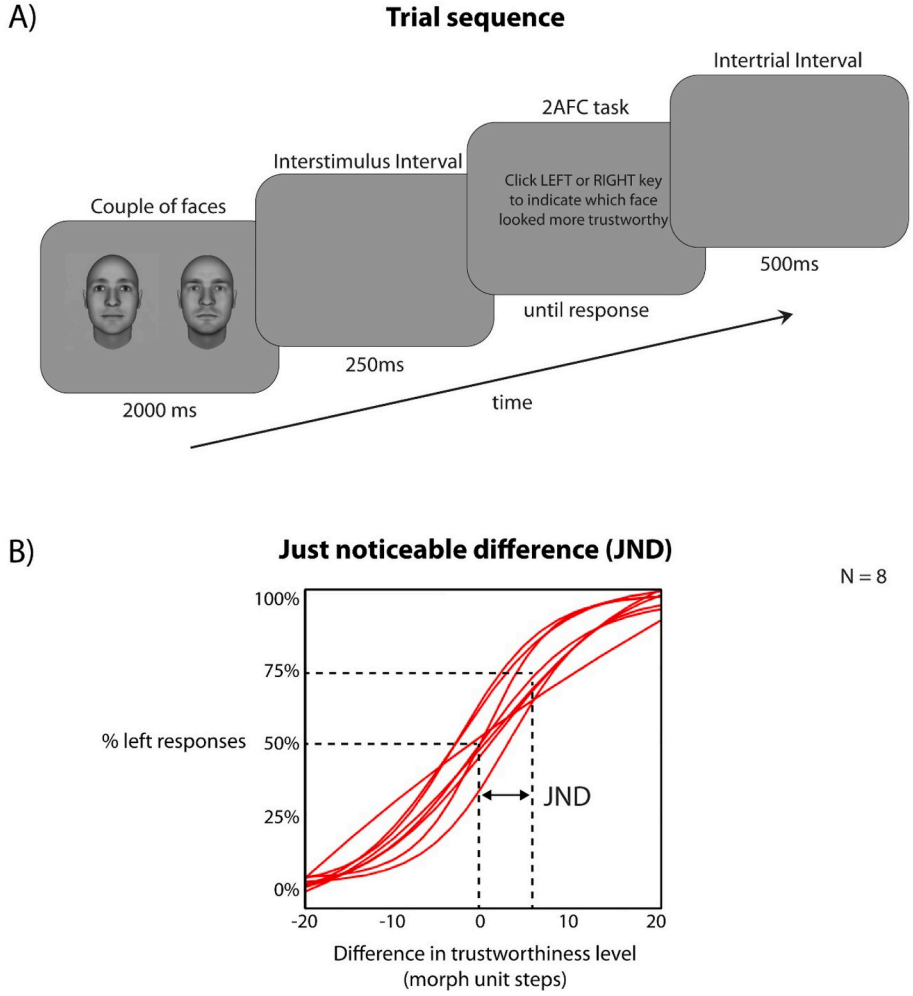
We thank R. Chakravarthi for his helpful advice and comments on an earlier version of the manuscript, and T. Burton for help in data collection. This research was supported by the Australian Research Council Discovery Project Grant 220101026. The data reported in the present manuscript were presented at ECVP conference 2022 and at Plymouth EPS meeting 2023. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

## Appendix A. Appendix

### Discriminability of Stimuli Experiment

This control experiment aimed to investigate whether participants were able to discriminate the trustworthiness level of the faces of the morph continuum used in this study. Eight participants (5 females, 3 males, M = 36 years, s.d. = 13.4 years) were included in this control study. The stimuli were the computer-generated face morphs used in the other Experiments.

On each trial, observers were presented with two faces from the morph continuum of the same identity on the left and on the right of the screen. Each couple of faces was composed by two morphs with a particular morph-unit distance in the morph continuum (−20; −15; −10; −5; 0; 5, 15, 20). After an ISI of 250 ms, participants performed a two-alternative forced choice task (2AFC) by indicating with the left or right arrow key which of the two faces was more trustworthy looking. After an ITI of 500 ms the next trial started. We tested if participants' ability to discriminate the more trustworthy looking morph face was a function of the distance in trustworthiness appearance between the two faces.




