

Facial Stereotype Bias Is Mitigated by Training

Social Psychological and
Personality Science
1-10
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1948550620972550
journals.sagepub.com/home/spp



Kao-Wei Chua¹  and Jonathan B. Freeman¹

Abstract

People automatically infer others' personality (e.g., trustworthiness) based on facial appearance, and such facial stereotype biases predict real-world consequences across political, legal, and business domains. The present research tested whether these biases can be mitigated through counterstereotype training aimed at reconfiguring the associations between specific facial appearances and social traits. Across six studies and a replication, a behavioral counterstereotype training consistently reduced or eliminated facial stereotype biases for White male faces in the context of economic trust games, hiring decisions, and even automatic evaluations assessed via evaluative priming. Together, the results demonstrate a fundamental malleability in facial stereotyping related to trustworthiness, with a minimal training able to mitigate the tendency to activate and apply long-held, highly automated facial stereotypes. These findings suggest that face impressions are more flexible than typically appreciated, and they provide a potential inroad toward combating our ingrained biases based on facial appearance.

Keywords

face perception, stereotypes, statistical learning, impression formation

At a glance, people quickly evaluate others' faces based on personality traits such as trustworthiness. People make these trait inferences as quickly as 100 ms (Bar et al., 2006) and do so automatically and without conscious awareness (Engell et al., 2007; Freeman et al., 2014). Although rarely diagnostic of a target's actual personality (Rule et al., 2013), trait impressions evoked by faces tend to be reliable across perceivers (Oosterhof & Todorov, 2008). These trait impressions can be consequential, as facial trustworthiness predicts outcomes from electoral success (Todorov et al., 2005) to criminal-sentencing decisions (Wilson & Rule, 2015). Here, we investigate the malleability of such social judgments, testing whether learning can reduce the tendency to spontaneously judge faces.

To date, face impression research has focused on bottom-up structural aspects, highlighting specific arrangements of features serving as "facial stereotypes" for particular traits. Ecological accounts of person perception have long emphasized evolutionarily important arrangements of facial features, with trait impressions reflecting overgeneralizations of functionally adaptive cues, such as emotional expressions (McArthur & Baron, 1983). As such, a face's physical resemblance to emotional expressions conveys specific trait impressions, for example, joyful cues (e.g., upturned mouth) convey trustworthiness while angry cues (e.g., furrowed brow) convey untrustworthiness (Oosterhof & Todorov, 2009; Zebrowitz et al., 2003). Data-driven models have identified specific facial features associated with countless perceived traits (Oosterhof & Todorov, 2008). Even 3-year-old children reliably evaluate

traits from faces, leading some to argue that such trait associations may be congenitally determined (Cogsdill et al., 2014).

However, experience and learning could also play a role. Although perceivers are unlikely to form trait associations with facial appearance based on targets' own behavior (as facial appearance does not reliably correspond to one's personality traits; Rule et al., 2013), they may implicitly learn regularities between certain facial features (e.g., furrowed brow) and how others judge and react to individuals with those features (e.g., untrustworthy), as these face-based judgments and reactions tend to be highly consistent (Oosterhof & Todorov, 2008). Accordingly, such trait associations with facial appearance would come to reflect preconceived notions derived from the social environment. Indeed, researchers have increasingly documented the effects of learning and experience on trait associations of faces (e.g., Dotsch et al., 2016; Hehman et al., 2017; Sofer et al., 2017; Stolier et al., 2018, 2020).

In the present research, we leverage such sensitivity to learning to explore an intervention to reduce facial stereotyping. Given face impressions' lack of correspondence with actual personality and their impact on real-world outcomes

¹ Department of Psychology, New York University, NY, USA

Corresponding Author:

Kao-Wei Chua, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA.
Email: kc3644@nyu.edu

(Olivola et al., 2014), it is important to examine whether facial stereotype biases can be ameliorated. A rich literature has leveraged counterstereotype interventions to reduce social biases like racial or gender bias (Lai et al., 2016; Paluck & Green, 2009), but trait impressions have not typically been conceptualized as a bias requiring intervention (despite operating as facial stereotypes). Counterstereotype interventions have typically had participants engage in counterstereotypical imagery or be presented with counterstereotypical exemplars or pairings of exemplars and traits (e.g., a female name and “strong”; for reviews, Blair, 2002, Forscher et al., 2019; Gawronski & Bodenhausen, 2006). In the face impressions literature, studies have long examined how trait impressions of a face are updated based on newly learned behaviors related to that trait (Todorov & Uleman, 2002, Bliss-Moreau et al., 2008). For instance, face identities paired with positive or negative behaviors are subsequently judged as more or less trustworthy, respectively (Todorov & Olson, 2008). This impression updating occurs because behavioral information spontaneously triggers trait inferences that alter impressions (Uleman et al., 1996). As learned behaviors can affect trait associations for individual face identities, it may be possible to leverage such behavior-based learning to flexibly reassociate trait associations with more general facial appearances, that is, facial stereotypes. Indeed, previous work found that facial features become rapidly associated with a trait if those features are paired with the trait’s related behaviors (Lick et al., 2018).

Here, we integrate the bias intervention and impression updating literatures to investigate whether a counterstereotype training paradigm, which pairs facial features and behaviors, has the potential to reduce facial stereotyping. A training task was used to either reverse (trained group) or maintain (control group) the mapping between untrustworthy/trustworthy facial features and untrustworthy/trustworthy behaviors. We focus on trustworthiness because it is the primary dimension by which faces are judged and is a proxy for general face valence (Oosterhof & Todorov, 2008). In the trained group, untrustworthy faces were associated with trustworthy behaviors 80% of the time and trustworthy faces were associated with untrustworthy behaviors 80% of the time. In the control group, the same faces were instead presented with a name label that provided similar individuating information without countering or affirming facial stereotypes, allowing this group to act as a baseline for trustworthiness evaluations. Overall, we found that the behavioral counterstereotype training reduced facial stereotype biases in an economic trust game (Studies 1A/1B), hiring decisions (Studies 2A/2B), and even in automatic evaluations assessed via evaluative priming (Studies 3A/3B), thereby showing that even long-held facial stereotype associations can be mitigated through associative learning mechanisms.

Study 1

We test whether training reduces facial stereotype bias in an economic trust game, assessed via payments to computer-generated (Study 1A) or real (Study 1B) faces.

