Journal of Experimental Social Psychology 91 (2020) 104047

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Facial trustworthiness predicts ingroup inclusion decisions

Ryan E. Tracy[a,*], John Paul Wilson[b], Michael L. Slepian[c], Steven G. Young[a,d]

[a] The Graduate Center, City University of New York, United States of America
[b] Montclair State University, United States of America
[c] Columbia University, United States of America
[d] Baruch College, City University of New York, United States of America

ARTICLE INFO

Keywords:
Facial trustworthiness
Ingroup inclusion
Person perception
Face perception

ABSTRACT

Perceivers tend to be reluctant to admit new members into their ingroups—unless there is some potential for prospective group members to provide value to the group. In the present research, we examine the effect of facial trustworthiness on ingroup inclusion decisions. Five studies demonstrate that facial trustworthiness exerts a powerful bottom-up perceptual cue that conveys this necessary "positive information," resulting in an increased likelihood of ingroup acceptance. This effect was first found for a homogenous sample of White male faces (Study 1), but was also found independent of sex (Study 2), and independent of race (Studies 3a, 3b, & 4), whereby facial trustworthiness influenced inclusion decisions more than salient aspects of group membership (i.e., sex and race).

1. Introduction

Group membership is an important dimension of a person's social world. Ingroups provide a variety of benefits that support wellbeing (e.g., protection, social support, self-esteem maintenance, resource sharing; e.g., Brewer, 2004; Correll & Park, 2005; Greenaway et al., 2015; Jetten et al., 2015), to the point that group affiliation has been argued to be a fundamental human need (e.g., Baumeister & Leary, 1995; Williams, 2009). Group affiliation also appears to buffer against existential anxiety, as prior research on group affiliation has found that participants primed with ideas of death demonstrated an increased level of ingroup favoritism and saw their ingroups as more entitative (Castano, Yzerbyt, Paladino, & Sacchi, 2002).

Consequently, group belonging has numerous effects on social cognition and motivation. Social Identity Theory argues that the self-concept is partly determined by group memberships (e.g., Tajfel & Turner, 1986) and that self-esteem is derived from membership in positively valenced groups (e.g., Abrams & Hogg, 1988; Leary, Tambor, Terdal, & Downs, 1995). Accordingly, people show ingroup favoritism in a variety of evaluative contexts, liking ingroup members more than outgroup members (Brewer, 1999; Sherman, Klein, Laskey, & Wyer, 1998; Turner, Brown, & Tajfel, 1979). Ingroup favoritism also biases behaviors, resulting in outcomes such as greater resource sharing and cooperation with ingroup members relative to outgroup members (e.g., Gaertner & Dovidio, 2000; Tajfel & Turner, 1986).

Notably, these effects occur when group boundaries are sharply defined and based on highly salient characteristics like race, sex, and socioeconomic status (e.g., Stephan & Stephan, 2000), but also when intergroup distinctions are largely inconsequential (e.g., university rivalries) or even arbitrary (e.g., minimal group paradigms, see Brewer, 1979; Tajfel, Billig, Bundy, & Flament, 1971). Such evidence reveals a fundamental motivation to belong to and identify with social groups.

2. Ingroup overexclusion

Given the importance of group membership, it is not surprising that people are sometimes selective when determining who should enjoy the benefits of ingroup membership. Indeed, people often reserve ingroup membership for targets who possess the positive qualities believed to be typical of fellow ingroup members. Conversely, those viewed as lacking on some dimension are excluded from the group (more commonly than they are included), a phenomenon termed "ingroup overexclusion" (e.g., Leyens & Yzerbyt, 1992; Yzerbyt, Leyens, & Bellour, 1995). As with other intergroup biases, ingroup overexclusion can occur in a variety of contexts.

Leyens and Yzerbyt's (1992) original demonstration of the effect involved a sample of Flemish and Walloon students reading descriptions of novel individuals and making decisions about which targets belonged to which group. These are two distinct regional groups in Belgium that have a history of ingroup favoritism and outgroup

* Corresponding author.
E-mail address: rtracy1@gradcenter.cuny.edu (R.E. Tracy).

https://doi.org/10.1016/j.jesp.2020.104047
Received 9 April 2020; Received in revised form 17 August 2020; Accepted 19 August 2020

derogation between them (Leyens & Yzerbyt, 1992; Yzerbyt et al., 1995). In this original work, both groups of participants systematically categorized more targets as outgroup than ingroup members, reserving ingroup categorization only for targets with highly positive descriptions.

More recent research finds that similar effects emerge in racial categorization. When White participants categorize Black/White racially ambiguous faces as either racial ingroup or outgroup members, intermixed with White faces and Black faces, ingroup categorizations are most commonly granted to unambiguously White faces (Knowles & Peng, 2005). Those who are not clearly in the ingroup are excluded. Racially ambiguous targets are categorized as outgroups particularly when displaying negative facial expressions (e.g., anger; Peery & Bodenhausen, 2008), thereby allowing people to maintain a more "positive" perception of their ingroup.

Aside from conflict-laden contexts, Rubin and Paolini (2014) used an arbitrary group distinction akin to a minimal group paradigm (Tajfel et al., 1971) and found that when people believed the ingroup to be positive (and the outgroup negative), people tended to include relatively few targets in their novel ingroup. These results demonstrate that the ingroup overexclusion effect is likely a product of ingroup favoritism, borne out of a desire to maintain the ingroup's positive image and protect the ingroup from infiltration by undesirable others (e.g., Hutchison & Abrams, 2003; Leyens & Yzerbyt, 1992; Rubin & Paolini, 2014).

This tendency is functional. Any general orientation to see oneself or one's group as positive, relative to outgroup members, would help maintain positive beliefs about the ingroup (and consequently the self) by maximizing psychological distinctions between the ingroup and outgroup (Brewer, 1999). Moreover, ingroup overexclusion could serve a practical function as well, as group cohesion and stability can be disrupted by bad actors (e.g., Hutchison, Abrams, & Christian, 2007; Kurzban & Leary, 2001; O'Boyle, Forsyth, & O'Boyle, 2011), leading people to set high standards before granting ingroup status to others, lest the group itself be threatened.

3. Who is granted ingroup membership?

As the ingroup overexclusion effect demonstrates, ingroup categorizations are not liberally granted but are instead reserved for those described in positive ways (Leyens & Yzerbyt, 1992). Moreover, even without descriptions conveying positive characteristics, similar effects emerge when information is conveyed through visual routes. For example, ingroup membership is often granted to targets unambiguously possessing physical characteristics of group membership (e.g., race; Knowles & Peng, 2005; see also Thorstenson, Pazda, Young, & Slepian, 2019), to targets perceived as being high on competence and warmth (Ponsi, Panasiti, Scandola, & Aglioti, 2016), or even simply displaying happy facial expressions (Peery & Bodenhausen, 2008). Other research has found that fluently processed faces are afforded ingroup membership regardless of race (Claypool, Housley, Hugenberg, Bernstein, & Mackie, 2012), suggesting that the positivity associated with fluent (i.e., easy) processing mitigates the tendency to overexclude targets.

These findings are especially relevant to the current research, as they illustrate that various aspects of face processing can affect inclusion/exclusion tendencies. However, this prior work has focused on group categorizations along racial dimensions (e.g., Black/White category decisions) and with perceptually obvious cues to group membership (e.g., easily identified facial expressions; variations in skin tone). In other words, from prior work, it is unclear whether the facial features had any influence on group categorization, or if instead, it was simply the clear social categories conveyed by those faces that had their influence (cf. Macrae & Martin, 2006; Maddox, 2004). In fact, to our knowledge, only one study has looked at how facial trait cues influence ingroup inclusion decisions. Ponsi et al. (2016) found that participants' ratings of warmth and competence correlated with their ingroup decisions. Having participants make both the psychological and the ingroup ratings, however, leaves open the possibility that the correlations exist at the conceptual level, rather than the facial feature level (see Stolier, Hehman, Keller, Walker, & Freeman, 2018). Indeed, in this prior study the faces were confined to images that were pre-rated to be neutral in trustworthiness.

Although judgments of facial trustworthiness are subject to idiosyncratic perceiver impressions, consensus is also relatively high when it comes to judging trustworthiness from faces (Hehman, Sutherland, Flake, & Slepian, 2017). Even aside from accuracy, such consensus has important psychological consequences (Slepian & Ames, 2016). Moving beyond prior work, we thus leveraged consensus in judging trustworthiness by presenting participants with faces pre-rated as high or low in trustworthiness (Experiments 1–3) or varying continuously along this dimension (Experiment 4). Accordingly, in contrast to prior work, we examine whether relatively subtle facial cues can shift the threshold for group inclusion. Specifically, we examined whether perceived facial trustworthiness conveys enough positive information to be included into the ingroup. We focus on facial trustworthiness for several reasons enumerated below.

Facial trustworthiness is argued to be one of two key dimensions of person perception (along with dominance, e.g., Oosterhof & Todorov, 2008). Moreover, although there is individual variation in trait judgments from faces (Hehman, Sutherland, et al., 2017; Kramer, Mileva, & Ritchie, 2018), trustworthiness judgments nevertheless do show high levels of interrater agreement (e.g., Rule, Krendl, Ivcevic, & Ambady, 2013). The tendency to differentiate faces according to trustworthiness emerges in as little as 34 ms (e.g., South Palomares & Young, 2017; Todorov, Mende-Siedlecki, & Dotsch, 2013; Todorov, Pakrashi, & Oosterhof, 2009) or 100 ms during more consequential situations (i.e., during trust games; De Neys, Hopfensitz, & Bonnefon, 2017; see also Willis & Todorov, 2006). Relative to faces considered untrustworthy, trustworthy faces tend to have larger, rounder eyes, a larger mouth with slightly upturned corners, a larger forehead, and a shorter chin (e.g., Kleisner, Priplatova, Frost, & Flegr, 2013).

Faces perceived as trustworthy enjoy numerous social benefits. For example, trustworthy faces also potentiate affiliative and prosocial approach behaviors (Slepian, Young, & Harmon-Jones, 2017; Slepian, Young, Rule, Weisbuch, & Ambady, 2012). Additionally, van't Wout and Sanfey (2008) found that participants made greater offers to partners with trustworthy faces than those with untrustworthy faces in an ultimatum game. Real-world effects have been documented as well, as facial trustworthiness is associated with increased pay (Fruhen, Watkins, & Jones, 2015) and assumptions of fairness in managerial decision making (Holtz, 2014). The inverse of these findings is that untrustworthy defendants are more likely to receive harsh sentences for crimes (Sutherland, Cojocariu, Day, & Hehman, 2020; Wilson & Rule, 2016). Thus, understanding how rapidly made and consensually agreed upon judgments of facial trustworthiness influence group categorizations is an important, but heretofore unexplored area of research.

