THIS MANUSCRIPT IS CURRENTLY UNDER REVIEW

Mitigating Facial Stereotyping:

Counter-Stereotype Exposure Succeeds Where Explicit Instruction Fails

Samuel A. W. Klein[1] and Jonathan B. Freeman[1]

[1]Department of Psychology, Columbia University, NY, USA

Author Note

This research was supported by National Science Foundation Grant BCS 2235130.

Correspondence concerning this article should be addressed to Samuel A. W. Klein, Columbia University, 1190 Amsterdam Ave, New York, NY, 10026, United States. Email: sak2290@columbia.edu

The data presented in this manuscript were collected in experiments approved by the Columbia University Institutional Review Board (protocol ID: AAAU5505).

Samuel A. W. Klein is postdoctoral research scientist in the Social Cognitive & Neural Sciences Lab, at Columbia University. His research primarily examines the cognitive processes behind how we form, update, and translate social impressions into behavior.

Jonathan B. Freeman is a Professor of Psychology at Columbia University and the director of the Social Cognitive & Neural Sciences Lab. His research primarily examines the cognitive and neural mechanisms underlying person perception, bias and stereotyping, and the real-time dynamics of social and emotional judgments.

Wordcount: 4,969

Abstract

Perceivers often draw inferences of personality from facial appearance, exerting pervasive influence on social cognition and behavior, despite their limited validity. These biases have proven remarkably resistant to mitigation efforts. This research examines two common interventions against facial stereotyping: counter-stereotype exposure and explicit instructions to control the bias. Across three experiments (N=1,632), participants made trust-based decisions about people whose facial appearance was manipulated to be more or less trustworthy, following counter-stereotype exposure, debiasing instructions, or a control procedure. Exposure reliably reduced facial stereotyping, remaining effective across a diverse set of faces, under high cognitive demand, and independent of explicit beliefs about facial stereotyping. Instructions, by contrast, consistently failed and became counterproductive under high cognitive demands. This dissociation suggests facial stereotyping is best modified through experience that may reshape the underlying associations between facial features and trait concepts, and helps explain why facial stereotyping often resists relatively deliberative debiasing strategies.

Keywords: face perception; person perception; impression formation; stereotyping

Mitigating Facial Stereotyping:
Counter-Stereotype Exposure Succeeds Where Explicit Instruction Fails

The human face is one of the most potent sources of social information, and its appearance readily cues inferences about a person's personality (Oosterhof & Todorov, 2008). Such inferences may partly reflect overgeneralization from adaptive emotion-detection systems, such that features resembling anger, for example, signal untrustworthiness (Zebrowitz, 1996; Zebrowitz & Montepare, 2006). These face stereotypes have little-to-no correspondence with target individuals' actual personality or behavior (Krendl et al., 2014; Rule et al., 2013). Nonetheless, they predict consequential decisions, including voting, hiring, and criminal sentencing (Olivola et al., 2014; Wilson & Rule, 2015, 2016). Studying how to modulate facial stereotyping is, therefore, both theoretically and practically important.

Malleability of Facial Stereotyping

Correcting against facial stereotyping has proven difficult. Giving decision-makers more time to reflect did not limit face stereotypes from biasing economic exchanges, nor did offering more decision-relevant information (Jaeger et al., 2019). Educating decision-makers about the inaccuracy and harms of facial stereotyping similarly failed, as did delaying exposure to a person's face until after an initial evidence-based decision had been made; when subsequently given the opportunity to revise that judgment with the face revealed, participants nevertheless exhibited facial stereotyping (Jaeger et al., 2020).

Despite these failures, recent work has identified an alternative channel through which to intervene on facial stereotyping. Repeated exposure to counter-stereotypic exemplars (e.g., untrustworthy-looking faces presented with trustworthy behaviors) produced markedly less trustworthiness-based facial stereotyping (Chua & Freeman, 2021; Hong et al., 2024). Because the facial identities during exposure were distinct from those in later judgment tasks, the reduced

use of facial trustworthiness information generalizes beyond individual identities and reflects a broader shift in facial stereotyping. However, the mechanisms responsible for such bias reduction effects remain poorly understood. Nor has counter-stereotype exposure been directly compared to explicit instructional approaches, limiting our understanding of the malleability in facial stereotyping.

Mechanisms of Facial Stereotyping Malleability

Counter-stereotype exposure and direct instructions against facial stereotyping may owe their differential effectiveness to engaging distinct mental processes: relatively automatic associative processing versus more deliberative, resource-demanding processing (Gawronski & Bodenhausen, 2006, 2018). Face stereotypes are associations between facial features and trait inferences. They activate rapidly (Bar et al., 2006; Willis & Todorov, 2006), under limited cognitive capacity (Bonnefon et al., 2013), and without conscious awareness or intention (Chua & Freeman, 2021; Engell et al., 2007; Freeman et al., 2014; Hong et al., 2024; Klapper et al., 2016), which may make them particularly resistant to deliberative control alone. Counter-stereotype exposure may succeed by weakening existing face stereotypes through associative learning (Chua & Freeman, 2021; Hong et al., 2024). Indeed, novel face stereotypes have been found to emerge through associative learning, such as when facial features statistically predict trait-relevant information (Chua & Freeman, 2022; Lick et al., 2018).