Method

Participants

Without a direct precedent, for all studies, we used a target sample size of 200 participants (100 participants/group), the sample necessary to detect a small-to-medium effect size ($d = .3$) at 80% power. All studies were performed on Mechanical Turk. Participants received monetary compensation.

Two hundred twenty-five total participants performed Study 1A. For final analyses, 101 participants were in the control group (age: $M = 33.7$ years, $SD = 10.3$ years; 43 male; race: 77 White, nine Black, eight Asian, seven Other; Hispanic/Latino ethnicity: 10), and 100 participants were in the trained group (age: $M = 35.5$, $SD = 11.1$; sex: 39 male; race: 80 White, nine Black, eight Asian, three Other; Hispanic/Latino ethnicity: 6). Two hundred twenty total participants completed Study 1B; 102 participants were in the final control group (age: $M = 33.3$, $SD = 10.3$; 57 male; 80 White, 15 Black, seven Asian; Hispanic/Latino ethnicity: 12), and 100 participants were in the final trained group (age: $M = 34.7$, $SD = 10.0$; sex: 58 male; race: 79 White, 13 Black, eight Asian; Hispanic/Latino ethnicity: 10).

Procedure

Participants engaged in a two-part task. The first task involved a learning phase purporting to test face memory. For control participants, the learning phase involved 20 different faces paired with a name label. The trained participants viewed 20 faces paired with one-sentence trustworthy/untrustworthy behavioral descriptions (see Supplementary Material for details). Participants were instructed to memorize the face–behavior or face–name pairings for a later test of face memory.

In Study 1A, the target faces were computer-generated White male faces generated via FaceGen (Blanz & Vetter, 1999). The faces were manipulated along the trustworthiness trait dimension (Oosterhof & Todorov, 2008), such that they were $+2 SD$ or $-2 SD$ in trustworthiness. Faces were cropped such that only internal face features were visible. Participants only saw one variant of each facial identity in the study (trustworthy or untrustworthy). One set of faces was used during training, and a different set of faces was used for the evaluation phase.

In Study 1B, the stimuli were natural White male faces from the Basel Face Database (Walker et al., 2018). These faces were systematically manipulated on the communion dimension, which is virtually identical to the trustworthiness dimension in facial feature space, acting as a semantic analogue to trustworthiness (Stolier et al., 2018). The stimuli consisted of White male faces whose features were increased $+2 SD$ or decreased $-2 SD$ in trustworthiness/communion, making them a functional analogue to the Study 1A faces (Figure 1).

Both control and trained participants saw 10 unique trustworthy faces and 10 unique untrustworthy faces during the learning phase. For the trained group, the 10 trustworthy faces were paired with untrustworthy behaviors 80% of the time and

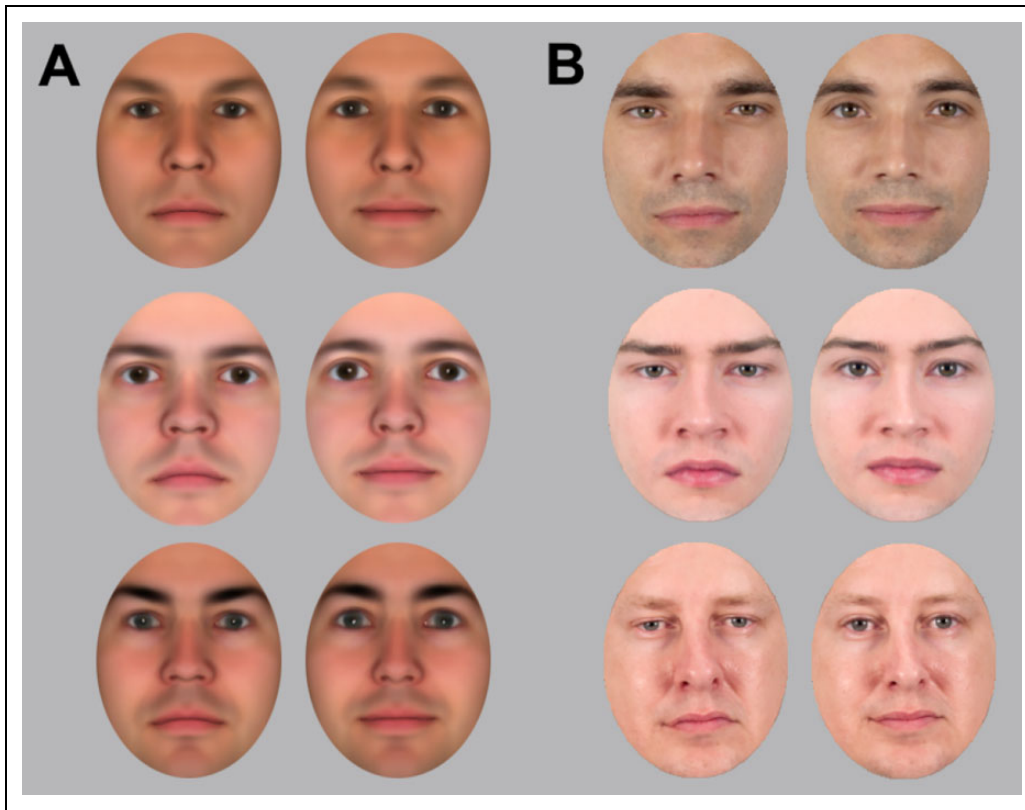


Figure 1. Examples of computer-generated faces (A) and real faces (B) that systematically vary in trustworthiness. Note. The faces on the left columns are 2 *SD* below the population mean in trustworthiness and faces on the right are 2 *SD* above the population mean.

the 10 untrustworthy faces were paired with trustworthy behaviors 80% of the time. The face–behavior pairings were randomized for each subject. Each trial was self-paced and presented in a randomized order. A time-out warning (meant to ensure participants carefully read the face–behavior/face–name pairings) remained on screen for 2,000 ms if participants finished a trial in under 500 ms. Each face–behavior pairing was repeated three times, resulting in 60 learning trials.