4. Facial trustworthiness and target race and sex

While facial trustworthiness exerts powerful impacts on how people are evaluated and treated, so too do other structural face cues, most notably race and sex. These two variables have been argued to dominate early stages of person perception (Bruce & Young, 1986; Ito & Urland, 2003) and form basic categories into which faces are rapidly allocated (e.g., Levin, 1996, 2000; Wild et al., 2000). However, how these two factors interact with variations in facial trustworthiness to predict overexclusion effects is an important yet largely unexamined question. Nevertheless, several outcomes are plausible based on research to date. Prior work has found that categorical information (e.g., race, sex) is extracted more efficiently than individuating information (Cloutier, Mason, & Macrae, 2005). This would suggest that when making ingroup inclusion decisions, participants should be

predominantly guided by a target's categorical facial information rather than their more individuating features that correspond with perceived traits (e.g., facial trustworthiness). Alternatively, more target-specific facial information can also be quickly processed and does affect social judgments (e.g., phenotypical face traits; Blair, Judd, Sadler, & Jenkins, 2002; Maddox, 2004), and facial trustworthiness is processed so efficiently as to perhaps have an obligatory and spontaneous effect on social perception (Klapper, Dotsch, van Rooij, & Wigboldus, 2016).

Considering face cues to sex first, prior research affords some tentative predictions. For example, a large literature on gender stereotypes shows that woman are believed to be especially communal and interpersonally warm (Brown, Phills, Mercurio IV, Olah, & Veilleux, 2018; Fiske, Cuddy, & Glick, 2007). Moreover, recent research has found that many first impressions of female faces are driven by perceptions of trustworthiness and/or warmth (cf. male faces, which are driven by both trustworthiness and dominance perceptions; Oh, Dotsch, Porter, & Todorov, 2020). In general, violating these stereotypic expectations is judged negatively (e.g., Flannigan, Miles, Quadflieg, & Macrae, 2013). More specific to face processing research, counter-stereotypical female faces (e.g., those with masculine features or expressing anger) are rated negatively relative to stereotype-congruent female targets, partly because these female faces defy expectations of warmth and prosociality (e.g., Bayet et al., 2015; Sutherland, Young, Mootz, & Oldmeadow, 2015). Although untrustworthy faces are not dominant or masculine per se, they nevertheless may appear angry (e.g., Said, Sebe, & Todorov, 2009; Slepian & Carr, 2019) and convey negative intentions that contradict female stereotypes. If so, then would predict that untrustworthy female faces would be over-excluded even more often than untrustworthy male faces. A competing possibility, however, is that the stereotypical prosociality linked with women will override any cues to (un)trustworthiness in the faces, leading to relatively high rates of ingroup inclusion for both trustworthy and untrustworthy females faces relative to male faces. That said, prior research has shown that the preconscious processing of facial trustworthiness is no different for male and female faces (Wang, Tong, Shang, & Chen, 2019). As facial trustworthiness is equally identifiable in both male and female faces, it is also possible that facial trustworthiness will matter more for participants' inclusion decisions than does target sex.

With respect to race, it is possible that this highly salient and potent cue to group membership (e.g., Levin, 1996; Peery & Bodenhausen, 2008) will outweigh trustworthiness when participants render ingroup decisions. Specifically, a largely White sample viewing Black faces might over-exclude Black targets regardless of facial cues to trustworthiness, due to general evaluative racial biases. Prior arguments suggest that perceivers typically attend to cues of skin color and racial phenotypicality in other-race faces, specifically those of Black targets (e.g., Levin, 1996, 2000; Maddox, 2004). Should this indeed be the case, we would expect facial trustworthiness to have a minimal effect on ingroup inclusion for Black targets, who would be largely excluded from the ingroup (by primarily non-Black participants) as a result of race-specifying facial cues, rather than trait-specifying features. This could either result in a main effect of face race or an interaction between facial trustworthiness and race where trustworthiness only affects group categorization of White faces (with Black faces generally excluded from the ingroup). However, it is also possible that trustworthiness will have similar effects for Black and White faces. Indeed, recent work finds that both Black and White faces pre-rated to appear trustworthy are rated as “nice” when children and adult White participants are afforded time to make explicit ratings (Charlesworth & Banaji, 2019). Such findings suggest that, when allowed to correct any biasing effect of race, face-trait cues are processed and inform judgments of same- and other-race faces alike (see also Cassidy & Krendl, 2018; Wilson, Young, Rule, & Hugenberg, 2018), which would produce a main effect of facial trustworthiness on exclusion.

5. The current research

Our overarching hypothesis is that perceptions of facial trustworthiness will offer enough positive information to surpass the threshold necessary for participants to determine that a target face is an ingroup member. We tested this hypothesis using a minimal group paradigm. To our knowledge, only one prior study has used groups that are not preexisting ethnic or salient racial identities (Rubin & Paolini, 2014) to test the threshold of ingroup inclusion. In that study, the new minimal groups were manipulated to be positive or negative, in essence recreating the valence already applied to existing category divides. This raises the possibility that minimal group manipulations may not yield a robust overexclusion effect unless the new outgroup is also made to be perceived in a negative light.

In the present work, we leave even this information ambiguous, and instead manipulate subtle features of the individual targets for categorization. The minimal group paradigm has been used to reliably demonstrate intergroup bias via ingroup favoritism (e.g., Hertel & Kerr, 2001; Lemyre & Smith, 1985; Otten & Moskowitz, 2000; Tajfel et al., 1971), thus allowing for the possibility that even with no information about the two groups (i.e., the ingroup and outgroup), features of its potential members may determine who is allowed into the ingroup.

In our first experiment, after a minimal group induction, we presented participants with White male faces (thus holding race and sex constant) that were pre-rated to vary in trustworthiness. We asked participants to indicate which faces were fellow ingroup members vs. outgroup members. Follow-up experiments expanded the scope to include target sex (Study 2) and finally target race (Studies 3a, 3b, & 4). Study 4 included computer-generated faces existing along a range of facial trustworthiness to allow for a more fine-grained analysis of ingroup inclusion decisions along a continuum of facial trustworthiness, from very untrustworthy to very trustworthy.

The current approach allows us to examine several theoretically and practically meaningful questions beyond how trustworthiness influences ingroup inclusion decisions. We also examine how additional (and more visually arresting) facial information like sex and race interact with morphological cues to trustworthiness. Collectively, these experiments integrate several research traditions (e.g., intergroup categorization, face perception, identity) to examine how facial trustworthiness exerts main and interactive effects on whether novel faces are included or excluded from newly formed, minimalistic ingroups. We report all measures, manipulations, and exclusions in these studies. Sample size was determined before any data analysis was conducted. In all studies, we sought to recruit 100 participants via Amazon's Mechanical Turk (with 80% power, this sample size can detect an effect size $r = 0.276$ at $\alpha = 0.05$; calculated with R package “pwr”; Champely, 2020; see Fritz, Morris, & Richler, 2012). Occasionally, participants did not enter a completion code, resulting in some studies having sample sizes over the target N. We analyzed all collected data when this happened.

6. Study 1

6.1. Method

6.1.1. Participants

In Study 1, data were collected from 108 online participants. Data from four participants were removed for duplicated response IDs and data from an additional two participants were removed for displaying nonvariance (i.e., responding with the same value across all trials). This left us with a final sample of 102 participants ($N_{males} = 54$, $N_{females} = 48$, 80.4% White, 9.8% Asian, 5.9% Black, 2.9% Latino/a, and 0.98% other).

6.1.2. Stimuli

Stimuli consisted of 30 White male faces that were pre-rated by a

sample of independent raters to be either trustworthy or untrustworthy (see Slepian et al., 2012). Faces were matched for saturation and luminance and did not differ on ratings of facial attractiveness or dominance, thus isolating trustworthiness as the key difference between face sets. Faces also displayed a neutral expression.

6.1.3. Procedure

Participants were first subjected to a minimal group dot-estimation task where they responded to a series of images displaying an array of dots (see Tajfel et al., 1971). Each image was displayed for 3 s, followed by a prompt asking participants to estimate how many dots they saw in the image. After three trials, participants were given randomly generated feedback that categorized them as overestimators or underestimators. This feedback only informed them that they tended to either overestimate or underestimate the number of dots in the images and contained no information that was suggestive of population guessing characteristics.

For the main task, participants categorized each faceindividually, indicating whether the presented face was an overestimator or an underestimator.Participants were given unlimited time to make their responses, with instructions to go with their "gut" intuition. Following the 30 group categorization trials, participants responded to a series of demographic questions and were debriefed.

6.1.4. Analysis

Participants' inclusion responses were scored as a binary variable. When a face was categorized as ingroup, the response was scored as 1 (i.e., categorized as an overestimator for those labelled as *overestimators*, or underestimator for those labelled as *underestimators*). When a face was categorized as outgroup, the response was scored as 0 (i.e., categorized as an underestimator for those labelled as *overestimators*, or overestimator for those labelled as *underestimators*).

These responses were then included in a generalized linear mixed model (GLMM) that included the predictor of perceived facial trustworthiness (trustworthy = 0.5, untrustworthy = −0.5) to determine the probability that a trustworthy (versus untrustworthy) face would be included in participants' ingroups.

Following this, we took the proportion of perceived trustworthy and untrustworthy faces counted as ingroup members to test these inclusion rates against chance. This analysis was done to see if participants demonstrated a systematic bias to include trustworthy faces at above-chance rates while including untrustworthy faces at below-chance rates. Chance-level comparisons (i.e., 50% inclusion) have been noted as the demarcation line of an overexclusion effect in prior studies on ingroup inclusion decisions (e.g., Castano, Yzerbyt, Bourguignon, & Seron, 2002; Claypool et al., 2012). This analysis procedure was repeated for all studies included in the manuscript.

6.2. Results

To analyze the data, we fit a GLMM specifying random intercepts for both participants and stimuli using the "lme4" package in R (Bates, Maechler, Bolker, & Walker, 2015), also using the "lmerTest" package to estimate *p*-values (Kuznetsova, Brockhoff, & Christensen, 2017). This cross-classified model resulted in a singular fit owing to zero variance accounted for by the stimuli, so the random factor for stimuli was dropped from the model in favor of one fitting a random slope for facial trustworthiness on inclusion for each participant.

6.2.1. Ingroup categorizations

We first examined the relative proportion of trustworthy and untrustworthy faces categorized as ingroup members. This model revealed the predicted effect of perceived facial trustworthiness, $b = 0.43$, $SE = 0.12$, $z = 3.67$, $p < .001$, $OR = 1.54$, 95% CI = [1.22, 1.94], where targets with trustworthy faces were 54% more likely to be included into the ingroup than those with untrustworthy faces.

6.2.2. Comparing to chance

We next compared the inclusion rates for trustworthy and untrustworthy faces against chance. Here, we found that trustworthy faces were included at rates that did not significantly differ from chance ($M = 0.52$, $SD = 0.18$, 95% CI_{mean} [0.49, 0.56]), $t(101) = 1.27$, $p = .21$, $d = 0.18$, while untrustworthy faces were included at rates significantly below chance ($M = 0.42$, $SD = 0.18$, 95% CI_{mean} = [0.39, 0.46]), $t(101) = -4.22$, $p < .001$, $d = -0.59$.