In contrast, explicit instructions to suppress or avoid facial stereotyping may engage deliberative control processes that leave underlying feature-trait associations intact. Such control is typically resource-intensive and failure-prone under limited cognitive capacity (Amodio & Swencionis, 2018; Braver, 2012; Galinsky & Moskowitz, 2007; Wang et al., 2020; Wegner, 1994). Instructions may also face headwinds from perceivers' own beliefs. Unlike stereotypes about race or gender, many people endorse facial appearance as a valid signal of personality

(Jaeger et al., 2022), and stronger endorsement predicts stronger application of face stereotypes (Jaeger et al., 2020). Perceivers who regard facial appearance as genuinely informative are likely not ordinarily motivated to regulate their use of it, which may make interventions requiring deliberative control especially demanding—a vulnerability that interventions targeting associative processing may sidestep.

The Present Research

The following three experiments investigate the mechanisms of facial stereotyping malleability. Experiment 1 offers the first test of whether counter-stereotype exposure mitigates facial stereotyping independent of explicit beliefs about face stereotypes, while simultaneously testing its generality to racially and gender diverse faces. Experiment 2 offers a direct comparison between counter-stereotype exposure and explicit instructions to avoid facial stereotyping. Experiment 3 then examines whether the impact of each intervention is robust or vulnerable to relatively high cognitive demands. Together, these experiments implicate distinct mechanisms through which facial stereotyping may be changed and the conditions under which those mechanisms succeed or fail.

Experiment 1

Experiment 1 tested whether counter-stereotype exposure reduces facial stereotyping independently of explicit endorsement for face stereotypes. Participants completed an economic trust game following counter-stereotype exposure or a control procedure. If exposure reduces the impact of facial appearance on trust-based decisions without changing endorsement, successful intervention would not depend on changing perceivers' explicit beliefs about the validity of face stereotypes.

In addition, although face stereotypes vary systematically by social groups (Hong & Freeman, 2024; Xie et al., 2019; Xie et al., 2021), White male faces have been exclusively

examined in prior work on facial stereotyping interventions (Chua & Freeman, 2021, 2022; Hong et al., 2024; Jaeger et al., 2019, 2020), raising the possibility that exposure effects observed in White male faces may not generalize. Salient social categories may also compete for attention during person perception, potentially interfering with learning about other facial dimensions (cf. Klein & Sherman, 2024). Experiment 1 therefore employed a racially and gender diverse set of faces to address these issues.

Method

Transparency and Openness

Across all experiments, we report sample sizes, data exclusions, and demographic characteristics. Sensitivity analyses estimated the effects detectable with greater than 80% power for each experiment's focal analysis. Analyses were conducted using the *lme4* and *lmerTest* R packages (Bates et al., 2015; Kuznetsova et al., 2017). The experiments were not preregistered. Data, code, and model files are available at https://osf.io/yf4ar.

Participants

In Experiment 1, we aimed to recruit approximately 400 participants (200 per intervention condition) and collected data from 416 participants. After excluding those who failed two or more (out of five) attention checks (n=4), the final sample consisted of 412 participants (203 in the control condition and 209 in the exposure condition; M_{age}=37.81 years, SD_{age}=11.40 years; 211 male, 196 female, 5 non-binary or unspecified; 284 White, 72 Black, 34 Asian, 5 Native American, 17 unspecified; 368 not Hispanic, 38 Hispanic, 6 unspecified). A simulation-based sensitivity analysis demonstrated that Experiment 1 had more than 80% power to detect the critical facial trustworthiness × intervention interaction effect at $|b|$=0.02.

Stimuli

Stimuli consisted of 160 face images originating from two samples of 40 face images with neutral expressions originating from the Chicago Face Database (CFD; Ma et al., 2015). Each set of faces matched U.S. demographic base rates (U.S. Census Bureau, 2023), including 14 White, 6 Black, 2 Asian, and 8 Hispanic men and women, in each of the two sets. Considerable work demonstrates that variations in task-irrelevant featural dimensions can readily interfere with face-based statistical learning, especially when that variation is salient, categorical, and coherent (Richler et al., 2011; Chua et al., 2015). Varying face stimuli systematically in accordance with national demographic prevalences, may thereby avoid undue salience on any one category.

Target faces were transformed along a trustworthiness vector between high- and low-trustworthiness prototypes from Oosterhof and Todorov (2008). Relying on corresponding landmark template files for the prototypes (Jaeger et al., 2020) and CFD faces (Singh et al., 2023), we implemented the transformations in *Psychomorph* (Tiddeman et al., 2001). Each face was transformed to appear either highly trustworthy (+3 SD) or highly untrustworthy (−3 SD), resulting in two sets of racially and gender diverse sets of 80 faces (40 untrustworthy faces, 40 trustworthy faces). Face images were oval-cropped around the face to minimize the influence of external features such as hair, neck, and shoulders. Each set was randomly assigned to either the exposure (or control) task or the trust game.

