Supplementary Materials

Reducing the reliance on facial stereotypes in consequential social judgments: Intervention success with White male faces

Youngki Hong[1], Kao-Wei Chua[1], & Jonathan B. Freeman[1]

[1]Columbia University

Supplemental Methods: Training Paradigm (Studies 1-4)

Behavioral descriptions. We used 10 trustworthy and 10 untrustworthy behaviors for training (Supplementary Table 1; Chua & Freeman, 2021, 2022) in all four studies. 30 independent Mechanical Turk raters judged each behavioral description on trustworthiness using a 7-point Likert scale to confirm the sentences' intended levels of trustworthiness. The agreement between raters was strong (intraclass correlation coefficient [ICC] = 0.92). The trustworthy behavioral descriptions were rated as significantly more trustworthy (M = 5.03) than the untrustworthy behavioral descriptions (M = 2.83), $t(18) = 19.75$, $p < 0.0001$, Cohen's $d = 9.31$). Both trustworthy and untrustworthy behaviors were similarly balanced and did not significantly differ in their distance from the midpoint of the Likert scale, $t(18) = 1.30$, $p = 0.21$, Cohen's $d = 0.61$.

Supplementary Table 1. Training behaviors and their ratings on trustworthiness.

Behavior	Behavioral description	Trustworthiness
Trustworthy	Surprised their significant other at work with flowers	4.77
Trustworthy	Volunteered at a homeless shelter	5.03
Trustworthy	Helped an elderly person cross a street	4.93
Trustworthy	Helped their friend plan a birthday party for their child	4.77
Trustworthy	Visited a sick friend at the hospital	4.73
Trustworthy	Let a friend stay on their couch who lost their apartment	5.03
Trustworthy	Returned $20 to someone who dropped it	5.63
Trustworthy	Performed a surgery free for someone who couldn't afford	5.13
Trustworthy	Let a friend win at cards because they had no money.	5.30
Trustworthy	Protected their little brother from bullies	4.98
Untrustworthy	Rigged a lottery to steal from old people	2.33
Untrustworthy	Spat in another person's face	2.80
Untrustworthy	Threw a rock at a neighbor's window	2.90
Untrustworthy	Screamed at a scared kindergartener	2.87
Untrustworthy	Sprayed curse words on someone's fence	3.00
Untrustworthy	Ate their friend's leftovers from the refrigerator	3.10
Untrustworthy	Took a bribe to give a student a better grade	2.70
Untrustworthy	Skipped a work shift they committed to covering	3.03
Untrustworthy	Cheated on their spouse while on a business trip	2.67
Untrustworthy	Got a promotion by lying about coworkers	2.85

Supplemental Analysis S1

To explore the effects of the training (control = -0.5, training = 0.5) on trustworthiness ratings (Study 1A), attractiveness ratings (Study 1B), sentencing recommendations (Study 2), and implicit trustworthiness evaluations (Study 3) directly, depending on targets' real-world sentencing outcome (-0.5 = life in prison, 0.5 = death), we used linear mixed-effects models to predict these dependent measures. An additional advantage of these supplemental analyses is that the logistic mixed-effects models that predicted real-world sentencing outcome as the dependent measure in Studies 1-3 were not able to specify random slopes for stimuli due to the one-to-one mapping between the stimulus and the dependent measure. Here we circumvent that problem, allowing maximal specification of the random effects in our design (Barr et al., 2013). Thus, these models allowed for random intercepts and random slopes of the training condition for both participants and stimuli.

***Study 1A*.** The main effect of sentencing outcome was significant, $b = -.06$, $SE = .03$, $z = 2.12$, $p = .03$, 95% CI [-.12, -.00], indicating that inmates who were sentenced to death tended to be rated more untrustworthy than those sentenced to life in prison. The main effect of training was not significant, $b = -.23$, $SE = .16$, $z = 1.43$, $p = .15$, 95% CI [-.54, .08]. Critically, there was a significant interaction, $b = .09$, $SE = .02$, $z = 5.85$, $p < .001$, 95% CI [.06, .12]. The control participants rated inmates who were sentenced to death significantly less trustworthy ($M = 3.48$, $SE = .12$) than those who were sentenced to life in prison ($M = 3.58$, $SE = .12$), $b = .10$, $SE = .03$, $z = 3.61$, $p < .001$, 95% CI [.05, .16]. However, the trained participants rated inmates who were sentenced to death ($M = 3.31$, $SE = .11$) equally trustworthy as those were sentenced to life in prison ($M = 3.30$, $SE = .11$), $b = .01$, $SE = .03$, $z = .47$, $p = .64$, 95% CI [-.04, .07].