6.3. Discussion

In summary, Study 1 showed the predicted effect of facial trustworthiness on ingroup inclusion decisions. Trustworthy-looking faces were included as ingroup members at a significantly increased likelihood over their untrustworthy-looking counterparts. Moreover, these effects were not merely relative differences in ingroup categorization, as participants included untrustworthy faces at rates below chance. These results provide initial evidence that perceived facial trustworthiness conveys enough positive information to grant someone membership into the ingroup.

To further explore this effect, we designed Study 2 with two objectives in mind. First, we sought to replicate the results of Study 1. Second, we sought to test whether the facial trustworthiness effect would interact or overcome another salient identity: target sex.

7. Study 2

In Study 1, faces that had trustworthy appearances were judged more often as ingroup, whereas faces that had untrustworthy appearances were judged more often as outgroup. Study 2 served as a replication and extension of Study 1 by introducing a more salient group dimension (target sex) to examine the robustness of this facial trustworthiness effect.

7.1. Method

7.1.1. Participants

Data were collected from a sample of 140 participants via MTurk. After excluding participants who failed the attention check assessing which group they were assigned to, we were left with a final sample of 125 participants ($M_{age} = 33.92$, $SD_{age} = 9.80$; $N_{men} = 65$, $N_{women} = 59$, $N_{other} = 1$; 70.4% White, 10.4% Black, 10.4% Asian, 5.6% Latino/a, 3.2% Other).

7.1.2. Stimuli

Stimuli used for this study were taken from the Oslo Face Database (Chelnokova et al., 2014). 20 White male and 20 White female faces were selected for this study, where, for each sex, 10 faces were rated to be highly trustworthy and 10 were rated to be untrustworthy, yielding a balanced 2 × 2 factorial design. Ratings were drawn from norming data included in the download of the face database. These ratings were entered into a 2 × 2 ANOVA that included trustworthiness (i.e., trustworthy, untrustworthy) and target sex (male, female) as independent variables. This model found only a main effect of facial trustworthiness, $F(1, 36) = 326.03$, $p < .001$. Faces categorized as *trustworthy* ($M = 6.30$, $SD = 0.19$) were rated as significantly more trustworthy than those categorized as *untrustworthy* ($M = 4.46$, $SD = 0.24$), $t(36) = 26.53$, $p < .001$. Neither a main effect of sex nor a sex × group interaction emerged ($Fs < 1.5$, $ps > 0.24$). Male and female faces pre-rated as *trustworthy* did not differ in this metric ($M_{males} = 6.36$, $SD_{males} = 0.22$; $M_{females} = 6.24$, $SD_{females} = 0.15$), $t(18) = 1.41$, $p = .18$. Similarly, male and female faces pre-rated as *untrustworthy* were also equally untrustworthy ($M_{males} = 4.47$, $SD_{males} = 0.25$; $M_{females} = 4.47$, $SD_{females} = 0.24$), $t(18) = -0.23$, $p = .82$. Faces displayed neutral expressions and were again matched for luminance, saturation, attractiveness, and dominance to remove plausible

confounds.

7.1.3. Procedure

The procedure for Study 2 was the same as that for Study 1. Participants first underwent the minimal group paradigm dot estimation task where they received bogus feedback categorizing them as either overestimators or underestimators, then they were given the main task. Participants were presented with a single face for each trial responding to a prompt asking them which group they believed each target belonged to. After 40 trials, participants were thanked and debriefed.

7.2. Results

7.2.1. Ingroup categorizations

As in Study 1, we fit a GLMM to assess the likelihood of a face being included as an ingroup member. This model specified a random slope for facial trustworthiness on inclusion for each participant, this time including target sex as a predictor (males = 0.5, females = −0.5) along with facial trustworthiness (trustworthy = 0.5, untrustworthy = −0.5). This model yielded only a main effect of perceived facial trustworthiness, $b = 0.27$, $SE = 0.11$, $z = 2.53$, $p = .01$, $OR = 1.30$, 95% CI [1.06, 1.61]. There was no main effect of target sex, $b = -0.04$, $SE = 0.06$, $z = 0.60$, $p = .55$. Furthermore, we found no interaction between target sex and trustworthiness, $b = -0.01$, $SE = 0.12$, $z = -0.11$, $p = .91$. Independent of target sex, we found that trustworthy faces were 30% more likely to be included than untrustworthy faces.

Because we had a nearly equal number of male and female participants, we conducted additional analyses that examined whether participants were more likely to include targets who were the same or different sex. We added this factor into a model (same sex = 0.5, different sex = −0.5) that also included facial trustworthiness as a predictor. This model again found only a main effect of facial trustworthiness, $b = 0.27$, $SE = 0.11$, $z = 2.53$, $p = .01$, $OR = 1.31$, 95% CI [1.06, 1.61]. There was no main effect of same/different sex, $b = -0.05$, $SE = 0.06$, $z = -0.90$, $p = .37$. Furthermore, we found no interaction between same/different sex and trustworthiness, $b = 0.10$, $SE = 0.12$, $z = 0.89$, $p = .37$.[1]

7.2.2. Comparing to chance

As in Study 1, we next compared inclusion rates against chance. We first collapsed across target sex to examine the effect of facial trustworthiness. This analysis found that trustworthy faces were included at rates significantly higher than chance ($M = 0.53$, $SD = 0.15$, 95% CI_{mean} [0.51, 0.56]), $t(124) = 2.60$, $p = .01$, $d = 0.33$, while untrustworthy faces were included at rates marginally below chance ($M = 0.47$, $SD = 0.18$, 95% CI_{mean} [0.44, 0.50]), $t(124) = -1.69$, $p = .09$, $d = -0.21$.

We next collapsed across facial trustworthiness to examine the effects of target sex against chance. This analysis found that male faces were included at rates no different than chance ($M = 0.50$, $SD = 0.15$, 95% CI_{mean} [0.47, 0.53]), $t(124) = 0.02$, $p = .99$, $d = 0.002$. Similarly, inclusion rates for female faces did not differ from chance ($M = 0.51$, $SD = 0.15$, 95% CI_{mean} [0.48, 0.54]), $t(124) = 0.61$, $p = .55$, $d = 0.08$.

Finally, we looked at inclusion rates for same-sex and other-sex faces, i.e., coding whether the target sex was congruent with participant sex (1 participant did not identify as female or male). Same-sex faces were not included at rates significantly different from chance ($M = 0.50$, $SD = 0.15$, 95% CI_{mean} [0.47, 0.52]), $t(123) = -0.12$, $p = .91$, $d = -0.02$,[2] as was the case for other-sex faces ($M = 0.51$, $SD = 0.14$, 95% CI_{mean} [0.49, 0.53]), $t(124) = 0.79$, $p = .43$, $d = 0.10$.

7.3. Discussion

The results from Study 2 replicate the effect found in Study 1, where trustworthy-looking targets were systematically counted as ingroup members more than untrustworthy-looking targets. The lack of an interaction between target sex and facial trustworthiness suggests that the effect of facial trustworthiness does not depend on more obviously salient face characteristics such as sex. Furthermore, this effect appears to be independent of whether participants were of the same or different sex as the target, given that the same/different sex factor also did not interact with facial trustworthiness. Instead, it was facial trustworthiness that seemed to sway ingroup inclusion decisions. Here is a situation where features seem to preside over categorical information (cf. Cloutier et al., 2005). Instead of relying on the sex category information provided by the face, our data suggest that participants relied more on the individuating information conveyed by facial trustworthiness.

Pre-existing stereotypes see women as warm and trustworthy (Sutherland et al., 2015; Sutherland, Oldmeadow, & Young, 2014). Hence, if the present effects were bound to stereotypes of existing category divides (e.g., men vs. women), we might see participants allowing more women into their newly formed minimal ingroups. The lack of this effect suggests that stereotypes of salient social categories (of which a target is clearly in) do not impinge on these results. Rather, it was other and more subtle features of the targets that determined whether they were allowed into the ingroup: facial trustworthiness.

8. Study 3a

The purpose of Study 3a (and its direct replication, Study 3b) was to explore whether target race interacted with facial trustworthiness when participants made ingroup inclusion decisions. Target race was chosen as a final factor given that it is another salient aspect of group distinction that typically results in stark differences in ingroup inclusion and has been implicated in prior work examining ingroup overexclusion (e.g., Knowles & Peng, 2005; though see Claypool et al., 2012). Although we expected that race might exert a main effect on inclusion decisions, such that participants might be less likely to include Black targets in the ingroup than White targets, we did not predict that race would moderate the effect of facial trustworthiness, given that target sex did not serve as a moderator to Study 2's results. However, competing possibilities exist. For example, race could interact with facial trustworthiness, such that Black targets are excluded regardless of facial trustworthiness, whereas trustworthiness influences the inclusion of White faces (i.e., replicating the effects seen in Studies 1 and 2).

8.1. Method

8.1.1. Participants

We collected data from 107 individuals via MTurk, removing the responses of 15 individuals who failed our attention check question. Responses were removed from an additional four participants for demonstrating nonvariance (i.e., responding with the same value across all trials). This left us a final sample of 88 participants ($N_{males} = 55$, $N_{females} = 33$, 79.5% White, 3.4% Black, 9.1% Asian, 5.7% Latino/a, 2.3% Other).

[1] We conducted additional analyses that entered the continuous mean ratings of trustworthiness from the OFD for each target face into the same GLMM used for the dichotomized trustworthiness variables. This analysis was repeated for the studies involving Black White faces. All models yielded results similar to the ones noted above and can be found in the Supplementary Materials.

[2] One participant did not identify as either male or female, and so could not be counted as being "same-sex." This person was instead counted as "other-sex" by default, hence the disparity in the degrees of freedom.

8.1.2. Stimuli

Faces used for this study were taken from the Eberhardt Face Database (Eberhardt, Goff, Purdie, & Davies, 2004). We selected the 10 most trustworthy faces and the 10 least trustworthy faces per each category (20 White and 20 Black male faces in total), yielding a balanced 2 × 2 design. As in Study 2, we entered the trustworthiness ratings from the database into a 2 × 2 ANOVA comparing the effects of race (Black, White) and trustworthiness (trustworthy, untrustworthy). This model again found only a main effect of facial trustworthiness, $F(1, 36) = 352.73$, $p < .001$. Faces categorized as *trustworthy* ($M = 4.28$, $SD = 0.16$) were rated as significantly more trustworthy than those categorized as *untrustworthy* ($M = 2.65$, $SD = 0.22$), $t(36) = 26.33$, $p < .001$. Neither a main effect of race nor a race × trustworthiness interaction emerged ($Fs < 1$, $ps > 0.70$). White and Black faces pre-rated as trustworthy were equivalent on this metric, ($M_{White} = 4.30$, $SD_{White} = 0.15$; $M_{Black} = 4.27$, $SD_{Black} = 0.15$), $t(18) = 0.48$, $p = .64$, as were White and Black faces pre-rated as untrustworthy ($M_{White} = 2.68$, $SD_{White} = 0.29$; $M_{Black} = 2.62$, $SD_{Black} = 0.29$), $t(18) = 0.61$, $p = .55$. Faces again displayed neutral expressions and were matched for luminance and saturation and perceptions of attractiveness and dominance to remove any potential moderators.