Procedure

Experiment 1 consisted of two phases. The first phase involved a counter-stereotype exposure protocol or a control protocol. Counter-stereotype exposure followed the design of Chua & Freeman (2021; see also Hong et al., 2024). Participants were instructed to memorize the faces and behaviors of 20 individuals, with each face paired with a one-sentence behavioral description presented underneath their displayed face image. Behavioral descriptions conveyed either trustworthy or untrustworthy actions (e.g., volunteering at a homeless shelter vs. accepting

a bribe from a student). These behaviors were previously tested to ensure they differed reliably in perceived trustworthiness and were matched in strength (Chua & Freeman, 2021; Hong et al., 2024). Face-behavior pairings were sampled such that 80% of trustworthy-looking faces were paired with untrustworthy behaviors and 80% of untrustworthy-looking faces were paired with trustworthy behaviors. Trial order was randomized, and a 2,000 ms feedback page (i.e., "Please spend more time on each image") appeared if participants advanced in under 2,000 ms. Each pairing was shown three times, for a total of 60 learning trials. The design of the control condition was similar to that of the exposure condition, with the only difference being the information paired with the faces. Rather than behavioral descriptions, faces were paired with the target's supposed favorite color.[1] Participants were randomly assigned to one of the two conditions.

In the second phase, participants completed an economic trust game adapted from prior work (Bonnefon et al., 2013; Chua & Freeman, 2021; Rezlescu et al., 2012; van't Wout & Sanfey, 2008). Participants were told they were playing for real money and, on each round, received $1.00 and decided how much ($0-$1.00 in $0.25 increments) to entrust to a partner represented by a face avatar. Entrusted money was tripled, after which the partner could return some portion to the participant. Each round began after a randomized 0–5,000-ms delay intended to simulate locating another player. See Supplemental Materials for full details.

The trust game included 32 main rounds, on which participants were paired with 16 trustworthy-looking and 16 untrustworthy-looking partner faces. Critically, by sampling target faces from the other pre-constructed set of 80 faces, these target faces were distinct from those in

[1] We used favorite colors to avoid introducing unintended race and gender cues. The name labels used in previous work were exclusively male-gendered and appeared to be predominantly White-associated (Chua & Freeman, 2021; Hong et al., 2024), making them incompatible with our racially and gender-diverse stimulus set.

the learning phase, allowing assessment of generalization to novel identities. Five attention checks were randomly interspersed among the 32 main rounds.

At the end of each experiment, participants completed a well-validated, three-item scale assessing their explicit endorsement of facial stereotyping, the Physiognomic Belief Scale (PBS; Jaeger et al., 2020; 2022). Demographic information was also collected.

Analytic Approach

Trials with RTs indicating a lack of attention (RTs > 30,000 ms; ~1% of trials) were excluded. This exclusion criterion did not meaningfully alter our conclusions. Responses were analyzed using a linear mixed-effects model (LMEM) with by-participant and by-stimulus random intercepts, the latter accounting for the racial and gender diversity of the stimulus set and supporting generalization across facial identities. Fixed effects included contrast-coded variables for facial trustworthiness (-0.5=untrustworthy-looking, 0.5=trustworthy-looking) and intervention condition (-0.5=control, 0.5=exposure), along with their interaction. Fixed effects also included a main effect for PBS (mean-centered), as well as a facial trustworthiness × PBS interaction, to test whether facial stereotyping is stronger among those who endorse it to a greater extent as shown in prior work. Finally, a *t*-test examined whether counter-stereotype exposure may have reduced explicit endorsement (PBS) relative to the control condition.

Results

There was a significant main effect of facial trustworthiness, b=0.04, SE=.01, $t(91)$=5.98, p<.001, whereby participants entrusted more money to partners with trustworthy-looking faces than to those with untrustworthy-looking faces. There was no significant main effect of intervention, b=-0.03, SE=.03, $t(408)$=-0.87, p=.385, indicating that counter-stereotype exposure

did not alter overall levels of money entrusted compared to the control. More importantly, there was a significant facial trustworthiness × intervention interaction, b=-0.03, SE=.01, t(12,580)=-4.02, p<.001. Those in the counter-stereotype exposure condition demonstrated significantly less facial stereotyping, simple b=-0.02, SE=.01, t(147)=-3.40, p=.001, than those in the control condition, simple b=-0.05, SE=.01, t(151)=-7.15, p<.001 (see Figure 1), with the exposure condition reducing facial stereotyping by more than half.

Additional analyses confirmed that the impact of counter-stereotype exposure (facial trustworthiness × intervention effect) was stable across targets' race and gender (see Supplemental Materials for details).

There was also a significant main effect of PBS, b=-0.01, SE<0.01, t(408.0)=-3.14, p=.002, which was qualified by a significant facial trustworthiness × PBS interaction, b=-0.05, SE=.01, t(12,580)=4.65, p<.001, indicating that those who reported greater endorsement of facial stereotyping demonstrated greater facial stereotyping, consistent with prior findings.

Finally, we tested whether explicit endorsement of facial stereotyping (PBS) was affected by counter-stereotype exposure, directly comparing PBS scores between intervention conditions. No significant difference in PBS emerged following the control versus exposure conditions, t(408)=1.39, p=.164, suggesting that the reduced facial stereotyping observed in the exposure condition was not accompanied by a reduction in participants' explicit endorsement or beliefs about the validity of face stereotypes.