Participants then performed a second task, ostensibly unrelated to the learning phase. This task was an economic trust game involving decisions to allocate money to another player, similar to previous studies (Bonnefon et al., 2013; Rezliescu et al., 2012; van’t Wout & Sanfey, 2008). To improve the task’s realism, participants were told they were playing for real money, and they were urged to maximize profits. Participants selected a face avatar to represent themselves, and they were told other human players were represented by similar avatars. The time between each round of the trust game began with a delay of random duration (1,000–8,000 ms) to simulate the latency in matching with another player. In actuality, there were no human players and participants were presented with a fixed set of faces in a randomized order. Additional data suggested that participants believed the cover story (see Supplementary Material).

Before the game began, participants were told they were assigned as Player 1, the “giver.” Consistent with previous

studies (van’t Wout & Sanfey, 2008), for each trial, they received US\$1.00 and could give a portion to another player (options: US\$0.00, US\$0.25, US\$0.50, US\$0.75, and US\$1.00). Whatever amount given was tripled, and their partner could return as much or as little back to the participant. For example, if the participant gave the full US\$1.00, the other player would have US\$3.00 and could return whatever portion of US\$3.00 back to the participant. However, if the participant found the player’s avatar untrustworthy, they could give less money and keep the remainder.

The trust game involved 32 total rounds. For 16 rounds, participants were paired with 16 unique partners. These 16 partners were either high or low in trustworthiness (+2 and –2 *SDs*), resulting in eight trustworthy faces and eight untrustworthy faces. The remaining 16 rounds consisted of female avatars with no systematic variability in trustworthiness, included to reduce the task’s transparency. Critically, the faces in the trust game were distinct from those in the learning phase so we could test whether training generalized to novel target faces that shared trustworthy and untrustworthy face features (but differed in identity).

Five attention checks were interspersed between trust game rounds to ensure participants maintained attention. The participants were instructed to press a number (1–5). After the trust game, participants learned there would be no test of face memory and were debriefed about the study’s aims.

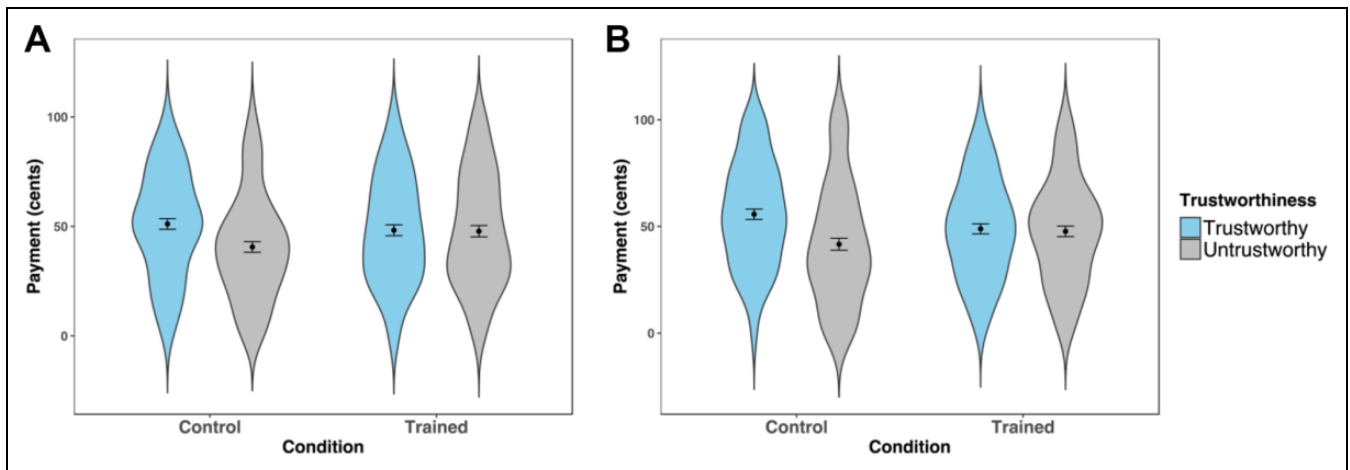


Figure 2. Violin plots representing probability densities for payment allocated to trustworthy and untrustworthy faces, separately for the control and trained groups for Study 1A (A) and Study 1B (B). Note. Error bars denote standard error of the mean. Note that values may exceed the possible 0–100 cents range because probability densities are depicted.

Results and Discussion

Participants who triggered timeout errors on >50% of learning phase trials (indicating they were not carefully encoding face–behavior/face–name pairings) were excluded (Study 1A: 14 participants; Study 1B: six participants), as were participants who failed any attention check (Study 1A: 10 participants; Study 1B: 12 participants). Data for all studies are available at <https://osf.io/3x5qn/>

In Study 1A with computer-generated faces, a 2 (group: control vs. trained) \times 2 (facial trustworthiness: trustworthy vs. untrustworthy) mixed-model analysis of variance (ANOVA) on payment elicited a main effect of trustworthiness, $F(1, 199) = 9.79, p = .002, \eta_p^2 = .047$, with greater payment to trustworthy versus untrustworthy faces, replicating previous findings (Bonnefon et al., 2013). There was no main effect of group, $F(1, 199) = 0.48, p = .49, \eta_p^2 = .02$. Critically, there was a significant Group \times Facial Trustworthiness interaction, $F(1, 199) = 11.57, p = .001, \eta_p^2 = .055$ (Figure 2). As predicted, the control group paid targets with trustworthy faces more than those with untrustworthy faces, $F(1, 199) = 21.4, p < .0001, 95\% \text{ CI } [6.00, 14.91], \eta_p^2 = .10$, but this bias was extinguished for the trained group, $F(1, 199) = 0.04, p = 0.85, 95\% \text{ CI } [-4.04, 4.92], \eta_p^2 = .00$.

Similarly, in Study 1B with real faces, a 2 (group: control vs. trained) \times 2 (trustworthiness: trustworthy vs. untrustworthy) mixed-model ANOVA elicited a main effect of facial trustworthiness, $F(1, 200) = 32.54, p < .0001, \eta_p^2 = .14$, with greater payment to Trustworthy vs. Untrustworthy faces. There was no main effect of group, $F(1, 200) = 0.015, p = .90, \eta_p^2 < .00$. Importantly, the interaction was significant, $F(1, 200) = 23.26, p < .0001, \eta_p^2 = .104$ (Figure 2). The control group paid targets with trustworthy faces more than those with untrustworthy faces, $F(1, 200) = 56.0, p < .0001, 95\% \text{ CI } [10.34, 17.75], \eta_p^2 = .22$, and this bias was extinguished in the trained group, $F(1, 200) = 0.38, p = .54, 95\% \text{ CI } [-2.56, 4.91], \eta_p^2 = .002$.