8.1.3. Procedure

The procedure for Study 3 mirrored that of Study 2, with the exception of the race and sex of the faces used in the study. Participants first underwent the minimal group paradigm dot estimation task where they received bogus feedback categorizing them as either overestimators or underestimators, then they were given the main group categorization task. Again, participants were presented with a single face for each trial responding to a prompt asking them to which group they believe this person belonged. After 40 trials, participants were thanked and debriefed.

8.2. Results

8.2.1. Ingroup categorizations

As in our previous studies, we analyzed the data using a GLMM specifying random intercepts for participants with a random slope for facial trustworthiness on inclusion. This model found a main effect of face race, $b = 0.28$, $SE = 0.07$, $z = 4.00$, $p < .001$, $OR = 1.32$, 95% CI [1.15, 1.52], indicating that White faces were 32% more likely to be included than Black faces. In addition, we found a main effect of facial trustworthiness, $b = 0.32$, $SE = 0.12$, $z = 2.59$, $p = .01$ $OR = 1.38$, 95% CI [1.08, 1.76], where trustworthy faces were 38% more likely to be included in the ingroup than their untrustworthy counterparts. The race × trustworthiness interaction was not significant, $b = 0.09$, $SE = 0.14$, $z = 0.67$, $p = .51$.

8.2.2. Comparing to chance

We first collapsed across face race to compare inclusion rates to chance, as per the earlier studies. This showed that participants tended to include trustworthy faces at rates nonsignificantly different from chance ($M = 0.52$, $SD = 0.15$, 95% CI_{mean} [0.49, 0.56]), $t(87) = 1.57$, $p = .12$, $d = 0.24$, while including untrustworthy faces at significantly below-chance rates ($M = 0.45$, $SD = 0.19$, 95% CI_{mean} [0.41, 0.49]), $t(87) = -2.35$, $p = .02$, $d = -0.35$.

We next conducted these analyses by target race, collapsing across facial trustworthiness given the lack of interaction between face race and perceived trustworthiness. Inclusion rates for White faces did not differ from chance ($M = 0.52$, $SD = 0.15$, 95% CI_{mean} [0.49, 0.55]), $t(87) = 1.40$, $p = .17$, $d = 0.21$, while Black faces were included at significantly below-chance rates ($M = 0.46$, $SD = 0.19$, 95% CI_{mean} [0.42, 0.49]), $t(87) = -2.16$, $p = .03$, $d = -0.33$.

8.3. Discussion

The data from Study 3a leads to two conclusions. First, confirming the previous studies, facial trustworthiness is a cue that perceivers use to make ingroup inclusion decisions. Second, the effect of trustworthiness on inclusion does not appear to differ based on target race. Although White faces were included as ingroup members more than Black faces, race did not moderate the effect of trustworthiness. Rather, the race effect may be more reflective of a general racial bias (e.g., Alter, Stern, Granot, & Balcetis, 2016) that occurs independent of the trustworthiness bias. Taken together, these results suggest that trustworthiness may actually function similarly to more obvious coalitional cues like race when perceivers make group inclusion/exclusion decisions. To test the robustness of the race effect, we designed Study 3b as a direct replication of Study 3a.

9. Study 3b

Study 3b was a direct replication of Study 3a (AsPredicted# 21381). The main purpose of this replication was to determine whether the main effect of race found in Study 3a was indeed robust and that participants demonstrate a systematic bias toward including White individuals as ingroup members over Black individuals.

9.1. Method

9.1.1. Participants

We collected data from an additional 102 participants. We removed the data from 22 participants who failed the attention check asking them to which group they were assigned as well as one additional participant who completed the study twice, leaving us with a final sample of 79 participants ($N_{males} = 44$, $N_{females} = 35$, 69.6% White, 12.7% Black, 8.9% Latino/a, 8.9% Asian).

9.1.2. Stimuli

The same stimuli from Study 3a were used in Study 3b.

9.1.3. Procedure

We followed the same procedure for Study 3b as in Study 3a.

9.2. Results

9.2.1. Ingroup categorizations

We analyzed participants' overall inclusion decisions using a GLMM that specified a random slope for facial trustworthiness for each participant, this time also including a random intercept for stimuli. This model yielded only a main effect of facial trustworthiness, $b = 0.47$, $SE = 0.12$, $z = 3.90$, $p < .001$, $OR = 1.60$, 95% CI [1.26, 2.04], with trustworthy faces being 60% more likely to be included into the ingroup than untrustworthy faces. We found no main effect of race, $b = -0.10$, $SE = 0.08$, $z = -1.31$, $p = .19$. Furthermore, the race × trustworthiness interaction was not significant, $b = -0.06$, $SE = 0.16$, $z = -0.39$, $p = .70$, again demonstrating that the impact of facial trustworthiness on ingroup inclusion is not moderated by race.

9.2.2. Comparing to chance

As with our previous studies, we computed the mean inclusion rates for each participant to test against chance. We first collapsed these analyses across face race. In these analyses, we found that trustworthy faces were included at marginally above-chance rates ($M = 0.53$, $SD = 0.23$, 95% CI_{mean} [0.49, 0.57]), $t(78) = 1.82$, $p = .07$, $d = 0.29$, while untrustworthy faces were included at rates significantly below chance ($M = 0.42$, $SD = 0.24$, 95% CI_{mean} [0.39, 0.46]), $t(78) = -3.69$, $p < .001$, $d = -0.59$. We next collapsed these analyses across facial trustworthiness to examine the effect of race. Inclusion rates for Black faces ($M = 0.49$, $SD = 0.21$, 95% CI_{mean} [0.44,

0.54]), did not differ from chance, $t(78) = -0.48$, $p = .63$, $d = -0.08$. Similarly, inclusion rates for White faces ($M = 0.47$, $SD = 0.16$, 95% CI_{mean} [0.43, 0.50]) were marginally below chance, $t(78) = -1.88$, $p = .06$, $d = -0.30$.[3]

9.3. Discussion

Our direct replication failed to reproduce the race effect initially found in Study 3a, indicating that participants' tendencies to include White targets over Black targets into minimal and novel ingroups may not be reliable. Instead, we found further evidence that participants cue in to signals of facial trustworthiness when making ingroup inclusion decisions, and that this effect is not dependent on the race of the target being considered. To further explore the role of race, we include race as a factor in our final study, which also implements a new design that allowed facial trustworthiness to vary continuously.

10. Study 4

Thus far we have shown that facial cues to trustworthiness result in higher rates of ingroup inclusion relative to facial cues to untrustworthiness. In all prior experiments, by using faces that were prerated to be high or low in perceived facial trustworthiness, our designs treated facial trustworthiness as a dichotomous variable. This limits our ability to determine at what *level* of trustworthiness participants become more likely to accept an individual as an ingroup member, and it introduces the possibility that participants were sensitive to some unintended categorical difference between the faces. We therefore designed Study 4 using faces from a database of identities created using FaceGen Modeller (e.g., Todorov et al., 2013a). This allowed us to implement a continuous measure of facial trustworthiness to provide a more fine-grained analysis of the effect of facial trustworthiness on ingroup inclusion across many levels of trustworthiness.

10.1. Method

10.1.1. Participants

Beyond the change in stimuli (see next), the procedure was the same as the prior studies, but with one exception. At the time the study was conducted, there was a recent influx in poor-quality MTurk responders (see Arechar & Rand, 2020). Consequently, based off in-lab estimates of the percentage of poor-quality responders, we added an extra screening measure and increased our sample size by 100 participants. That is, we anticipated a high rate of failing the screening, and so we sought to collect data from 300 participants to reach our goal of 200 participants after exclusions.

At the end of the procedure, we asked participants to report in an open-ended item how they arrived at their inclusion decisions using at least one full sentence. Before analyzing the data, we removed the responses from 59 participants who gave nonsensical answers to this probe (e.g., by writing "GOOD" or by pasting the definition of a sentence). We further removed data from 30 participants who failed the group assignment attention check that asked which group they were assigned to (i.e., overestimator or underestimator), and one participant who displayed nonvariance in their responses (i.e., responding with the same value across all trials), yielding a final sample of 208 participants ($M_{age} = 36.2$, $SD_{age} = 10.8$; $N_{males} = 110$, $N_{females} = 95$, 3 did not specify; 66.8% White, 13.5% Black, 10.1% Asian, 7.2% Latino/a, 2.4% Other). An a priori power analysis conducted using the "simr" package in R (Green & MacLeod, 2016) determined that a sample of 210 participants afforded us > 99% power to detect a minimum effect of $OR = 1.22$ for the effect of facial trustworthiness (corresponding to $b = 0.20$, $p < .05$).

10.1.2. Stimuli

We selected 42 identities from two databases created using the FaceGen Modeller (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Of these identities, 21 were White male targets and 21 were Black male targets. There are seven versions of each specific facial identity in the database, manipulated to range in facial trustworthiness from −3 *SD* to +3 *SD* varying in steps of 1 *SD*. In other words, there were seven levels of trustworthiness for each target. This left us with a total of 294 stimuli, sorted into seven blocks. Each block randomly selected one variant of the facial identity. This ensured that participants saw each identity only one time, at one level of facial trustworthiness.

10.1.3. Procedure

The procedure was largely similar to that of the previous studies. Participants were first randomly sorted into minimal groups of either overestimators or underestimators, and then were shown a series of 42 faces (once per identity). Because the faces in Study 4 were not photographs of real people, we added a brief cover story. Here, we told participants the following:

> "We have entered photographs of prior participants' faces into a 3D face modeler program. These are computer generated images that recreate real people's faces. This can make the faces look more similar to each other than is the reality, but still the faces will vary, even if only slightly. To preserve prior participants' anonymity, we will only show you the computer-processed versions of their faces."

The ingroup categorization probe was the same as in the previous studies. Faces were presented one at a time at an approximate visual angle of 14.25.[4] After categorizing all faces, completing the attention check measures, and completing all demographic measures, participants were thanked and debriefed.

10.2. Results

10.2.1. Ingroup categorizations

We analyzed participants' ingroup categorizations using a similar GLMM as in the prior studies, this time including random intercepts for participants and stimuli with the continuous predictor of facial trustworthiness (−3 *SD* to +3 *SD* in steps of 1 *SD*), face race (White = 0.5, Black = −0.5), and their interaction term. This model yielded the predicted main effect of facial trustworthiness, $b = 0.12$, $SE = 0.01$, $z = 10.38$, $p < .001$, $OR = 1.13$, 95% CI [1.11, 1.16], indicating that participants' likelihood of counting a target as ingroup increased by 13% for each incremental increase in facial trustworthiness. The effect of race was not significant, $b = -0.07$, $SE = 0.04$, $z = -1.64$, $p = .10$, $OR = 0.93$, 95% CI [0.86, 1.01], nor was the race × trustworthiness interaction, $b = 0.03$, $SE = 0.02$, $z = 1.11$, $p = .27$.