Figure 1. *Money Entrusted Between Counter-Stereotype Exposure and Control Conditions (Experiment 1)*. Facial trustworthiness effect (money entrusted to trustworthy- minus untrustworthy-looking faces) by condition in Experiment 1. Larger markers reflect condition means and error bars reflect 95% confidence intervals. Smaller markers reflect participant-level difference scores. The dashed line indicates an effect of zero.

Discussion

Counter-stereotype exposure substantially reduced facial stereotyping, and critically, it did so without changing explicit endorsement, consistent with exposure operating independently of explicit beliefs. The effect also generalized across a demographically diverse stimulus set in which race, gender, and facial trustworthiness varied simultaneously.

Experiment 2

In Experiment 1, counter-stereotype exposure reduced facial stereotyping without affecting explicit beliefs about facial stereotyping. Experiment 2 builds on this by introducing the first

direct comparison between counter-stereotype exposure and explicit instructions to avoid facial stereotyping. Alongside a control condition, all participants completed a trust game analogous to that of Experiment 1, affording a test of whether past research's differential success with these two interventions replicate when evaluated in the same experiment.

Method

Participants

We aimed to recruit 500 participants and collected data from 533 participants. After excluding those who failed two or more of four attention checks (n=10), the final sample consisted of 523 participants (179 control, 162 instruction, and 182 exposure conditions; M_{age}=37.08 years, SD_{age}=11.53 years; 273 female, 246 male, 4 non-binary or unspecified; 332 White, 136 Black, 28 Asian, 5 Native American, 1 Native Hawaiian or Other Pacific Islander, 21 unspecified; 455 not Hispanic, 56 Hispanic, 12 unspecified). A simulation-based sensitivity analysis demonstrated that Experiment 2 had more than 80% power to detect each of the critical facial trustworthiness × intervention (control vs. exposure; control vs. instruction) interaction effects at $|b|$=0.03.

Procedure

The first phase of Experiment 2 consisted of a face-learning task with control and exposure conditions identical to those in Experiment 1, and a third condition in which participants were explicitly instructed to avoid facial stereotyping. In this instructions condition, participants were presented faces accompanied by arbitrary behavioral descriptions and asked to form an impression of how trustworthy each individual seemed (1–7 Likert-type scale; 1="Extremely Untrustworthy," 7="Extremely Trustworthy"). Behaviors were selected to be only weakly-to-moderately diagnostic of trustworthiness to avoid establishing any strong statistical associations

(Mickelberg et al., 2022).[2] After 14 trials, participants were interrupted with a message informing them that their responses reflected a bias toward perceiving certain faces as more trustworthy than other faces and that relying on facial appearance tends to be inaccurate and harmful (regardless of their actual responses or bias in the task). They were then instructed to avoid relying on trustworthiness-related facial appearance for the remainder of the experiment. Example images of trustworthy- and untrustworthy-looking faces were provided to help them recognize these cues (see Supplemental Materials for verbatim script).

The trust game in the second phase of the experiment followed Experiment 1, except that all participants saw randomly paired biographical information (a partner player's hometown and number of siblings) alongside each face, providing an alternative, nonfacial basis for decisions among participants instructed to avoid facial information; this biographical information was arbitrary and unrelated to trustworthiness or behavior. To further increase realism, participants reported their hometown and number of siblings before starting the game, implying that their partners would see the participant's information as well. Finally, whereas Experiment 1 included five response options, Experiment 2 included four, omitting the midpoint (i.e., $0.50) so that responses reflected a directional choice toward relatively greater trust or distrust. Four attention checks were included. Following the trust game, PBS and demographics were collected.

Analytic Approach

Analyses followed those of Experiment 1, with two changes to the model: two intervention contrasts (control vs. exposure; control vs. instruction) replaced the single exposure contrast, and a second by-stimulus random intercept was added for the biographical information paired with

[2] Mickelberg et al. (2022) specifically assessed moral vs. immoral behaviors, but morality impressions closely align with trustworthiness impressions (Brambilla et al., 2011; Jaeger et al., 2022; analyzed ratings of behavioral descriptions are presented in the Supplementary Materials).

each partner face. Explicit endorsement (PBS) was compared across all three conditions using a one-way ANOVA rather than a t-test.

Results

There was a significant main effect of facial trustworthiness, b=0.04, SE=.01, t(88)=7.19, p<.001, once again demonstrating facial stereotyping, with participants entrusting more money to partners with trustworthy-looking faces than untrustworthy-looking faces. Responses did not overall differ from those in the control condition due to either counter-stereotype exposure, b=0.02, SE=0.04, t(517)=0.49, p=.623, or direct instruction, b=0.02, SE=0.04, t(517)=0.40, p=.688. Critically, compared to the control condition (simple b=0.06, SE=.01, t(263)=7.57, p<.001), counter-stereotype exposure (simple b=0.03, SE=.01, t(258)=3.86, p<.001) reduced facial stereotyping, b=-0.02, SE=.01, t(15,930)=-2.53, p=.012, cutting its effect approximately in half. However, this reduction failed to emerge for the instruction condition (simple b=0.04, SE=.01, t(304)=4.92, p<.001), b=-0.01, SE=.01, t(15,940)=-0.60, p=.548 (see Figure 2). As in Experiment 1, there was also a significant main effect of PBS, b=-0.01, SE<.01, t(517)=-2.42, p=.016, and a significant facial trustworthiness × PBS interaction, b<.01, SE<.01, t(15,950)=4.09, p<.001, indicating that facial stereotyping was once again magnified for participants who reported greater endorsement of it.