Thus, a brief learning phase was sufficient to abolish the effects of facial appearance on trust game payments. While control participants were influenced by facial trustworthiness, trained participants paid trustworthy and untrustworthy faces equally. Moreover, because the training generalized to novel targets, these results show that the associations between specific facial appearances and trustworthy/untrustworthy traits were flexibly updated rather than simply updating impressions of specific individuals.

Study 2

In Study 2, we aim to extend the extinction of facial stereotype biases to downstream behaviors with important social consequences. Participants engaged in identical training as Study 1 and performed a mock hiring process in accordance with studies examining hiring bias (Rooth, 2010). Participants evaluated candidates for a job and were given biographical information alongside a candidate's face. They rated candidates' suitability for a job and the likelihood that they would interview them. Candidates' faces were either computer-generated (Study 2A) or real (Study 2B). Prior work has shown that facial stereotype biases persist even when more diagnostic cues are available (Jaeger et al., 2019; Olivola & Todorov, 2010). In Study 2, we test whether training reduces facial stereotype bias even in contexts where more diagnostic information is provided.

Method

Participants

Two hundred nineteen total participants completed Study 2A. About 104 participants were in the final control group (age: $M = 35.0, SD = 9.0$; sex: 60 male; race: 79 White, 16 Black, nine Other; Hispanic/Latino ethnicity: 15), and 102 participants were in the final trained group (age: $M = 36.3, SD = 9.8$; sex:

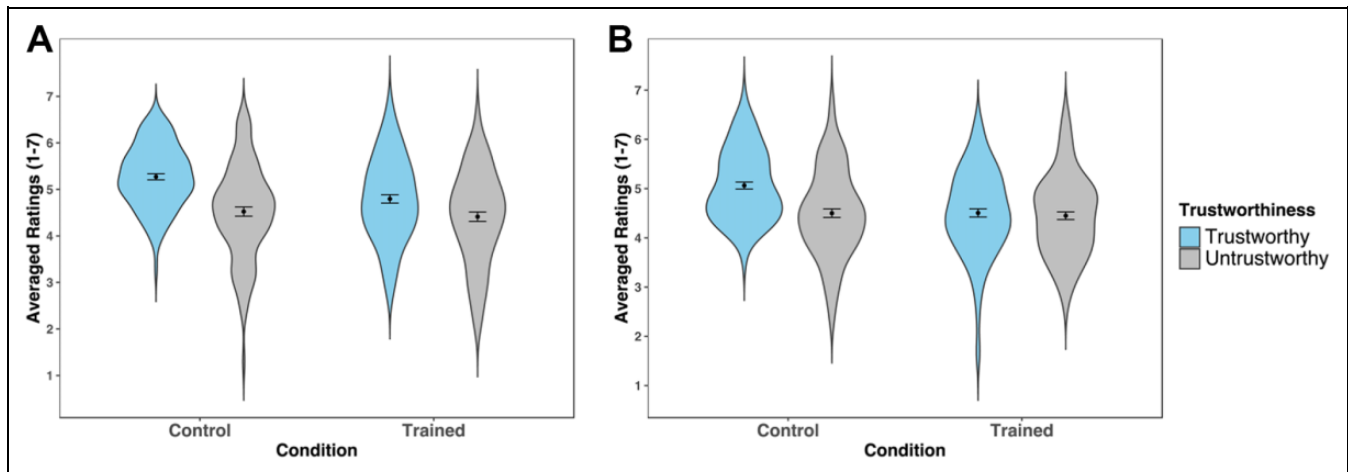


Figure 3. Violin plots representing probability densities for the job suitability metric, separately for trustworthy and untrustworthy faces and for the control and trained groups in Study 2A (A) and Study 2B (B). Note. Error bars denote standard error of the mean. Note that values may exceed the possible 1–7 ratings range because probability densities are depicted.

53 male; race: 82 White, 12 Black, eight Asian; Hispanic/Latino ethnicity: 15). 235 total participants completed Study 2B. Around 105 participants were in the control group (age: $M = 36.8$, $SD = 11.0$; sex: 48 male; race: 87 White, five Black, 10 Asian, three Other; Hispanic/Latino ethnicity: 15), and 112 participants were in the trained group (age: $M = 35.0$, $SD = 10.0$; sex: 57 male; race: 82 White, 17 Black, 11 Asian, two Other; Hispanic/Latino ethnicity: 14).

Procedure

Following the same training as Study 1, participants performed a task purportedly helping a company make hiring decisions. They were shown a picture of a candidate as well as their age, years of relevant work experience, and college attended. Ages were randomized from 24 to 26 years in 1-year increments. Relevant work experience was randomized from 2.5 to 3.5 years in half-year increments. Colleges were taken from a ranking of top 25 public universities (U.S. News and World Report, 2019). The combination of face, age, education, and work experience was randomized for each participant. Biographical information was presented in accordance with studies examining bias in hiring (Rooth, 2010).

The hiring task involved making two judgments about each candidate: rating how suitable they were for the job (1 = *not suitable at all* to 7 = *very suitable*) and how likely they were to invite them for an interview (1 = *not at all likely to interview* to 7 = *very likely to interview*). The candidates consisted of 10 trustworthy and 10 untrustworthy faces, yielding 20 candidates total. In Study 2A, candidates were computer-generated faces, and in Study 2B, candidates were real faces. Interspersed between trials were attention checks, wherein participants typed a specific number (1–7). After rating the candidates, participants were debriefed.

Results and Discussion

Participants were removed for >50% time-out trials in the learning phase (Study 2A: five participants; Study 2B: six participants), failing any attention checks (Study 2A: six participants; Study 2B: five participants), or for having fast mean reaction times (RT; <500 ms) during the hiring task (Study 2A: two participants; Study 2B: seven participants).