10.2.2. Comparing to chance

For the chance comparisons, we explored the effect of facial trustworthiness at each *SD* level. Faces at −3 *SD* in trustworthiness were included at significantly below-chance rates ($M = 0.41$, $SD = 0.29$,

[3] We conducted additional analyses on the combined data from Studies 3a and 3b. Specifically, we first explored only White participants' responses, creating a same/other race factor as we did for sex in Study 2. We next included Black and White participants' responses and explored the same/other race effect. Neither analysis found an effect for anything other than facial trustworthiness. These analyses are included in the Supplementary Materials.

[4] These are not exact values, as the experiments were conducted online, but the approximate values are as follows:
Study 1: 5.72° (6 cm height, estimated 65 cm viewing distance)
Study 2: 6.67° (7 cm height, estimated 65 cm viewing distance)
Study 3a/3b: 5.72°
Study 4: 14.25° (15 cm height, estimated 65 cm viewing distance).

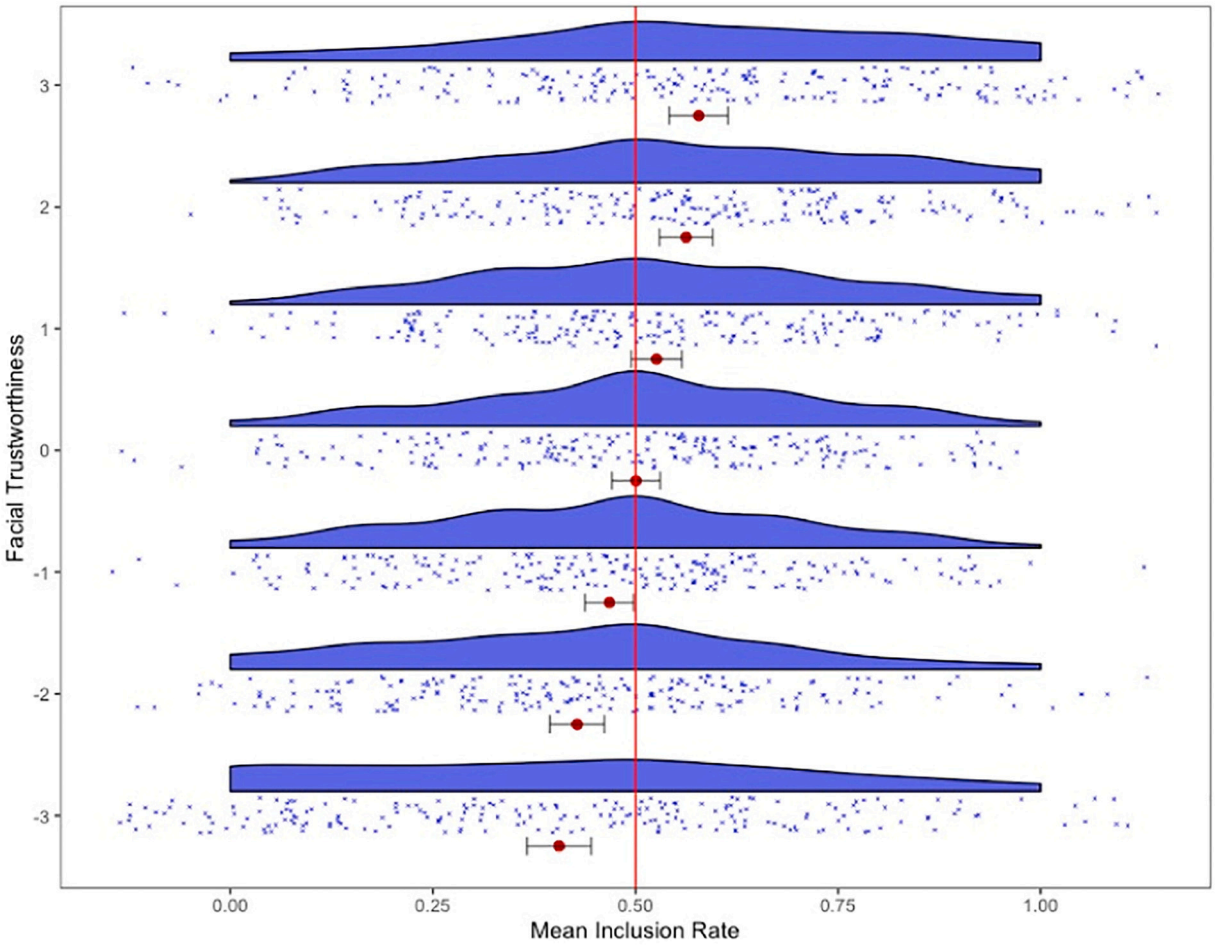

Fig. 1. The distribution of inclusion rates for each participant at all levels of trustworthiness, collapsed across target race. The red line indicates chance inclusion, while the red dots indicate the mean inclusion rate for each level of trustworthiness. Error bars indicate 95% CIs of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

95% CI_{mean} [0.37, 0.45]), $t(207) = -4.66$, $p < .001$, $d = -0.46$. Faces at -2 *SD* were similarly significantly included below chance ($M = 0.43$, $SD = 0.25$, 95% CI_{mean} [0.39, 0.46]), $t(207) = -4.21$, $p < .001$, $d = -0.41$. Faces at -1 *SD* were also included at rates significantly below chance ($M = 0.47$, $SD = 0.22$, 95% CI_{mean} [0.44, 0.50]), $t(207) = -2.10$, $p = .03$, $d = -0.21$. Faces at 0 *SD* did not differ from chance ($M = 0.50$, $SD = 0.22$, 95% CI_{mean} [0.47, 0.53]), $t(207) = 0.04$, $p = .97$, $d = 0.004$, nor did faces at $+1$ *SD*, ($M = 0.53$, $SD = 0.23$, 95% CI_{mean} [0.49, 0.56]), $t(207) = 1.64$, $p = .10$, $d = 0.16$. Faces at $+2$ *SD* in trustworthiness were included at rates significantly above chance ($M = 0.56$, $SD = 0.24$, 95% CI_{mean} [0.53, 0.60]), $t(207) = 3.72$, $p < .001$, $d = 0.37$, as were faces at $+3$ *SD* ($M = 0.58$, $SD = 0.27$, 95% CI [0.54, 0.61]), $t(207) = 4.20$, $p < .001$, $d = 0.41$. These results are summarized in Fig. 1.

Despite the lack of conventional significance in the race variable from the GLMM, we conducted chance comparisons on this factor as well, to parallel the other studies. Inclusion rates for White faces did not differ from chance ($M = 0.49$, $SD = 0.17$, 95% CI_{mean} [0.46, 0.51]), $t(207) = -1.09$, $p = .28$, $d = -0.11$. Similarly, inclusion rates for Black faces did not differ from chance ($M = 0.50$, $SD = 0.19$, 95% CI_{mean} [0.48, 0.53]), $t(207) = 0.33$, $p = .74$, $d = 0.03$.

In a final set of analyses, we conducted comparisons of the inclusion rates of Black and White targets at each level of trustworthiness. In comparing the inclusion rates for Black and White targets at 0 *SD* in trustworthiness (i.e., "neutral" faces), we found no differences in inclusion, $t(207) = -0.02$, $p = .98$, $d = -0.002$. Similarly, no differences were found for the inclusion rates of Black and White targets at $+1$ *SD* in trustworthiness, $t(207) = -0.29$, $p = .77$, $d = -0.03$, nor at -1 *SD* in trustworthiness, $t(207) = 0.61$, $p = .55$, $d = 0.06$. For Black and White faces at $+2$ *SD* in facial trustworthiness, inclusion rates were nearly identical, $t(207) < 0.01$, $p > .99$, $d < 0.01$. Inclusion rates for Black and White faces at -2 *SD* in trustworthiness were only marginally different from each other, $t(207) = 1.69$, $p = .09$, $d = 0.16$. At $+3$ *SD* in facial trustworthiness, participants did not differ in their inclusion rates for Black and White targets, $t(207) = 0.81$, $p = .42$, $d = 0.07$, a result similar to the inclusion rates for Black and White targets at -3 *SD* in trustworthiness, $t(207) = 1.18$, $p = .24$, $d = 0.08$.

Again, these data suggest that participants do not rely on categorical cues (i.e., race), but instead rely on individuating information (i.e., facial trustworthiness) when making ingroup inclusion decisions.

10.3. Discussion

The results from Study 4 provide additional evidence that facial trustworthiness is a key predictor of ingroup inclusion. The lack of an interaction between facial trustworthiness and face race suggests that perceivers' ability to act on facial trustworthiness in the context of inclusion decisions is independent of a target's race. This further illustrates that categorical cues to social group membership (e.g., race) are given less weight than cues that provide information about a target's putative intentions. This is surprising, given prior stereotypes that cast Black individuals as more threatening than White individuals (e.g., Correll, Park, Judd, & Wittenbrink, 2002; Cothran, 2011; Dunham, 2011; Hugenberg & Bodenhausen, 2003, 2004; Shapiro et al., 2009;

Skinner & Haas, 2016). Should participants rely on these stereotypes when making inclusion decisions, it would stand to reason that Black targets at the low end of the facial trustworthiness spectrum (i.e., −3 *SD* and − 2 *SD* from neutral) would be counted as ingroup members at significantly lower rates than White targets at these same levels of facial trustworthiness. Our data showed that this was not the case. In summary, when faces varied in terms of race and trustworthiness, it was trustworthiness—rather than race—that determined inclusion decisions.

11. General discussion

In the service of protecting positive beliefs about the ingroups, people often enact strict criteria for ingroup membership, reserving inclusion for those who likely provide a benefit to the group (Leyens & Yzerbyt, 1992). In five studies, we demonstrated that perceived facial trustworthiness is sufficient to increase the likelihood of being accepted as an ingroup member. In Study 1, participants judged faces that were trustworthy-looking as more likely to belong to the ingroup relative to those that were untrustworthy-looking. Participants preferred to include trustworthy faces as ingroup members while excluding untrustworthy faces at rates significantly above chance. This suggests that participants were consistently sensitive to trustworthiness cues and primarily weighted this information when making ingroup inclusion decisions. Prior research has shown that people exclude those they believe to be "bad" for the ingroup, including potential cheaters (e.g., Kurzban & Leary, 2001; Ponsi et al., 2016). Our data suggest that, in terms of facial morphology, positive (i.e., facial trustworthiness) and negative (i.e., facial untrustworthiness) face cues exert similar influences in the person perception process, providing sufficient information to guide participants' ingroup inclusion decisions.