Finally, a one-way ANOVA showed no effect of intervention on facial stereotyping endorsement, F(2, 520)=0.28, p=.755, η_p^2<.001. Tukey-adjusted pairwise comparisons showed no difference in PBS between those in the control condition and those in either the exposure condition, t(520)=−0.50, p=.870, d=−0.05, 95% CI [−0.26, 0.15], or the instruction condition, t(520)=−0.73, p=.746, d=−0.08, 95% CI [−0.29, 0.13].

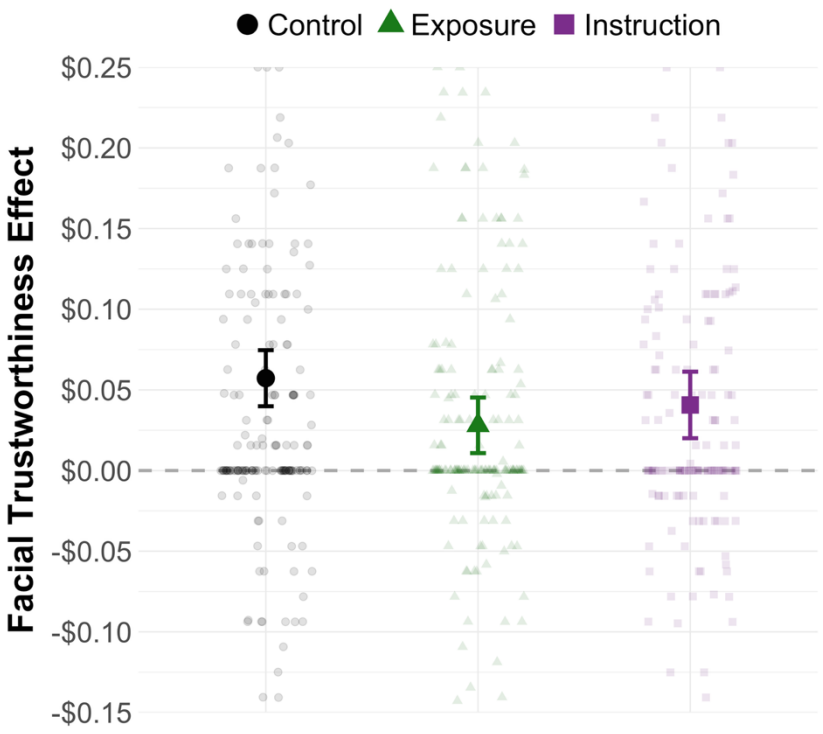

Figure 2. *Money Entrusted Between Intervention Conditions (Experiment 2)*. Facial trustworthiness effect (money entrusted to trustworthy- minus untrustworthy-looking faces) by condition in Experiment 2. Larger markers reflect condition means and error bars reflect 95% confidence intervals. Smaller markers reflect participant-level difference scores. The dashed line indicates an effect of zero.

Discussion

Counter-stereotype exposure again reduced facial stereotyping relative to the control condition. By contrast, explicit instructions did not significantly reduce facial stereotyping. It is notable that explicit endorsement of facial stereotyping did not vary across conditions. Even informing participants that their judgments exhibited inaccurate and harmful facial stereotyping, and instructing them to deliberately avoid it, did not alter their beliefs in the validity of face

stereotypes. This highlights the profound resilience of explicit beliefs that endorse facial stereotyping.

Experiment 3

The results thus far raise the possibility that the two interventions differ not only in effectiveness, but also in their dependence on available cognitive resources. Counter-stereotype exposure may engage processes, such as relatively automatic associative processing, that remain robust under relatively high cognitive demands, whereas explicit instructions may rely on more resource-intensive deliberative control over responses. Experiment 3 tested this possibility by imposing relatively high cognitive demands through both time pressure and a concurrent irrelevant task. If the effectiveness of counter-stereotype exposure is relatively resource-independent, it should continue to reduce facial stereotyping. By contrast, if explicit instructions depend on resource-intensive deliberative control, their effects may be especially vulnerable under load, likely failing to reduce facial stereotyping—or even ironically increasing the effect (Galinsky & Moskowitz, 2007; Wang et al., 2020; Wegner, 1994).

Method

Participants

We aimed to recruit at least 750 participants and collected data from 819 participants. After excluding 122 participants (14.9%) for failing both one-back trials, our final sample was comprised of 697 participants (228 control, 227 instruction, and 242 exposure conditions; M_{age}=40.80 years, SD_{age}=20.90 years; 387 female, 303 male, 7 non-binary or unspecified; 499 White, 122 Black, 45 Asian, 6 Native American, 25 unspecified; 633 not Hispanic, 53 Hispanic, 11 unspecified). A simulation-based sensitivity analysis demonstrated that Experiment 3 had

more than 80% power to detect the two crucial facial trustworthiness × intervention (control vs. exposure; control vs. instruction) interaction effects at $|b|$=0.02.