In Study 2A, the two measures were highly correlated ($r = .89$, $p < .0001$, 95% CI [0.863, 0.918]), so they were averaged together into an overall job suitability metric and submitted to a 2 (group: control vs. trained) \times 2 (trustworthiness: trustworthy vs. untrustworthy) mixed-model ANOVA. There was a main effect of trustworthiness, $F(1, 204) = 81.48$, $p < .0001$, $\eta_p^2 = .29$, with trustworthy-faced candidates deemed more suitable for jobs than untrustworthy-faced candidates. The main effect of group was also significant, $F(1, 204) = 7.11$, $p = .008$, $\eta_p^2 = .03$. Critically, there was a significant interaction, $F(1, 204) = 8.64$, $p = .004$, $\eta_p^2 = 0.041$ (Figure 3). The control group deemed trustworthy-faced candidates as more suitable than untrustworthy-faced candidates, $F(1, 204) = 72.3$, $p < .0001$, 95% CI [0.58, 0.92], $\eta_p^2 = .26$. However, this bias was attenuated in the trained group, $F(1, 204) = 18.35$, $p < .0001$, 95% CI [0.206, 0.556], $\eta_p^2 = .08$.

For Study 2B (real faces), the two measures were again highly correlated ($r = .87$, $p < .0001$, 95% CI [0.84, 0.90]), so were averaged together and submitted to a 2 (group) \times 2 (trustworthiness) mixed-model ANOVA. There was a main effect of trustworthiness, $F(1, 215) = 36.25$, $p < .0001$, $\eta_p^2 = .14$, with higher ratings for trustworthy versus untrustworthy faces. The main effect of group was significant, $F(1, 215) = 9.10$, $p = .003$, $\eta_p^2 = .04$. Critically, there was a significant Group \times Trustworthiness interaction, $F(1, 215) = 24.18$, $p < .0001$, $\eta_p^2 = .101$ (Figure 3). For the control group, candidates with trustworthy versus untrustworthy faces were deemed more suitable, $F(1, 215) = 57.95$, $p < .0001$, 95% CI

[0.42, 0.71], $\eta_p^2 = .21$, but this bias was extinguished in the trained group, $F(1, 215) = 0.63$, $p = 0.43$, 95% CI [-0.08, 0.20], $\eta_p^2 = .003$.

Extending the results of Study 1, the brief training reduced the effect of facial appearance in the domain of hiring decisions. Importantly, the findings also show that this bias reduction persists even in contexts where more diagnostic, relevant information is provided rather than a task where only faces are provided.

Study 3

In the previous studies, it is possible that participants were merely responding to learned information in a deliberate manner rather than genuinely shifting their evaluations. In Study 3, we use an evaluative priming task to test whether training reduces facial stereotype biases even in automatic face evaluations. Evaluative priming measures automatic evaluations in an indirect manner (Fazio et al., 1986), including for faces (Olson & Fazio, 2003) and social biases (Olson & Fazio, 2006). Participants were shown trustworthy or untrustworthy face primes followed by a positive or negatively valenced word (e.g., “happy,” “poison”). Given the automaticity of face evaluations (e.g., Freeman et al., 2014; Todorov et al., 2009), in the control group, we expect RT facilitation effects, whereby evaluatively congruent face–target pairs (trustworthy face/positive word and untrustworthy face/negative word) elicit faster RTs than incongruent pairs (trustworthy face/negative word and untrustworthy face/positive word). If the training is successful in remapping these associations and reducing facial stereotypes, we would expect reduced RT facilitation in the trained group.

Method

Participants

Two hundred forty-four total participants completed Study 3A on Mechanical Turk. Previous studies have successfully conducted evaluative priming tasks in online samples (e.g., Mattan et al., 2019; Green et al., 2019). About 110 participants were in the final control group (age: $M = 36.9$, $SD = 11.3$; sex: 51 male; race: 87 White, 13 Black, 10 Other; Hispanic/Latino ethnicity: 9), and 110 participants were in the final trained group (age: $M = 37.0$, $SD = 10.5$; sex: 55 male; race: 88 White, 15 Black, four Asian, three Other; Hispanic/Latino ethnicity: 8). Two hundred twenty-six total participants completed Study 3B, and 97 participants were in the final control group (age: $M = 38.4$, $SD = 12.3$; sex: 46 male; race: 75 White, 13 Black, six Asian, three Other; Hispanic/Latino ethnicity: 5), and 108 participants were in the final trained group (age: $M = 36.0$, $SD = 9.7$; 60 male; 79 White, 16 Black, seven Asian, six Other; Hispanic/Latino ethnicity: 13).

Procedure

Following the same training as previous studies, participants were told they would perform a language test. On each trial, a fixation cross was presented (500 ms), followed by a prime face (200 ms), followed by a blank screen (100 ms), followed

by target word that appeared until a response. The timing was based on recommendations from previous studies to achieve facilitation effects (Fazio et al., 1986; Hermans et al., 1994). The task was to classify the target word as positive or negative as quickly and accurately as possible by key press. The prime faces consisted of new trustworthy and untrustworthy faces that were not encountered during training, with eight unique identities for each condition, resulting in 16 prime faces that were computer-generated (Study 3A) or real (Study 3B). The target words were adjectives that had positive (e.g., “Good,” “Kind”) or negative (“Bad,” “Mean”) valences. There were 10 target words for each valence. Each prime face (16 faces) was paired with each target word once, resulting in 320 trials total.

Results and Discussion

Participants were removed for having >50% of learning trials register as time-outs (Study 3A: 11 participants; Study 3B: eight participants) or for at-chance performance (50% accuracy) in the evaluative priming task (Study 3A: 13 participants; Study 3B: 13 participants). For final analysis, we removed incorrect responses (Study 3A: 5% of trials; Study 3B: 6% of trials) and trials with RTs faster than 250 ms and slower than 3,000 ms (Study 3A: 6% of trials; Study 3B: 7% of trials).

For each participant, RT difference scores were computed for (negative–positive) words. In this calculation, positive values (greater than 0) denote facilitation for positive words and difference scores smaller than 0 denote facilitation for negative words. Separate RT difference scores were computed for trustworthy and untrustworthy faces.