In Study 2, we sought to replicate and extend the findings from Study 1 by including a more salient aspect of group membership: target sex. In this study, we again found that participants preferred to include trustworthy faces over untrustworthy faces as ingroup members, with no differences emerging based on target sex itself or whether the target's sex was the same as or different from the participants. This result suggests that when it comes to a novel and minimal group, participants rely on perceived cues that ostensibly relate to "character" or personality trait judgments rather than those that indicate pre-existing social categories.

We sought further support for the role of facial trustworthiness in modulating group inclusion by using a sample of Black and White male faces in Studies 3a and 3b, wherein we isolated the effect of race to test two plausible predictions. First, if race cues outweighed trustworthiness information, then our largely White sample would be expected to exclude other-race faces more than same-race faces without regard to facial trustworthiness. However, perceivers are sensitive to trustworthiness in both same and other-race faces (e.g., Cassidy et al., 2017; Wilson et al., 2018) and appear to use such information when forming evaluations of Black and White targets (Charlesworth & Banaji, 2019). While we did find the predicted effect of facial trustworthiness, a separate main effect also emerged in Experiment 3a, showing that White faces were more likely to be included than Black faces. This effect failed to replicate in Study 3b, which was a direct replication of Study 3a that included a separate sample of participants responding to the same stimuli. Combined, the results from Studies 3a and 3b suggest that facial trustworthiness exerts a powerful cue that is more consistent than race in driving decisions about which targets are fellow ingroup members when it comes to minimal and novel groups.

We designed Study 4 as a final test of our facial trustworthiness hypothesis. Rather than using dichotomized faces, we utilized a sample of highly controlled, computer-generated stimuli (Todorov, Dotsch, et al., 2013), selecting Black and White faces that ranged from −3 *SD* to +3 *SD* standardized trustworthiness in steps of 1 *SD*. Again, we found only an effect of facial trustworthiness, such that participants relied on facial cues associated with perceptions of trustworthiness to make their ingroup inclusion decisions, and not on categorical cues that could be linked with other related attributes (e.g., Black = "threatening"; Correll et al., 2002; Cothran, 2011; Dunham, 2011; Hugenberg & Bodenhausen, 2003, 2004; Shapiro et al., 2009; Skinner & Haas, 2016). Additionally, we did not find any interaction between race and trustworthiness, nor did we find any differences in inclusion rates for Black and White targets at *any* level of trustworthiness, ultimately suggesting the effect of trustworthiness on inclusion decisions was similar for both Black faces and White faces.

Across all studies, we conducted chance comparisons for inclusion rates to determine whether participants demonstrated a systematic tendency to include or exclude (un)trustworthy faces in their groups. This was done for several reasons. First, we were able to determine if participants demonstrated response biases toward exclusion or inclusion for any specific face categories. Second, we were able to directly test for overexclusion effects. The initial tests conducted via GLMMs are unable to parcel out whether participants demonstrated an overexclusion effect. Rather, they estimate whether faces at one level of trustworthiness (or race or sex) are more *likely* to be included than others. By conducting chance comparisons (i.e., comparisons to 0.50), we test whether trustworthy faces were systematically included at above-chance rates and whether untrustworthy faces were systematically included below chance. A chance cutoff has been used in prior literature on ingroup overexclusion (Castano, Yzerbyt, Bourguignon, & Seron, 2002; Claypool et al., 2012), and thus provides a valid benchmark to test participants' sensitivity to facial trustworthiness when making group categorization decisions. Despite some heterogeneity, we find evidence that suggests participants are particularly sensitive to facial untrustworthiness, as these targets were included at below-chance rates in all studies but Study 2. Our data from Study 4 is particularly telling, given that all targets below 0 *SD* in trustworthiness were included at significantly below-chance rates (i.e., significantly *excluded* more than chance), independent of race, suggesting that when all facial cues to disposition are controlled for (via computer-generated faces), facial trustworthiness information provides one with enough necessary information to make an ingroup inclusion decision.

Overall, we find support for motivational tendencies predicted by prior work on group membership inclusion and exclusion (e.g., Leyens & Yzerbyt, 1992; Peery & Bodenhausen, 2008; Ponsi et al., 2016; Rubin & Paolini, 2014; Rudert, Reutner, Greifeneder, & Walker, 2017; Yzerbyt et al., 1995). Ingroup membership is a key dimension for social functioning (e.g., Abrams & Hogg, 1988; Brewer, 2004; Correll & Park, 2005; Jetten et al., 2015; Tajfel & Turner, 1986). While obvious concerns such as resource access, protection, and affiliation are known to motivate strict ingroup inclusion decisions (e.g., Brewer, 2004; Correll & Park, 2005; Jetten et al., 2015), identity concerns are also important to current group members. Social Identity Theory argues that a person derives positive self-worth from his or her group memberships (Tajfel & Turner, 1986), with membership in positively valenced groups leading to increased self-esteem among members (e.g., Abrams & Hogg, 1988; Leary et al., 1995).

By including trustworthy-looking faces as ingroup members more frequently than untrustworthy-looking faces, it seems that participants seek to protect the ingroup from presumed "bad actors" (as argued by the literature on ingroup overexclusion; e.g., Castano, 2004; Hutchison & Abrams, 2003; Hutchison et al., 2007; Kurzban & Leary, 2001; Leyens & Yzerbyt, 1992). This is especially relevant in light of the results from Castano (2004), who found that perceived threats to the self increase instances of ingroup overexclusion. Prior research suggests that untrustworthy faces signal a possible intention of doing harm to the perceiver (e.g., Flowe, 2012; Oosterhof & Todorov, 2008), while trustworthy faces have been found to evoke an approach response (e.g., Radke, Kalt, Wagels, & Derntl, 2018). Situated within the context of this literature, the present research provides further support for the role of self-protection motives in ingroup inclusion decisions (see also Ponsi

et al., 2016; Rudert et al., 2017).

This research also provides a complement to recent studies suggesting that people evaluate their own ingroup as more trustworthy than the outgroup, as demonstrated by envisioning the facial features of one's own ingroup as having a more trustworthy appearance than members of the outgroup (Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). Consistent with this effect, we find that even when there are more salient cues to other group memberships (race, sex), it was primarily facial features associated with trustworthiness that determined who was allowed to enter a novel ingroup.

There has long been interest in how facial features and the social categories that are associated with them mutually influence social cognitive processes (Freeman et al., 2008; Macrae & Martin, 2006; Maddox, 2004). Prior research suggests that categorical information is more easily extracted than higher-order individuating information (e.g., Cloutier et al., 2005). Here, we present a domain in which these readily available cues do not take precedent. Rather, participants' ingroup inclusion decisions were reliably influenced by more subtle facial appearances of trustworthiness.

We propose that a social affordance account explains these findings (Zebrowitz, 2006). The current work is consistent with the theory that the configuration of facial features that people describe as trustworthy-looking reflects an affordance we see in someone who has that appearance. The facial features that we associate with interpersonal warmth (rounder eyes, a mouth with slightly upturned corners) are the ones we seem to base our inclusion decisions on. That is, facial appearances of trustworthiness afford us opportunities for affiliation and cooperation (Radke et al., 2018). Another non-mutually exclusive possibility is that trustworthy-looking faces are processed more fluently than untrustworthy-looking faces (e.g., trustworthy face features may be more face-typical or otherwise easily encoded than untrustworthy features), and this fluency may affect ingroup categorization decisions (see Claypool et al., 2012; Schwarz & Clore, 1996; Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Additionally, if trustworthy-looking faces don't call as much attention to their features, this might support more holistic processing, which has recently been shown to undergird a number of positive social judgments about targets (Fincher & Tetlock, 2016; Hugenberg et al., 2016; Wilson et al., 2018). Importantly, facial trustworthiness is distinct from other positively evaluated facial features such as babyfacedness and femininity. Future work would benefit from examining these and other morphological features of the face (e.g., facial width to height ratio) in the context of group categorizations to provide additional insight into ingroup acceptance and rejection decisions.

When it comes to deciding who should be let into a novel group, our participants recognized that target sex (whether male or female, or the same or different as the perceiver) does not reliably afford positive value. Participants recognized males and females alike should be included. While we acknowledge that there may be an issue with prototypicality where untrustworthy female targets are concerned (e.g., Flannigan et al., 2013; see also Oh et al., 2020), our findings suggest that facial trustworthiness is processed equally in both male and female faces when participants consider this information in the context of ingroup inclusion decisions. This is suggested by the lack of an interaction between target sex and trustworthiness in the initial model, and by the at-chance inclusion rates of both male and female faces.

To a similar point, while participants included White targets into the novel ingroup more often than Black targets in Study 3a, this effect did not replicate in Studies 3b and 4. Hence, it does not seem that participants are strongly basing ingroup decisions on race. Similar to the point above, one potential explanation could be the prototypicality of trustworthy versus untrustworthy Black faces. Participants could have viewed trustworthy Black faces as atypical for a Black individual (cf. Livingston & Pearce, 2009), which in turn may have influenced their inclusion decisions. As with the point above, the lack of an interaction between race and sex in three studies suggests that this may not be the case. Moreover, as we found no differences in inclusion rates between Black and White targets at multiple levels of facial trustworthiness in Study 4, it appears that target race (and sex) played little-to-no role in participants' ingroup inclusion decisions. Instead, participants' inclusion decisions were reliably swayed by facial trustworthiness. Is this behavior warranted? While some studies have found that people with trustworthy-looking faces do cooperate more (Stirrat & Perrett, 2010), other studies find no difference in behavioral trustworthiness across different facial appearances (Rule et al., 2013).

Given the broad consensus in perceptions of facial trustworthiness, people might interact much differently with trustworthy-looking targets than untrustworthy-looking targets, which could change the targets' behaviors (as a function of a self-fulfilling prophecy; see Slepian & Ames, 2016; see also Cooley, 1902; Rosenthal, 1994). A person fortunate enough to have a trustworthy-looking face may notice that people tend to include them more, whereas a person cursed with an untrustworthy-looking face may be more likely to feel socially rejected. Indeed, our participants were more willing to include trustworthy-looking targets into a novel ingroup. Thus, while facial trustworthiness does not determine the trustworthiness of one's behavior (e.g., Rule et al., 2013), that does not mean that observers' impressions of trustworthiness based on facial cues are trivial. Indeed, these first impressions lead to differential treatment of individuals with trustworthy and untrustworthy faces, as seen in past research (e.g., Slepian & Ames, 2016; van't Wout & Sanfey, 2008; Wilson & Rule, 2016) and in the present experiments.

When people are frequently bestowed memberships into others' groups, the world may seem like a kind and warm place. When people are frequently denied memberships into others' groups, the world may seem like an unkind and cold place. Prior work has documented how small, appearance-based biases can accumulate into meaningful outcomes (ranging from received salary to the severity sentences for crimes; Fruhen et al., 2015; Holtz, 2014; Wilson & Rule, 2016). In this vein, the current work suggests provocative future directions. Perhaps people with particularly trustworthy-looking faces are prone to trust others and cooperate as a function of feeling like they are valued by others. In contrast, perhaps people with particularly untrustworthy-looking faces are prone to distrust others and defect as a function of being ostracized. If these effects aggregate as do other appearance-based biases, perhaps through repeated experiences of inclusion or exclusion, facial appearance may be associated with constructs such as self-esteem and self-worth.