Procedure

The procedure of Experiment 3 largely followed that of Experiment 2 aside from the introduction of cognitive demands, through response-time pressure and the inclusion of occasional one-back trials that required carefully monitoring for repeated stimuli. To apply time pressure, we instructed participants to respond in in under 3,000 ms, roughly the median response time during the trust game in Experiment 2. Responses slower than the deadline initiated feedback to "Please respond faster next time!" Responses faster than 500 ms also initiated feedback to "Please take your time to consider your choices."

For the occasional one-back trials, participants were told that some rounds were "fake" and that the immediately preceding partner would reappear with different, fictitious biographical information. Participants were told that these "fake" rounds were meant to test whether they were paying attention. They were instructed to detect these repetitions and respond by clicking the partner's face rather than a monetary response option. Participants completed eight practice rounds (six regular, two one-back), followed by 40 regular trials and two randomly interspersed one-back trials. No additional attention checks were included.

Analytic Approach

Due to the speeded nature of the task, responses with RTs faster than 300 ms or slower than 5,000 ms were excluded from analyses. The analyses otherwise followed those of Experiment 2.

Results

There was a significant main effect of facial trustworthiness, b=0.02, SE<.01, $t(83)$=6.27, p<.001, demonstrating facial stereotyping. The money entrusted did not significantly differ from

the control condition due to either counter-stereotype exposure, b=-0.03, SE=0.03, t(691)=-0.88, p=.379, or direct instruction, b=-0.04, SE=0.03, t(691)=-1.32, p=.188. Critically, compared to the control condition (simple b=0.01, SE<.01, t(381)=3.22, p=.001), counter-stereotype exposure (simple b=0.01, SE<.01, t(337)=2.82, p=.005) once again reduced facial stereotyping, b=-0.01, SE<.01, t(26740)=-2.17, p=.030. Unlike the instruction condition's null effect in Experiment 2, here with high cognitive demands, direct instructions (simple b=0.03, SE<.01, t(388)=6.83, p<.001) *increased* the degree of facial stereotyping relative to control, b=0.02, SE<.01, t(26750)=3.55, p<.001, more than doubling its effect (see Figure 3). As in Experiments 1 & 2, there was also a significant main effect of PBS, b=-0.01, SE<.01, t(691)=-3.26, p=.001, and a significant facial trustworthiness × PBS interaction, b<.01, SE<.01, t(26,750)=5.27, p<.001, with stronger endorsers showing exacerbated facial stereotyping.

Finally, a one-way ANOVA revealed a significant effect of intervention condition on PBS scores, F(2, 694)=14.63, p <.001, η_p^2 =.040. Tukey-adjusted pairwise comparisons showed no difference in PBS between exposure and control conditions, t(694)=2.03, p=.107, d=0.15, 95% CI [0.00, 0.30]. However, participants in the instruction condition did show reduced endorsement relative to those in the control condition, t(694)=5.35, p<.001, d=0.41, 95% CI [0.26, 0.56].

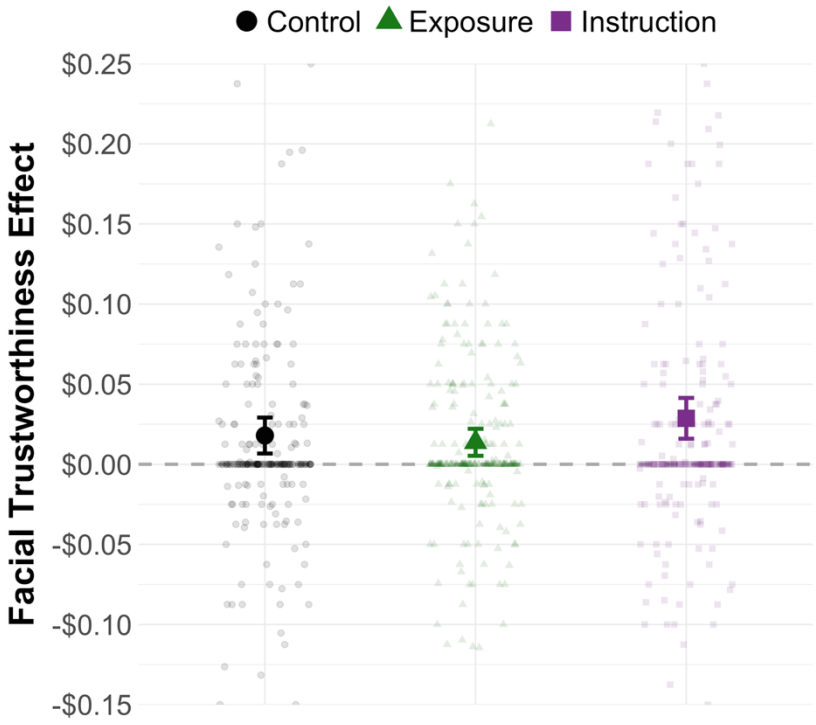

Figure 3. *Money Entrusted Between Intervention Conditions (Experiment 3)*. Facial trustworthiness effect (money entrusted to trustworthy- minus untrustworthy-looking faces) by condition in Experiment 3. Larger markers reflect condition means and error bars reflect 95% confidence intervals. Smaller markers reflect participant-level difference scores. The dashed line indicates an effect of zero.