For Study 3A, RT difference scores were submitted to a 2 (group) \times 2 (trustworthiness) mixed-model ANOVA, eliciting a main effect of facial trustworthiness, $F(1, 218) = 44.87$, $p < .0001$, $\eta_p^2 = .17$, and a significant interaction, $F(1, 218) = 14.11$, $p < .0001$, $\eta_p^2 = .061$ (Figure 4). There was no main effect of group, $F(1, 218) = 0.88$, $p = .35$, $\eta_p^2 = .004$. Control participants showed strong RT facilitation and automatic, evaluatively congruent evaluations, $F(1, 218) = 54.66$, $p < .0001$, 95% CI [28.3, 48.9], $\eta_p^2 = .20$, whereas this effect was considerably weaker in trained participants, $F(1, 218) = 4.33$, $p = .04$, 95% CI [0.57, 21.1], $\eta_p^2 = .02$.

For Study 3B, RT difference scores were submitted to a 2 (group) \times 2 (trustworthiness) mixed-model ANOVA, eliciting a main effect of trustworthiness, $F(1, 203) = 13.34$, $p < .0001$, $\eta_p^2 = .06$, and a significant interaction, $F(1, 203) = 7.86$, $p = .006$, $\eta_p^2 = .04$. There was no main effect of group, $F(1, 203) = 0.03$, $p = .87$, $\eta_p^2 = .000$. Control participants again showed strong evaluatively-congruent evaluations, $F(1, 203) = 19.77$, $p < .0001$, 95% CI [18.1, 46.9], $\eta_p^2 = .09$, which were reduced in trained participants, $F(1, 203) = 0.38$, $p = .54$, 95% CI [-9.4, 17.9], $\eta_p^2 = .002$.

Extending the results of the previous studies, we found training reduced not only payment and hiring decisions, but even participants’ automatically generated evaluations for trustworthy and untrustworthy targets, assessed in a more implicit manner.

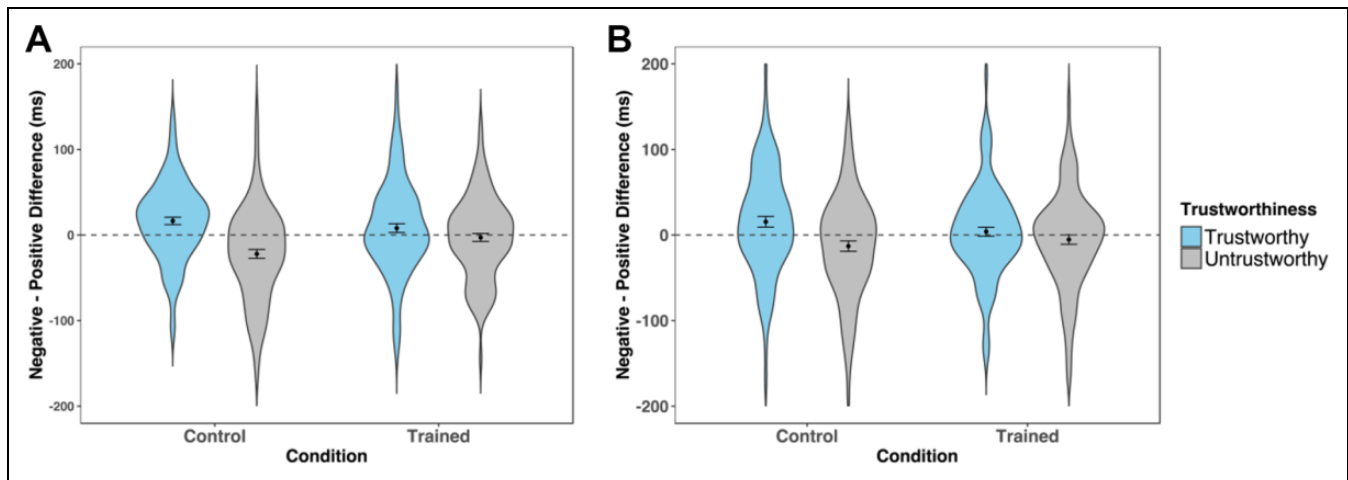


Figure 4. Violin plots representing probability densities for reaction time (RT) facilitation effects (negative RT–positive RT), separately for trustworthy and untrustworthy faces in the control and trained groups in Study 3A (A) and Study 3B (B). *Note.* The dashed line represents zero facilitation in either direction. Positive values indicate facilitation for positive words and negative values indicate facilitation for negative words. Error bars denote standard error of the mean.

Meta-Analysis and Replication

To assess the overall strength of the training effects, we meta-analyzed the six studies using fixed effects, with effect size weighted by sample size (Goh et al., 2016). The effect size for the Trustworthiness \times Group interaction in each study was converted to Cohen's d . The meta-analytic effect was significant, with a conventionally medium effect size, mean $d = .53$, $Z = 9.30$, $p < .00001$. The aggregated meta-analytic framework also permitted additional analyses described in the Supplementary Material. Participant race did not influence any results, and the training effects held when controlling for other facial dimensions such as attractiveness and competence. Finally, the training effects were replicated in an additional study ($n = 199$) that used a different set of training behaviors (see Supplementary Material).

General Discussion

Across six studies and a replication, we demonstrate flexibility in facial stereotyping, with a brief training reducing or extinguishing the activation and application of facial stereotypes. Expectedly, untrained control participants were strongly biased by facial appearance. When targets bore untrustworthy appearances, control participants allocated less money, found them to be less suitable for jobs, and harbored automatic evaluative biases against them. When associations between facial appearance and trustworthiness were reversed in trained participants, facial stereotype biases were reduced or eliminated. Critically, the test phase involved novel targets, so these results reflect a generalized bias reduction for untrustworthy versus trustworthy facial appearances rather than merely effects of impression updating for specific targets. This bias reduction was observed for both computer-generated and real faces and demonstrated across a wide set of contexts: when facial appearance was the only information provided (Study 1), when diagnostic,

decision-relevant information was also provided (Study 2), and even when automatic evaluative responses were assessed in a more implicit manner (Study 3).