12. Limitations and future directions

Perhaps the most notable limitation of the present experiments is that our sample was not racially diverse, and so the lack of a consistent race effect across Studies 3a and 3b is bound to majority-group members. More diverse samples may shed light on how participant race factors into ingroup inclusion decisions, especially when the targets are racially diverse. However, at least in the context of sex, we do not find an own-sex bias in ingroup categorization (Study 2). A second limitation of our study is the lack of assessment of participants' levels of sexual or racial prejudice. Future studies may wish to include such measures to further disentangle whether facial trustworthiness is indeed a stronger predictor of inclusion than categorical cues or whether this is subject to individual differences in prejudice.

Other opportunities for future research are suggested by the present findings. For example, face presentation in the current work was supraliminal and unconstrained, allowing participants to make consciously regulated choices about group membership. On the one hand, this makes the findings notable by showing that facial trustworthiness exerts an influence on decisions even when participants have the opportunity to discard biasing cues (the same is true of race in Studies 3a and 3b). Nevertheless, future research that limits face exposure time and/or the time to make group decisions could be valuable, not only to

examine more automatic aspects of overexclusion but also to pit competing face cues (e.g., trustworthiness crossed with sex, and/or race) against one another. Although research clearly shows that race, sex and trustworthiness are all extracted early in visual processing (Ito & Urland, 2003; Todorov et al., 2009; Willis & Todorov, 2006), it would be interesting whether these cues are processed in parallel or instead whether category information (race, sex) or trait information (trust) has precedence in early processing.

A final suggestion for future research is to include faces that vary on trait dimensions other than trustworthiness, for example dominance (e.g., Sutherland et al., 2015), or even approachability or youthful-attractiveness (e.g., Sutherland et al., 2018). Deciding whether to include dominant faces in the ingroup is highly context-dependent (Hehman et al., 2018)—unlike the more univalent trustworthiness—and is likely to interact with sex (e.g., Quist, Watkins, Smith, DeBruine, & Jones, 2011; Torrance, Wincenciak, Hahn, DeBruine, & Jones, 2014) and race in theoretically informative ways.

13. Conclusion

Across five studies, we consistently found that trustworthy faces were included as ingroup members more than untrustworthy faces, and that this effect did not interact with target sex (Study 2) or target race (Studies 3a, 3b, & 4). Thus, bottom-up perceptual cues to trustworthiness may exert a strong influence on fundamental decisions about who is granted or denied ingroup membership, even when competing information about sex and race is available. These findings offer novel insight into how early stages of face processing, including the extraction of both trait-connoting and category-specifying cues, feed forward into important social cognitive processes, like determining who is afforded ingroup membership.

Open practices

All data and analysis scripts are available at https://osf.io/bmfc8/?view_only=7562ebd4fd6149a1ae17d58d902ba0b6.

Credit author statement

All authors contributed equally to the research concept and design. Data analysis was conducted by R. E. Tracy. The manuscript was drafted and revised by R. E. Tracy, J. P. Wilson, M. L. Slepian, and S. G. Young. All authors approved the final version of the manuscript for submission.

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104047.

References

Abrams, D., & Hogg, M. A. (1988). Comments on the motivational status of self-esteem in social identity and intergroup discrimination. *European Journal of Social Psychology, 18*, 317–334. https://doi.org/10.1002/ejsp.2420180403.

Alter, A. L., Stern, C., Granot, Y., & Balcetis, E. (2016). The "bad is black" effect: Why people believe evildoers have darker skin than do-gooders. *Personality and Social Psychology Bulletin, 42*, 1653–1665. https://doi.org/10.1177/0146167216669123.

Arechar, A. A., & Rand, D. G. (2020). Turking in the time of COVID. PsyArXiv https://doi.org/10.31234/osf.io/vktqu.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529. https://doi.org/10.1037/0033-2909.117.3.497.

Bayet, L., Pascalis, O., Quinn, P. C., Lee, K., Gentaz, Ã., & Tanaka, J. W. (2015). Angry facial expressions bias gender categorization in children and adults: Behavioral and computational evidence. *Frontiers in Psychology, 6* https://doi.org/0.3389/fpsyg.2015.00346.

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology, 83*, 5–25. https://doi.org/10.1037/0022-3514.83.1.5.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive motivational analysis. *Psychological Bulletin, 86*, 307–324. https://doi.org/10.1037/0033-2909.86.2.307.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*, 429–444. https://doi.org/10.1111/0022-4537.00126.

Brewer, M. B. (2004). Taking the social origins of human nature seriously: Toward a more imperialist social psychology. *Personality and Social Psychology Review, 8*, 107–113. https://doi.org/10.1207/s15327957pspr0802_3.

Brown, E. R., Phills, C. E., Mercurio IV, D. G., Olah, M., & Veilleux, C. J. (2018). Ain't she a woman? How warmth and competence stereotypes about women and female politicians contribute to the warmth and competence traits ascribed to individual female politicians. *Analyses of Social Issues and Public Policy*, https://doi.org/https://doi.org/10.1111/asap.12151.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x.

Cassidy, B. S., & Krendl, A. C. (2018). Believing is seeing: Arbitrary stigma labels affect the visual representation of faces. *Social Cognition, 36*, 381–410. https://doi.org/10.1521/soco.2018.36.4.381.

Cassidy, B. S., Krendl, A. C., Stanko, K. A., Rydell, R. J., Young, S. G., & Hugenberg, K. (2017). Configural face processing impacts race disparities in humanization and trust. *Journal of Experimental Social Psychology, 73*, 111–124. https://doi.org/10.1016/j.jesp.2017.06.018.

Castano, E., Yzerbyt, V., Bourguignon, D., & Seron, E. (2002). Who may enter? The impact of in-group identification on in-group/out-group categorization. *Journal of Experimental Social Psychology, 38*, 315–322. https://doi.org/10.1006/jesp.2001.1512.

Castano, E., Yzerbyt, V., Paladino, M.-P., & Sacchi, S. (2002). I belong, therefore, I exist: Ingroup identification, ingroup entitativity, and ingroup bias. *Personality and Social Psychology Bulletin, 28*, 135–143. https://doi.org/10.1177/0146167202282001.

Champely, S. (2020). pwr: Basic functions for power analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr.

Charlesworth, T., & Banaji, M. (2019). Face-trait and face-race cues in adults' and children's social evaluations. *Social Cognition, 37*, 357–388. https://doi.org/10.1521/soco.2019.37.4.357.

Chelnokova, O., Laeng, B., Eikemo, M., Riegels, J., Løseth, G., Maurud, H., & Leknes, S. (2014). Rewards of beauty: The opioid system mediates social motivation in humans. *Molecular Psychiatry, 19*, 746–747. https://doi.org/10.1038/mp.2014.1.

Claypool, H. M., Housley, M. K., Hugenberg, K., Bernstein, M. J., & Mackie, D. M. (2012). Easing in: Fluent processing brings others into the ingroup. *Group Processes & Intergroup Relations, 15*, 441–455. https://doi.org/10.1177/1368430212439115.

Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology, 88*, 885–894. https://doi.org/10.1037/0022-3514.88.6.885.

Cooley, C. H. (1902). *Human nature and the social order*. Charles Scribner's Sons.

Correll, J., & Park, B. (2005). A model of the ingroup as a social resource. *Personality and Social Psychology Review, 9*, 341–359. https://doi.org/10.1207/s15327957pspr0904_4.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314.

Cothran, D. L. (2011). Facial affect and race influence threat perception. *Imagination, Cognition and Personality, 30*, 341–354. https://doi.org/10.2190/ic.30.3.g.

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2017). Split-second trustworthiness detection from faces in an economic game. *Experimental Psychology, 64*, 231–239. https://doi.org/10.1027/1618-3169/a000367.

Dunham, Y. (2011). An angry = Outgroup effect. *Journal of Experimental Social Psychology, 47*, 668–671. https://doi.org/10.1016/j.jesp.2011.01.003.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*, 876–893. https://doi.org/10.1037/0022-3514.87.6.876.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83. https://doi.org/10.1016/j.tics.2006.11.005.

Flannigan, N., Miles, L. K., Quadflieg, S., & Macrae, C. N. (2013). Seeing the unexpected: Counterstereotypes are implicitly bad. *Social Cognition, 31*, 712–720. https://doi.org/10.1521/soco.2013.31.6.172.

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

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2–18. https://doi.org/10.1037/a0024338.

Fruhen, L. S., Watkins, C. D., & Jones, B. C. (2015). Perceptions of facial dominance, trustworthiness and attractiveness predict managerial pay awards in experimental tasks. *The Leadership Quarterly, 26*, 1005–1016. https://doi.org/10.1016/j.leadqua.