Discussion

Under high cognitive demands, counter-stereotype exposure once again reduced facial stereotyping. In contrast, direct instructions not only failed to reduce facial stereotyping but exacerbated it. This effect is consistent with the engagement of resource-intensive deliberative control, which tends to backfire under increased cognitive demands (Wang et al., 2020). Notably, those instructed to control facial stereotyping under high cognitive demand demonstrated more facial stereotyping bias despite explicitly endorsing it *less*. This is consistent with a form of

compensatory overcorrection documented in ironic process theories (Wegner, 1994; Galinsky & Moskowitz, 2007).

General Discussion

Faces cue inferences of personality despite being poor signals of actual behavior (Krendl et al., 2014; Rule et al., 2013). Such facial stereotyping often resists deliberate control (Jaeger et al., 2019, 2020), yet repeated exposure to exemplars that violate face stereotypes can reliably reduce it (Chua & Freeman, 2021; Hong et al., 2024). The present research offers the first direct comparison of two intervention approaches, counter-stereotype exposure and explicit instructions to control the bias: exposure consistently reduced facial stereotyping whereas instruction did not. Their divergence implicates perceivers' explicit beliefs and the cognitive resources demanded by the processes each intervention engages.

Across all experiments, self-reported endorsement predicted stronger facial stereotyping (e.g., Jaeger et al., 2020). Yet counter-stereotype exposure strongly reduced facial stereotyping without reducing that endorsement, indicating that successful intervention need not first change perceivers' explicit beliefs about whether facial appearance is a valid cue to personality. The instruction condition sharpened this dissociation. In Experiment 2, telling participants their judgments reflected bias and instructing them to avoid it reduced neither stereotyping nor endorsement. In Experiment 3, under greater cognitive demand, instructions reduced explicit endorsement yet increased facial stereotyping, so behavior and stated belief moved in opposite directions. Participants may have recognized that they ought to disavow facial stereotyping while still failing to control it in behavior, consistent with evidence that people reaffirm egalitarian

standards once aware of a gap between those standards and their responses (Dunton & Fazio, 1997; Monteith, 1993; Plant & Devine, 1998); future research could test this possibility directly.

The divergent effects of the two interventions are consistent with different underlying mechanisms. Counter-stereotype exposure may reduce facial stereotyping by retuning associations between facial features and trait inferences, whereas explicit instruction requires perceivers to monitor and control the influence of those associations. Exposure remained effective under high cognitive demands, leaving explicit endorsement unchanged. Explicit instruction, by contrast, was ineffective under more modest demands (Experiment 2) and became counterproductive under heavier demands, increasing the very bias it targeted (Experiment 3). This reversal accords with evidence that deliberate control of unwanted responses is resource-intensive (Braver, 2012) and can rebound when cognitive capacity is taxed, ironically amplifying the response it is intended to suppress (Galinsky & Moskowitz, 2007; Wang et al., 2020; Wegner, 1994).

The role of explicit endorsement may further help explain why facial stereotyping poses a distinctive challenge for deliberative intervention. Unlike racial and gender bias, which people typically recognize as undesirable and may experience guilt or discomfort about (Monteith, 1993; Monteith et al., 2002), facial stereotyping is readily endorsed by many individuals (Jaeger et al., 2022). Recent research shows endorsement of facial stereotyping is consistently associated with a general faith in one's own intuitions (Jaeger et al., 2025). Such intuitive thinking is widely valued as authentic, natural, and wise (Alter & Oppenheimer, 2009; Epstein, 1994; Gigerenzer, 2007; Usher et al., 2011). Thus, asking perceivers to avoid facial stereotyping may require them to discount impressions that feel immediate, perceptual, and intuitive, rather than merely

suppress a response they already understand is inappropriate. This distinction may help situate the findings relative to interventions targeting more widely studied racial and gender bias, for which both associative and non-associative approaches can be effective (Calanchini et al., 2021; Röhner & Lai, 2021). Instructions to set aside such group-based biases can appeal to egalitarian values perceivers might already hold; instructions to set aside facial stereotyping may first need to confront the fact that perceivers do not necessarily represent the bias as a bias at all. Counter-stereotype exposure sidesteps this obstacle: rather than persuading people their intuitions are wrong, it alters the associations underlying them.

This research also provides the first evidence that counter-stereotype exposure can reduce facial stereotyping across demographically diverse targets. Variation in salient, categorical, and task-irrelevant dimensions interferes with statistical learning in face perception (Richler et al., 2011; Chua et al., 2015). We suspect that distributing race and gender categories in line with demographic base rates likely minimized their undue salience and preserved statistical learning of counter-stereotype associations. Future research should expand on such methodological factors affecting the success or failure of different interventions in diverse contexts.

Some limitations are worth noting. First, we examined only facial trustworthiness, a dimension closely tied to emotion overgeneralization (Zebrowitz & Montepare, 2006); less emotion-linked traits, which tend to be more strongly endorsed as physiognomic (Jaeger et al., 2022), might prove more tractable to instruction. Additionally, endorsement was measured as trait-general, potentially masking trustworthiness-specific belief change. Future work should test whether counter-stereotype exposure operates across a broader space of facial trait dimensions and how it relates to explicit endorsement of those dimensions. Moreover, the distinction

between associative versus deliberative processes offers a parsimonious explanation of why exposure succeeded where instruction failed, but future work should examine these processes more comprehensively (e.g., examining awareness).