Study 1B. The main effect of sentencing outcome was not significant, b = -.05, SE = .04, z = 1.25, p = .21, 95% CI [-.12, .03], indicating that inmates with different sentencing outcomes were not rated differently on attractiveness. The main effect of training was significant, b = .50, SE = .15, z = 3.40, p < .001, 95% CI [.21, .79], as participants in the training condition rated inmates higher on attractiveness overall than participants in the control condition did. Critically, the interaction was not significant, b = .01, SE = .02, z = .21, p = .83, 95% CI [-.04, .05]. The control participants rated inmates who were sentenced to death as less attractive (M = 3.07, SE = .11) than those who were sentenced to life in prison (M = 3.12, SE = .11), although the difference was not significant, b = .05, SE = .04, z = 1.23, p = .22, 95% CI [-.03, .13]. The same pattern of results was found among the trained participants. Trained participants rated inmates who were sentenced to death (M = 3.58, SE = .11) as less attractive than those were sentenced to life in prison (M = 3.63, SE = .11), although the difference was not significant, b = .05, SE = .04, z = 1.14, p = .25, 95% CI [-.03, .13].

Combining data from Studies 1A and 1B and conducting a three-way interaction analysis revealed that the absence of a significant interaction in attractiveness ratings (Study 1B) significantly differed from the significant interaction observed in trustworthiness ratings (Study 1A). Specifically, we used a logistic mixed effects model to predict ratings based on sentencing outcome (-0.5 = life sentence, 0.5 = death), training condition (control = -0.5, training = 0.5), trait (trustworthiness-Study 1A = -0.5, attractiveness-Study 1B = 0.5), and their interactions. We found a significant three-way interaction involving sentencing outcome, training condition, and trait (b = .09, SE = .02, z = 2.24, p = .02, CI [.05, .13]) indicating that the two-way interaction between sentencing outcome and training condition was highly significant for trustworthiness

ratings (Study 1A) (b = .09, SE = .02, z = 5.31, p < .001, 95% CI [.06, .13]), but not for attractiveness ratings (Study 1B) (b = .01, SE = .02, z = .30, p = .77, 95% CI [-.03, .04]).

Study 2. The main effect of sentencing outcome was significant, b = .08, SE = .02, z = 3.49, p < .001, 95% CI [.03, .12], indicating that participants tended to recommend harsher sentences for inmates sentenced to death vs. to life in prison. The main effect of training was not significant, b = .04, SE = .10, z = .34, p = .73, 95% CI [-.17, .24]. Critically, there was a significant interaction, b = .06, SE = .02, z = 3.31, p = .001, 95% CI [.02, .10]. Control participants recommended inmates who were sentenced to death to receive significantly harsher sentences (M = 4.63, SE = .06) than those who were sentenced to life in prison (M = 4.50, SE = .06), b = .14, SE = .03, z = 4.41, p < .001, 95% CI [.08, .20]. The trained participants also recommended inmates who were sentenced to death to receive significantly harsher sentences (M = 4.57, SE = .09) than those who were sentenced to life in prison (M = 4.49, SE = .09), b = .08, SE = .02, z = 3.49, p < .001, 95% CI [.03, .12], although the magnitude of difference among the trained participants was smaller than among the control participants.

Study 3. The main effect of sentencing outcome was not significant, b = 4.45, SE = 3.28, z = 1.36, p = .17, 95% CI [-1.97, 10.88], indicating that inmates who were sentenced to death elicited implicit trustworthiness evaluations that did not significantly differ from those who were sentenced to life in prison. The main effect of training was not significant, b = 4.57, SE = 6.41, z = .71, p = .48, 95% CI [-8.00, 17.14]. Critically, there was a significant interaction, b = .17.49, SE = 6.76, z = 2.59, p = .01, 95% CI [4.24, 30.73]. Among control participants, inmates who were sentenced to death elicited implicit evaluations that were significantly less trustworthy (M = 3.55 ms, SE = 5.68 ms) than those who were sentenced to life in prison (M = 16.75 ms, SE =

5.67 ms), b = 13.20, SE = 4.72, z = 2.80, p = .005, 95% CI [3.95, 22.44]. However, among trained participants, implicit trustworthiness evaluations did not differ between inmates who were sentenced to death (M = 7.73 ms, SE = 4.43 ms) vs. life in prison (M = 3.44 ms, SE = 4.44 ms), b = -4.29, SE = 4.70, z = .91, p = .36, 95% CI [-13.50, 4.92].