2015.07.001.
Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common ingroup identity model.* New York, NY, US: Psychology Press.
Green, P., & MacLeod, C. J. (2016). "simr": An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493–498. https://doi.org/10.1111/2041-210X.12504.
Greenaway, K. H., Haslam, S. A., Cruwys, T., Branscombe, N. R., Ysseldyk, R., & Heldreth, C. (2015). From "we" to "me": Group identification enhances perceived personal control with consequences for health and well-being. *Journal of Personality and Social Psychology, 109*, 53–74. https://doi.org/10.1037/pspi0000019.
Hehman, E., Sutherland, C. A., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology, 113*, 513–529. https://doi.org/10.1037/pspa0000090.
Hertel, G., & Kerr, N. L. (2001). Priming in-group favoritism: The impact of normative scripts in the minimal group paradigm. *Journal of Experimental Social Psychology, 37*, 316–324. https://doi.org/10.1006/jesp.2000.1447.
Holtz, B. C. (2014). From first impression to fairness perception: Investigating the impact of initial trustworthiness beliefs. *Personnel Psychology, 68*, 499–546. https://doi.org/10.1111/peps.12092.
Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of hostility. *Psychological Science, 14*, 640–643. https://doi.org/10.1046/j.0956-7976.2003.psci_1478.x.
Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization. The role of prejudice and facial affect in race categorization. *Psychological Science, 15*, 342–345. https://doi.org/10.1111/j.0956-7976.2004.00680.x.
Hutchison, P., & Abrams, D. (2003). Ingroup identification moderates stereotype change in reaction to ingroup deviance. *European Journal of Social Psychology, 33*, 497–506. https://doi.org/10.1002/ejsp.157.
Hutchison, P., Abrams, D., & Christian, J. (2007). The social psychology of exclusion. In D. Abrams, J. Christian, & D. Gordon (Eds.). *Multidisciplinary handbook of social exclusion research* (pp. 29–57). John Wiley & Sons, Ltd.
Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*, 616–626. https://doi.org/10.1037/0022-3514.85.4.616.
Jetten, J., Branscombe, N. R., Haslam, S. A., Haslam, C., Cruwys, T., Jones, J. M., … Zhang, A. (2015). Having a lot of a good thing: Multiple important group memberships as a source of self-esteem. *PLoS One, 10*, Article e0124609. https://doi.org/10.1371/journal.pone.0124609.
Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology, 111*, 655–664. https://doi.org/10.1037/pspa0000062.
Kleisner, K., Priplatova, L., Frost, P., & Flegr, J. (2013). Trustworthy-looking face meets brown eyes. *PLoS One, 8*, Article e53285. https://doi.org/10.1371/journal.pone.0053285.
Knowles, E. D., & Peng, K. (2005). White selves: Conceptualizing and measuring a dominant-group identity. *Journal of Personality and Social Psychology, 89*, 223–241. https://doi.org/10.1037/0022/3514.89.2.223.
Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgments from faces. *PLoS One, 13*, Article e0202655 https://doi.org/10.1371.journal.pone.0202655.
Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin, 127*, 187–208. https://doi.org/10.1037/0033-2909.127.2.187.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82, 1–26. https://doi.org/10.18637/jss.v082.i13.
Leary, M. R., Tambor, E. S., Terdal, S. K., & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology, 68*, 518–530. https://doi.org/10.1037/0022-3515.68.3.518.
Lemyre, L., & Smith, P. M. (1985). Intergroup discrimination and self-esteem in the minimal group paradigm. *Journal of Personality and Social Psychology, 49*, 660–670. https://doi.org/10.1037/0022-3514.49.3.660.
Levin, D. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1364–1382. https://doi.org/10.1037/0278-7393.22.6.1364.
Levin, D. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General, 129*, 559–574. https://doi.org/10.1037/0096-3445.129.4.559.
Leyens, J.-P., & Yzerbyt, V. Y. (1992). The ingroup overexclusion effect: Impact of valence and confirmation on stereotypical information search. *European Journal of Social Psychology, 22*, 549–569. https://doi.org/10.1002/ejsp.2420220604.
Livingston, R. W., & Pearce, N. A. (2009). The teddy-bear effect: Does having a baby face benefit black chief executive officers? *Psychological Science, 20*, 1229–1236. https://doi.org/10.1111/j.1467-9280.2009.02431.x.
Macrae, C. N., & Martin, D. (2006). A boy primed sue: Feature-based processing and person construal. *European Journal of Social Psychology, 37*, 793–805. https://doi.org/10.1002/ejsp.406.
Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8*, 383–401. https://doi.org/10.1207/s15327957pspr0804_4.
O'Boyle, E. H., Forsyth, D. R., & O'Boyle, A. S. (2011). Bad apples or bad barrels: An examination of group- and organizational-level effects in the study of counterproductive work behavior. *Group & Organization Management, 36*, 39–69. https://doi.org/10.1177/1059601110390998.
Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General, 149*, 323. https://doi.org/10.1037/xge0000638.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences. 105. Proceedings of the National Academy of Sciences* (pp. 11087–11092). . https://doi.org/10.1073/pnas.0805664105.
Otten, S., & Moskowitz, G. B. (2000). Evidence for implicit evaluative in-group bias: Affect-biased spontaneous trait inference in a minimal group paradigm. *Journal of Experimental Social Psychology, 36*, 77–89. https://doi.org/10.1006/jesp.1999.1399.
Peery, D., & Bodenhausen, G. V. (2008). Black + white = black. *Psychological Science, 19*, 973–977. https://doi.org/10.1111/j.1467-9280.2008.02185.x.
Ponsi, G., Panasiti, M. S., Scandola, M., & Aglioti, S. M. (2016). Influence of warmth and competence on the promotion of safe in-group selection: Stereotype content model and social categorization of faces. *Quarterly Journal of Experimental Psychology, 69*, 1464–1479. https://doi.org/10.1080/17470218.2015.1084339.
Quist, M. C., Watkins, C. D., Smith, F. G., DeBruine, L. M., & Jones, B. C. (2011). Facial masculinity is a cue to women's dominance. *Personality and Individual Differences, 50*, 1089–1093. https://doi.org/10.1016/j.paid.2011.01.032.
Radke, S., Kalt, T., Wagels, L., & Derntl, B. (2018). Implicit and explicit motivational tendencies to faces varying in trustworthiness and dominance in men. *Frontiers in Behavioral Neuroscience, 12*, 1–10. https://doi.org/10.3389/fnbeh.2018.00008.
Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology, 106*, 897–911. https://doi.org/10.1037/a0036498.
Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3*, 176–179. https://doi.org/10.1111/1467-8721.ep10770698.
Rubin, M., & Paolini, S. (2014). Out-group flies in the in-group's ointment. *Social Psychology, 45*, 265–273. https://doi.org/10.1027/1864-9335/a000171.
Rudert, S. C., Reutner, L., Greifeneder, R., & Walker, M. (2017). Faced with exclusion: Perceived facial warmth and competence influence moral judgments of social exclusion. *Journal of Experimental Social Psychology, 68*, 101–112. https://doi.org/10.1016/j.jesp.2016.06.005.
Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*, 409–426. https://doi.org/10.1037/a0031050.
Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*, 260–264. https://doi.org/10.1037/a0014681.
Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. In E. T. Higgins, & A. W. Kruglanski (Eds.). *Social psychology: Handbook of basic principles* (pp. 433–465). The Guilford Press.
Shapiro, J. R., Ackerman, J. M., Neuberg, S. L., Maner, J. K., Vaughn Becker, D., & Kenrick, D. T. (2009). Following in the wake of anger: When not discriminating is discriminating. *Personality and Social Psychology Bulletin, 35*, 1356–1367. https://doi.org/10.1177/0146167209339627.
Sherman, J. W., Klein, S. B., Laskey, A., & Wyer, N. A. (1998). Intergroup bias in group judgment processes: The role of behavioral memories. *Journal of Experimental Social Psychology, 34*, 51–65. https://doi.org/10.1006/jesp.1997.1342.
Skinner, A. L., & Haas, I. J. (2016). Perceived threat associated with police officers and Black men predicts support for policing policy reform. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01057.
Slepian, M. L., & Ames, D. R. (2016). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological Science, 27*, 282–288. https://doi.org/10.1177/0956797615594897.
Slepian, M. L., & Carr, E. W. (2019). Facial expressions of authenticity: Emotion variability increases judgments of trustworthiness and leadership. *Cognition, 183*, 82098. https://doi.org/10.1016/j.cognition.2018.10.009.
Slepian, M. L., Young, S. G., & Harmon-Jones, E. (2017). An approach-avoidance motivational model of trustworthiness judgments. *Motivation Science, 3*, 91097. https://doi.org/10.1037/mot0000046.
Slepian, M. L., Young, S. G., Rule, N. O., Weisbuch, M., & Ambady, N. (2012). Embodied impression formation: Social judgments and motor cues to approach and avoidance. *Social Cognition, 30*, 232–240. https://doi.org/10.1521/soco.2012.30.2.232.
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*, 990–1000. https://doi.org/10.1177/1948550617732388.
Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.). *The Claremont symposium on applied social psychology: Reducing prejudice and discrimination* (pp. 23–45). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust. *Psychological Science, 21*, 349–354. https://doi.org/10.1177/0956797610362647.
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, 115*, 9210-9215. https://doi.org/10.31234/osf.io/d9ha7.
Sutherland, C., Oldmeadow, J., & Young, A. (2014). Are first impressions the same for male and female faces? *Journal of Vision, 14*, 1275. https://doi.org/10.1167/14.10.1275.
Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin, 44*, 521–537. https://doi.org/10.1177/0146167217744194.

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*, 186–208. https://doi.org/10.1111/bjop.12085.

Sutherland, J. E., Cojocariu, A. M., Day, D. M., & Hehman, E. (2020). Youths' facial appearance distinguishes leaders from followers in group-perpetrated criminal offenses and is associated with sentencing outcomes. *Criminal Justice and Behavior, 47*, 187–207. https://doi.org/10.1177/0093854819889645.

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*, 796–803. https://doi.org/10.1037/e722292011-008.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology, 1*, 149–178. https://doi.org/10.1002/ejsp.2420010202.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel, & W. Austin (Eds.). *Psychology of intergroup relations* (pp. 7–24). Nelson Hall.

Thorstenson, C. A., Pazda, A. D., Young, S. G., & Slepian, M. L. (2019). Incidental cues to threat and racial categorization. *Social Cognition, 37*, 389–404. https://doi.org/10.1521/soco.2019.37.4.389.

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*, 724–738. https://doi.org/10.1037/a0032335.

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology, 23*, 373–380. https://doi.org/10.1016/j.conb.2012.12.010.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*, 813–833. https://doi.org/10.1521/soco.2009.27.6.813.

Torrance, J. S., Wincenciak, J., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2014). The relative contributions of facial shape and surface information to perceptions of attractive and dominance. *PLoS One, 9*, Article e104415. https://doi.org/10.1371/journal.pone.0104415.

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology, 9*, 187–204. https://doi.org/10.1002/ejsp.2420090207.

Wang, H., Tong, S., Shang, J., & Chen, W. (2019). The role of gender in the preconscious processing of facial trustworthiness and dominance. *Frontiers in Psychology, 10*, 1–12. https://doi.org/10.3389/fpsyg.2019.02565.

Wild, H. A., Barrett, S. E., Spence, M. J., O'Toole, A. J., Cheng, Y. D., & Brooke, J. (2000). Recognition and sex categorization of adults' and children's faces: Examining performance in the absence of sex-stereotyped cues. *Journal of Experimental Child Psychology, 77*, 269–291. https://doi.org/10.1006/jecp.1999.2554.

Williams, K. D. (2009). Ostracism: A temporal need-threat model. In M. P. Zanna (Ed.). *Advances in experimental social psychology* (pp. 275–314). Elsevier Academic Press.

Willis, J., & Todorov, A. (2006). First impressions. *Psychological Science, 17*, 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x.

Wilson, J. P., & Rule, N. O. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes. *Social Psychological and Personality Science, 7*, 331–338. https://doi.org/10.1177/1948550615624142.

Wilson, J. P., Young, S. G., Rule, N. O., & Hugenberg, K. (2018). Configural processing and social judgments: Face inversion particularly disrupts inferences of human-relevant traits. *Journal of Experimental Social Psychology, 74*, 1–7. https://doi.org/10.1016/j.jesp.2017.07.007.

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch, & K. C. Klauer (Eds.). *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 197–226). Lawrence Earlbaum Associates, Inc.

Yzerbyt, V. Y., Leyens, J.-P., & Bellour, F. (1995). The ingroup overexclusion effect: Identity concerns in decisions about group membership. *European Journal of Social Psychology, 25*, 1–16. https://doi.org/10.1002/ejsp.2420250102.

Zebrowitz, L. A. (2006). Finally, faces find favor. *Social Cognition, 24*, 657–701. https://doi.org/10.1521/soco.2006.24.5.65.