In summary, facial stereotyping may be more effectively mitigated by changing the underlying associations that give rise to biased impressions than by asking perceivers to deliberately override them. Such findings highlight the importance of matching intervention strategies to the different psychological mechanisms underlying the biases we seek to change.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review*, *13*(3), 219-235.

Amodio, D. M., & Swencionis, J. K. (2018). Proactive control of implicit bias: A theoretical model and implications for behavior change. *Journal of Personality and Social Psychology*, *115*(2), 255.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269.

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

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General*, *142*(1), 143.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European journal of social psychology*, *41*(2), 135-143.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, *16*(2), 106-113.

Calanchini, J., Lai, C. K., & Klauer, K. C. (2021). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology*, *121*(4), 796.

Chua, K. W., Richler, J. J., & Gauthier, I. (2015). Holistic processing from learned attention to parts. *Journal of Experimental Psychology: General*, *144*(4), 723.

Chua, K. W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. *Social Psychological and Personality Science*, *12*(7), 1335-1344.

Chua, K. W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology*, *98*, 104225.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316-326.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, *49*(8), 709.

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.

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *The Journal of Neuroscience*, *34*(32), 10573-10581.

Galinsky, A. D., & Moskowitz, G. B. (2007). Further ironies of suppression: Stereotype and counterstereotype accessibility. *Journal of Experimental Social Psychology*, *43*(5), 833-841.

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.

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, *13*(3), 1-33.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Hong, Y., & Freeman, J. B. (2024). Shifts in facial impression structures across group boundaries. *Social Psychological and Personality Science*, *15*(6), 619-629.

Hong, Y., Chua, K. W., & Freeman, J. B. (2024). Reducing facial stereotype bias in consequential social judgments: intervention success with white male faces. *Psychological Science*, *35*(1), 21-33.

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology: General*, *148*(6), 1008.

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 104004.

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2022). Understanding the role of faces in person perception: Increased reliance on facial appearance when judging sociability. *Journal of Experimental Social Psychology*, *100*, 104288.

Jaeger, B., Evans, A., Stel, M., & van Beest, I. (2025). Intuitive thinking is associated with stronger belief in physiognomy and confidence in the accuracy of facial impressions.

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance?. *Journal of Personality and Social Psychology*, *111*(5), 655.

Klein, S. A., & Sherman, J. W. (2024). Measuring the impact of multiple social cues to advance theory in person perception research. *Psychological Review*.

Krendl, A. C., Rule, N. O., & Ambady, N. (2014). Does aging impair first impression accuracy? Differentiating emotion recognition from complex social inferences. *Psychology and aging*, *29*(3), 482.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

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.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, *47*(4), 1122-1135.

Mickelberg, A., Walker, B., Ecker, U. K., Howe, P., Perfors, A., & Fay, N. (2022). Impression formation stimuli: A corpus of behavior statements rated on morality, competence, informativeness, and believability. *PLoS One*, *17*(6), e0269393.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of personality and social psychology*, *65*(3), 469.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: on the development and operation of cues for control. *Journal of personality and social psychology*, *83*(5), 1029.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in cognitive sciences*, *18*(11), 566-570.

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

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, *75*(3), 811.

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.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological science*, *22*(4), 464-471.

Röhner, J., & Lai, C. K. (2021). A diffusion model approach for understanding the impact of 17 interventions on the race Implicit Association Test. *Personality and Social Psychology Bulletin*, *47*(9), 1374-1389.

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.

Singh, B., Gambrell, A., & Correll, J. (2023). Face templates for the Chicago face database. *Behavior Research Methods*, *55*(2), 639-645.

Tiddeman, B., Burt, M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE computer graphics and applications*, *21*(5), 42-50.

U.S. Census Bureau. (2023). *American Community Survey 1-year estimates, Table B03002: Hispanic or Latino origin by race*

Usher, M., Russo, Z., Weyers, M., Brauner, R., & Zakay, D. (2011). The impact of the mode of thought in complex decisions: Intuitive decisions are better. *Frontiers in psychology*, *2*, 37.

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.

Wang, D., Hagger, M. S., & Chatzisarantis, N. L. (2020). Ironic effects of thought suppression: A meta-analysis. *Perspectives on Psychological Science*, *15*(3), 778-793.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological review*, *101*(1), 34.

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

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science*, *26*(8), 1325-1331.

Wilson, J. P., & Rule, N. O. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes: The role of facial trustworthiness. *Social Psychological and Personality Science*, *7*(4), 331-338.

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.

Xie, S. Y., Flake, J. K., Stolier, R. M., Freeman, J. B., & Hehman, E. (2021). Facial impressions are predicted by the structure of group stereotypes. *Psychological Science*, *32*(12), 1979-1993.

Zebrowitz, L. A. (1996). Physical appearance as a basis of stereotyping. *Stereotypes and stereotyping*, 79-120.

Zebrowitz, L. A., & Montepare, J. (2006). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. In *Evolution and Social Psychology* (pp. 81-113). Psychology Press.