**Appendix Fig. 1.** Trial sequence and results of Experiment on discriminability of stimuli. (A) In each trial participants were shown a couple of randomly selected computer-generated faces of the same identity. After an ISI of 250 ms, they were asked to perform a two-alternative forced choice task (2AFC), by indicating which of the two faces displayed on the left and right of the screen was more trustworthy looking. After an ITI of 500 ms the next trial started. (B) Just noticeable difference (JND) analysis of the 2AFC data. The just noticeable difference for the face stimuli used across the experiments was calculated for each participant. The morph unit difference between the faces presented on the left and right sides of the screen in the 2AFC discrimination task is plotted on the x-axis. A positive x-axis value indicates that the morph face displayed on the left side of the screen was more trustworthy looking than the one presented on the right. The percentage of "left" responses is plotted on the y-axis. The average JND across participants was 6.56 morph unit. This result suggests that participants were able to discriminate the morph faces across the morph continuum for their trustworthiness level.

### Results

For each participant, we calculated the percentage of "left" responses (trials in which the morph face presented on the left was judged as more trustworthy looking than the one on the right). Then, we fitted a Cumulative Gaussian function on the percentage of "left" responses as a function of the morph-unit distance between the couples of morph faces. We calculated the just noticeable difference (JND) (the smallest face morph difference that allows to reliably distinguish two face morphs) by considering the distance between the 50th and 75th percentile of the Gaussian distribution. The mean JND across participants was 6.56 morph unit with standard deviation 2.21 morph unit. Given that the overall morph continuum was composed by 50 grayscale images created between the trustworthy and untrustworthy looking initial faces, these results suggest that participants were able to discriminate the trustworthiness level of morph faces across the morph continuum.

## Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105540.

## References


Afraz, S. R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, 48*(1), 42–54. https://doi.org/10.1016/J.VISRES.2007.10.028

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

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. https://doi.org/10.1016/J.TICS.2011.01.003

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Bai, Y., Leib, A. Y., Puri, A. M., Whitney, D., & Peng, K. (2015). Gender differences in crowd perception. *Frontiers in Psychology, 6*, 1300. https://doi.org/10.3389/FPSYG.2015.01300/BIBTEX

Baudouin, J. Y., & Tiberghien, G. (2004). Symmetry, averageness, and feature size in the facial attractiveness of women. *Acta Psychologica, 117*(3), 313–332. https://doi.org/10.1016/J.ACTPSY.2004.07.002

Carragher, D. J., Thomas, N. A., & Nicholls, M. E. R. (2021). The dissociable influence of social context on judgements of facial attractiveness and trustworthiness. *British Journal of Psychology, 112*(4), 902–933. https://doi.org/10.1111/BJOP.12501

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900. https://doi.org/10.1016/J.VISRES.2004.10.004

Epstein, M. L., & Emmanouil, T. A. (2021). Ensemble statistics can be available before individual item properties: Electroencephalography evidence using the oddball paradigm. *Journal of Cognitive Neuroscience, 33*(6), 1056–1068. Epstein, ML, 33(6), 1056–1068 https://direct.mit.edu/jocn/article-abstract/33/6/1056/98113. Epstein, ML, 33(6), 1056–1068.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389–1398. https://doi.org/10.1152/JN.00904.2010

Flowe, H. D. (2012). Do characteristics of faces that convey trustworthiness and dominance underlie perceptions of criminality? *PLoS One, 7*(6), Article e37253. https://doi.org/10.1371/JOURNAL.PONE.0037253

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716–1722. https://doi.org/10.1080/17470210902811249

Foo, Y. Z., Sutherland, C. A. M., Burton, N. S., Nakagawa, S., & Rhodes, G. (2021). *Accuracy in facial trustworthiness impressions: Kernel of truth or modern physiognomy? A meta-analysis.* https://doi.org/10.1177/01461672211048110