Previous work has shown that face impressions update when new behavioral information about specific identities is encountered (Bliss-Moreau et al., 2008; Todorov & Olson, 2008; Todorov & Uleman, 2002). The present results demonstrate a distinct, more fundamental form of updating underlying face impressions, whereby specific facial appearances themselves can be flexibly reassociated with new trait information. In the context of counterstereotype training paradigms used to reduce racial or gender bias (Blair, 2002; Forscher et al., 2019; Gawronski & Bodenhausen, 2006; Lai et al., 2016; Paluck & Green, 2009), it is noteworthy that our training paired exemplars with behaviors rather than traits (as is common in these paradigms), and yet a brief training period yielded particularly strong effects. Because behavioral information spontaneously triggers trait inferences (Uleman et al., 1996), it is likely that these inferred traits may alter stereotypical associations through similar mechanisms as presenting the traits directly (e.g., Gawronski & Bodenhausen, 2006; Gawronski et al., 2008). However, in theory, behavioral pairings in a counterstereotype training context may operate doubly as individuating information, leading to greater individuation of targets than would standard trait pairings, thereby aiding in bias reduction (Brewer, 1988; Fiske & Neuberg, 1990). If true, such behavioral counterstereotype training may capitalize on a synergy between counterstereotypical learning and individuation processes and represent a promising paradigm for bias reduction research more generally. It is also possible that facial stereotype bias is simply easier to reduce than other biases. We believe this is unlikely, as previous attempts to reduce facial stereotyping have failed (Jaeger et al., 2019).

Previous attempts to reduce facial stereotyping have focused on either educating participants to avoid facial stereotypes or

“nudging” them to use more diagnostic cues, which proved unsuccessful (Jaeger et al., 2019). Unlike the present work, this previous study attempted to reduce bias by raising participants’ awareness of their general stereotypes. However, raising awareness of a general rather than specific bias can often be ineffective (Axt et al., 2019). Moreover, while interventions focusing on participants’ goals to reduce bias can be effective (Fazio & Towles-Schwen, 1999; Forscher et al., 2019), counterstereotype interventions like those used in the present research (aimed at changing underlying associations) tend to elicit stronger bias-reduction effects (Lai et al., 2016). It is also likely that the former interventions rely more on deliberate, propositional processes and the latter on more automatic, associative processes, and this distinction may have implications for successful bias reduction (Gawronski & Bodenhausen, 2006). Future research could directly compare multiple candidate interventions in the novel context of facial stereotyping as well as test their boundary conditions.

A critical question is whether the mitigation of facial stereotype biases persists over time. Lab-based interventions have successfully reduced implicit racial bias, but these effects tend to vanish after 3–4 days (Lai et al., 2016). Given a lifetime of acquired facial stereotype associations, it might prove difficult for single-shot interventions to result in long-term evaluative shifts. In the context of racial bias, researchers have incorporated similar kinds of lab-based counterstereotype training into longer, more intensive interventions under the framework of bias habit-breaking, which have demonstrated long-term bias reductions (Devine et al., 2012; McNulty et al., 2017). Our findings provide a critical first demonstration that facial stereotype bias is indeed malleable rather than fixed and susceptible to counterstereotype interventions. Theoretically, this is important because while other social biases like racial bias are widely acknowledged to be malleable, as they reflect cultural learning (which interventions seek to reverse), facial stereotype bias is generally assumed to be relatively fixed due to evolutionary adaptation. Thus, demonstrating a generalized bias reduction in activating and applying facial stereotypes even in the short-term represents an important advance. Future research could build on the present work to test whether the present training, in isolation or in conjunction with more temporally extended interventions, could have demonstrable long-term bias reduction effects.

There are several limitations of this work. Our stimuli were limited to White male faces so as to avoid potential confounds such as individual differences in gender and racial bias. Previous work has shown that target race and gender affect facial trustworthiness evaluations; however, substantial variance in these evaluations is still driven by trustworthiness-related features that are independent of race and gender (Hehman et al., 2019; Xie et al., 2019). Future studies could build on this work to explore how the training effects generalize to faces of other social groups and potentially interact with other social dimensions. Moreover, an alternative interpretation of our overall results is that rather than the training weakening associations between specific features and traits, it violated participants’

expectations and decreased their general reliance on facial stereotypes. Future research could directly examine the specificity versus generality of these training effects, and this could have implications for different intervention strategies.

In summary, the present work provides evidence for an associative learning mechanism that shapes our evaluations of others’ faces and that this can be exploited to reduce the activation and application of facial stereotypes. These findings not only show that face impressions are more flexible than typically appreciated, but they also provide an inroad toward combating ingrained biases based on facial appearance.

Acknowledgments

We thank Michael Berkebile and Maryam bin Meshar for their help with the studies and stimulus preparation. We thank Dr. Chu Chang Chua for her continued guidance.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by NSF BCS-1654731 (J.B.F) and SBE-1911860 (K.C.) grants.

ORCID iD

Kao-Wei Chua  <https://orcid.org/0000-0002-2035-1030>

Supplemental Material

The supplemental material is available in the online version of the article.

References

- Axt, J. R., Casola, G., & Nosek, B. A. (2019). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin*, 45(8), 1232–1251. <https://www.ncbi.nlm.nih.gov/pubmed/30520340>. <https://doi.org/10.1177/0146167218814003>
- Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, 6(2), 269–278. <https://www.ncbi.nlm.nih.gov/pubmed/16768559>. <https://doi.org/10.1037/1528-3542.6.2.269>
- Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6(3), 242–261. https://doi.org/10.1207/S15327957pspr0603_8
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*.
- Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion*, 8(4), 479–493. <https://www.ncbi.nlm.nih.gov/pubmed/18729580>. <https://doi.org/10.1037/1528-3542.8.4.479>
- Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental*