Furtak, M., Mudrik, L., & Bola, M. (2022). The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. *Cognition, 221*, 104983. https://doi.org/10.1016/j.cognition.2021.104983

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. https://doi.org/10.1037/XGE0000450

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11). https://doi.org/10.1167/9.11.1, 1–1.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753. https://doi.org/10.1016/J.CUB.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718–734. https://doi.org/10.1037/A0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*(7), 1825–1838. https://doi.org/10.3758/APP.72.7.1825

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin and Review, 18*(5), 855–859. https://doi.org/10.3758/S13423-011-0125-6

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In *From perception to consciousness: Searching with Anne Treisman.* Oxford University Press.

Han, L., Yamanashi Leib, A., Chen, Z., & Whitney, D. (2021). Holistic ensemble perception. *Attention, Perception, & Psychophysics, 83*(3), 998–1013. https://doi.org/10.3758/S13414-020-02173-1

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791–804. https://doi.org/10.1016/S0896-6273(02)01091-7

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2015). Global statistics are not neglected. *Journal of Vision, 15*(4), 7. https://doi.org/10.1167/15.4.7

Im, H. Y., Albohn, D. N., Steiner, T. G., Cushing, C. A., Adams, R. B., & Kveraga, K. (2017). Differential hemispheric and visual stream contributions to ensemble coding of crowd emotion. *Nature Human Behaviour, 1*(11), 828–842. https://doi.org/10.1038/s41562-017-0225-z

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology, 90*, Article 104004. https://doi.org/10.1016/J.JESP.2020.104004

Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of vision, 18*(3), 17-17 https://doi.org/10.1167/18.3.17.

Keles, U., Lin, C., & Adolphs, R. (2021). A cautionary note on predicting social judgments from faces with deep neural networks. *Affective Science, 2*(4), 438–454. https://doi.org/10.1007/S42761-021-00075-5

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision, 18*(9), 23. https://doi.org/10.1167/18.9.23

Korva, N., Porter, S., O'Connor, B. P., Shaw, J., & ten Brinke, L. (2013). Dangerous decisions: Influence of juror attitudes and defendant appearance on legal decision-making. *20*(3), 384–398. https://doi.org/10.1080/13218719.2012.692931

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8), 26. https://doi.org/10.1167/14.8.26

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology, 7*(SEP), 1332. https://doi.org/10.3389/FPSYG.2016.01332/BIBTEX

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications, 12*(1), 1–15. https://doi.org/10.1038/s41467-021-25500-y

Linke, L., Saribay, S. A., & Kleisner, K. (2016). Perceived trustworthiness is associated with position in a corporate hierarchy. *Personality and Individual Differences, 99*, 22–27. https://doi.org/10.1016/J.PAID.2016.04.076

Little, A. C., Jones, B. C., & Debruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 366*(1571), 1638–1659. https://doi.org/10.1098/RSTB.2010.0404

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281. https://doi.org/10.1038/36846

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision, 18*(8), 7. https://doi.org/10.1167/18.8.7

Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes: Serial dependence in ensemble representations. *Scientific Reports, 7*(1), 1–9. https://doi.org/10.1038/s41598-017-02201-5

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193. https://doi.org/10.2307/4609267

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology, 11*(4), 219–226. https://doi.org/10.1037/H0083717

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–788. https://doi.org/10.3758/PP.70.5.772

Oliveira, M., Garcia-Marques, T., Garcia-Marques, L., & Dotsch, R. (2020). Good to bad or bad to bad? What is the relationship between valence and the trait content of the big two? *European Journal of Social Psychology, 50*(2), 463–483. https://doi.org/10.1002/EJSP.2618

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*(11), 566–570. https://doi.org/10.1016/J.TICS.2014.09.007

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. https://doi.org/10.1073/PNAS.0805664105

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*(1), 128–133. https://doi.org/10.1037/A0014520

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/S13428-018-01193-Y

Puce, A., McNeely, M. E., Berrebi, M. E., Thompson, J. C., Hardee, J., & Brefczynski-Lewis, J. (2013). Multiple faces elicit augmented neural activity. *Frontiers in Human Neuroscience, 0*(MAY), 282. https://doi.org/10.3389/FNHUM.2013.00282/BIBTEX

Roberts, T., Cant, J. S., & Nestor, A. (2019). Elucidating the neural representation and the processing dynamics of face ensembles. *Journal of Neuroscience, 39*(39), 7737–7747. https://doi.org/10.1523/JNEUROSCI.0471-19.2019

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits: Short report. *Psychological Science, 19*(2), 109–111. https://doi.org/10.1111/j.1467-9280.2008.02054.x

Siddique, S., Sutherland, C. A. M., Palermo, R., Foo, Y. Z., Swe, D. C., & Jeffery, L. (2022). Development of face-based trustworthiness impressions in childhood: A systematic review and metaanalysis. *Cognitive Development, 61*, Article 101131. https://doi.org/10.1016/J.COGDEV.2021.101131

South Palomares, J. K., & Young, A. W. (2017). Facial first impressions of partner preference traits: Trustworthiness, status, and attractiveness. *Social Psychological and Personality Science, 9*(8), 990–1000. https://doi.org/10.1177/1948550617732388

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America, 115*(37), 9210–9215. https://doi.org/10.1073/PNAS.1807222115

Sun, J., & Chong, S. C. (2020). Power of averaging: Noise reduction by ensemble coding of multiple faces. *Journal of Experimental Psychology: General, 149*(3), 550. https://doi.org/10.1037/xge0000667

Sutherland. (2015). A basic guide to Psychomorph. https://aura.abdn.ac.uk/handle/2164/12696.

Sutherland, C. A. M., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., … Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences of the United States of America, 117*(19), 10218–10224. https://doi.org/10.1073/PNAS.1920131117

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105–118. https://doi.org/10.1016/J.COGNITION.2012.12.001

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *8*(5), 538–551. https://doi.org/10.1177/1948550617697176

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology*. https://doi.org/10.1111/BJOP.12583

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106*(2), 186–208. https://doi.org/10.1111/BJOP.12085

Swe, D. C., Palermo, R., Gwinn, O. S., Rhodes, G., Neumann, M., Payart, S., & Sutherland, C. A. M. (2020). An objective and reliable electrophysiological marker for implicit trustworthiness perception. *Social Cognitive and Affective Neuroscience, 15*(3), 337–346. https://doi.org/10.1093/SCAN/NSAA043

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., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science, 18*(4), 556–568. https://doi.org/10.1111/DESC.12239

Tiddeman, B., Burt, M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21*(5), 42–50. https://doi.org/10.1109/38.946630

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13*(4), 724–738. https://doi.org/10.1037/A0032335

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*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). For example, inferences of competence, based solely on fa-cial appearance. *Montepare & Zebrowitz, 27*(6), 813–833.

Twele, A. C., & Mondloch, C. J. (2022). The dimensions underlying first impressions of older adult faces are similar, but not identical, for young and older adult perceivers. *British Journal of Psychology*. https://doi.org/10.1111/BJOP.12568

Valentine, K. A., Li, N. P., Meltzer, A. L., & Tsai, M. H. (2020). Mate preferences for warmth-trustworthiness predict romantic attraction in the early stages of mate selection and satisfaction in ongoing relationships. *Personality and Social Psychology Bulletin, 46*(2), 298–311. https://doi.org/10.1177/0146167219855048

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences of the United States of America, 111*(32). https://doi.org/10.1073/PNAS.1409860111

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science, 25*(1), 230–235. https://doi.org/10.1177/0956797613497969

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

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. https://doi.org/10.3758/BRM.42.3.671

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology. General, 149*(10), 1811–1822. https://doi.org/10.1037/XGE0000745

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. https://doi.org/10.1037/XGE0000564