- Psychology: General*, 142(1), 143–150. <https://www.ncbi.nlm.nih.gov/pubmed/22686638>. <https://doi.org/10.1037/a0028930>
- Brewer, M. (1988). A dual process model of impression formation. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Advances in social cognition. A dual process model of impression formation* (Vol. 1, pp. 1–36). Lawrence Erlbaum Associates.
- Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, 25(5), 1132–1139. <https://www.ncbi.nlm.nih.gov/pubmed/24570261>. <https://doi.org/10.1177/0956797614523297>
- Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, 48(6), 1267–1278. <https://www.ncbi.nlm.nih.gov/pubmed/23524616>. <https://doi.org/10.1016/j.jesp.2012.06.003>
- Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, 1(1). <https://doi.org/10.1038/s41562-016-0001>
- Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19(9), 1508–1519. <https://www.ncbi.nlm.nih.gov/pubmed/17714012>. <https://doi.org/10.1162/jocn.2007.19.9.1508>
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50(2), 229–238. <https://doi.org/10.1037/0022-3514.50.2.229>
- Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 97–116). The Guilford Press.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier.
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3), 522–559. <https://www.ncbi.nlm.nih.gov/pubmed/31192631>. <https://doi.org/10.1037/pspa0000160>
- Freeman, J. B., Stolier, R. M., Ingbreetsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *The Journal of Neuroscience*, 34(32), 10573–10581. <https://www.ncbi.nlm.nih.gov/pubmed/25100591>. <https://doi.org/10.1523/JNEUROSCI.5063-13.2014>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5), 692–731. <https://www.ncbi.nlm.nih.gov/pubmed/16910748>. <https://doi.org/10.1037/0033-2909.132.5.692>
- Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When “just say no” is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, 44(2), 370–377.
- Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, 10(10), 535–549.
- Green, L. J. S., Luck, C. C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. <https://www.ncbi.nlm.nih.gov/pubmed/31750707>. <https://doi.org/10.1037/emo0000701>
- Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass*, 13(2). <https://doi.org/ARTNe1243110.1111/spc3.12431>
- Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, 113(4), 513–529. <https://www.ncbi.nlm.nih.gov/pubmed/28481616>. <https://doi.org/10.1037/pspa0000090>
- Hermans, D., Houwer, J. D., & Eelen, P. (1994). The affective priming effect: Automatic activation of evaluative information in memory. *Cognition & Emotion*, 8(6), 515–533.
- Jaeger, B., Todorov, A., Evans, A., & van Beest, I. (2019). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, 90, 104004.
- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., ... Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, 145(8), 1001–1016. <https://www.ncbi.nlm.nih.gov/pubmed/27454041>. <https://doi.org/10.1037/xge0000179>
- Lick, D. J., Alter, A. L., & Freeman, J. B. (2018). Superior pattern detectors efficiently learn, activate, apply, and update social stereotypes. *Journal of Experimental Psychology: General*, 147(2), 209–227. <https://www.ncbi.nlm.nih.gov/pubmed/28726438>. <https://doi.org/10.1037/xge0000349>
- Mattan, B. D., Kubota, J. T., Li, T., Venezia, S. A., & Cloutier, J. (2019). Implicit evaluative biases toward targets varying in race and socioeconomic status. *Personality and Social Psychology Bulletin*, 45(10), 1512–1527.
- Mcarthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social-perception. *Psychological Review*, 90(3), 215–238. <https://doi.org/10.1037/0033-295x.90.3.215>
- McNulty, J. K., Olson, M. A., Jones, R. E., & Acosta, L. M. (2017). Automatic associations between one’s partner and one’s affect as the proximal mechanism of change in relationship satisfaction: Evidence from evaluative conditioning. *Psychological Science*, 28(8), 1031–1040.
- Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, 18(11), 566–570. <https://doi.org/10.1016/j.tics.2014.09.007>
- Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, 46(2), 315–324.

- Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*(6), 636–639. <https://www.ncbi.nlm.nih.gov/pubmed/14629698>. https://journals.sagepub.com/doi/10.1046/j.0956-7976.2003.psci_1477.x?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed&. https://doi.org/10.1046/j.0956-7976.2003.psci_1477.x
- Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421–433. <https://www.ncbi.nlm.nih.gov/pubmed/16513796>. <https://doi.org/10.1177/0146167205284004>
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092. <https://www.ncbi.nlm.nih.gov/pubmed/18685089>. <https://doi.org/10.1073/pnas.0805664105>
- Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*(1), 128–133. <https://www.ncbi.nlm.nih.gov/pubmed/19186926>. <https://doi.org/10.1037/a0014520>
- Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology, 60*, 339–367. <https://www.ncbi.nlm.nih.gov/pubmed/18851685>. <https://doi.org/10.1146/annurev.psych.60.110707.163607>
- Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLOS ONE, 7*(3), e34293. <https://www.ncbi.nlm.nih.gov/pubmed/22470553>. <https://doi.org/10.1371/journal.pone.0034293>
- Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523–534. <https://doi.org/10.1016/j.labeco.2009.04.005>
- 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*(3), 409–426. <https://doi.org/10.1037/a0031050>
- Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception, 46*(8), 914–928. <https://www.ncbi.nlm.nih.gov/pubmed/28152651>. <https://doi.org/10.1177/0301006617691786>
- Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behavior, 4*(4), 361–371.
- Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America, 115*(37), 9210–9215. <https://www.ncbi.nlm.nih.gov/pubmed/30139918>. <https://doi.org/10.1073/pnas.1807222115>
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626. <https://www.ncbi.nlm.nih.gov/pubmed/15947187>. <https://doi.org/10.1126/science.1110589>
- Todorov, A., & Olson, I. R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience, 3*(3), 195–203. <https://www.ncbi.nlm.nih.gov/pubmed/19015111>. <https://doi.org/10.1093/scan/nsn013>
- Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833.
- Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051–1065. <https://www.ncbi.nlm.nih.gov/pubmed/12416911>
- Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 211–279). Academic Press.
- U.S. News and World Report (2019) *Top Public Schools*. <https://www.usnews.com/best-colleges/rankings/national-universities/top-public>
- van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796–803. <https://www.ncbi.nlm.nih.gov/pubmed/18721917>. <https://doi.org/10.1016/j.cognition.2008.07.002>
- Walker, M., Schonborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in big two and big five personality dimensions. *PLOS ONE, 13*(3), e0193190. <https://www.ncbi.nlm.nih.gov/pubmed/29590124>. <https://doi.org/10.1371/journal.pone.0193190>
- Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325–1331. <https://www.ncbi.nlm.nih.gov/pubmed/26162847>. <https://doi.org/10.1177/0956797615590992>
- Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology, 117*(2), 364–385. <https://www.ncbi.nlm.nih.gov/pubmed/30550328>. <https://doi.org/10.1037/pspi0000160>
- Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review, 7*(3), 194–215.

Author Biographies

Kao-Wei Chua is a postdoctoral scholar at the New York University, Department of Psychology, specializing in perceptual expertise and the impact of learning and experience on face and object perception.

Jonathan B. Freeman is an associate professor of psychology and neural science at New York University and the director of the Social Cognitive & Neural Sciences Lab. His research specializes in social perception, face categorization, and the impact of conceptual knowledge on face processing.

Handling Editor: Margo Monteith