Supplemental Analysis S2

Study 1A found a significant interaction between trustworthiness ratings and training condition, while Study 1B found a non-significant interaction between attractiveness ratings and training condition. To establish that the two interactions are significantly different from each other, we combined the datasets from the respective studies and used a logistic mixed effects model to predict sentencing outcome (0 = life sentence, 1 = death) based on rating, training condition (control = -0.5, training = 0.5), trait (trustworthiness-Study 1A = -0.5, attractiveness-Study 1B = 0.5), and their interactions. The models allowed for random intercepts for participants and random slopes of rating for participants.

The main effect of rating was significant, $b = -.04$, $SE = .01$, $z = 6.60$, $p < .001$, $OR = .97$, 95% CI [.96, .98], indicating that faces that were rated lower on trustworthiness or attractiveness were more likely to belong to individuals who were sentenced to death than life in prison. We also found a significant interaction between rating and training condition, $b = .03$ $SE = .01$, $z = 2.93$, $p = .003$, $OR = 1.03$, 95% CI [1.01, 1.05], which was driven by the trustworthiness ratings (Study 1A). Specifically, a significant three-way interaction among rating, training condition, and trait, $b = -.05$, $SE = .02$, $z = 2.24$, $p = .02$, $OR = .95$, CI [.91, .99] indicated that the two-way interaction between rating and training condition was highly significant for trustworthiness ratings (Study 1A), $b = .06$ $SE = .02$, $z = 3.66$, $p = < .001$, $OR = 1.06$, 95% CI [1.03, 1.09], but not significant for attractiveness ratings (Study 1B), $b = .01$ $SE = .02$, $z = .50$, $p = .62$, $OR = 1.01$, 95% CI [.98, 1.04]. No other effects were significant ($p > .05$).

Supplemental Analysis S3

In Study 3, although we expected accuracy in categorization of the target words to be very high and not to differ by the condition, for completeness we examined categorization accuracy using a logistic mixed effects model (0 = incorrect, 1 = correct) based on the sentencing condition of the facial primes (-0.5 = life in prison, 0.5 = death sentence), training condition (-0.5 = control, 0.5 = training), and their interaction. The model included random intercepts for participants and stimuli, and random slopes of sentencing condition for participants. The main effect of sentencing was not significant, b = .0003, SE = 04, z = .01, p = .99, OR = 1.00, 95% CI [.93, 1.07], nor was the main effect of training, b = .04, SE = .12, z = .35, p = .73, OR = 1.04, 95% CI [.82, 1.33]. The interaction was also not significant, b = .04, SE = .06, z = .73, p = .47, OR = .96, 95% CI [.86, 1.07]. There were comparable levels of high accuracy across all conditions: $\text{control}_{\text{death}}$ (M = 94.12%, SE = .82%), $\text{control}_{\text{life-in-prison}}$ (M = 94.21%, SE = .77%), $\text{trained}_{\text{death}}$ (M = 94.70%, SE = .69%), and $\text{trained}_{\text{life-in-prison}}$ (M = 94.94%, SE = .66%).

Supplemental Analysis S4

To test whether participant exclusions were consistent across conditions in Studies 1-4, we conducted chi-squared tests. First, we tested whether there was any difference in the number of participants excluded between the control and the training conditions. Because we had very few participants excluded in each study, we summed all the numbers of excluded participants in each condition across the four studies. This included 39 participants excluded from the control condition (11, 5, 12, 8, and 3 participants from Studies 1A, 1B, 2, 3, and 4), and 32 participants excluded from the training condition (9, 7, 7, 7, and 2 participants from Studies 1A, 1B, 2, 3, and 4). A chi-squared goodness of fit test revealed that the number of excluded participants did not differ between conditions, χ^2 (1) = .69, p = .41. Next, we conducted a chi-squared test of independence to examine whether the number of excluded participants for the two conditions systematically differed between studies. We found no evidence of dependence between conditions and studies, χ^2 (1) = 1.44, p = .84. These results indicate that participants were excluded consistently across experimental conditions and studies.

Additionally, because Study 3 included by-trial exclusions, we examined whether the number of excluded trials differed between conditions. The number of excluded trials for the control condition (M = 19.6, SE = 2.46) did not significantly differ from the number of excluded

trials for the training condition (M = 19.35, SE = 2.29), t(383) = .07, p = .94, 95% CI [-6.35, 6.87], Cohen's d = .01.

Supplemental Discussion: Reproducibility of Wilson & Rule (2015)

Some readers may note a discrepancy between the consistent replication of Wilson and Rule (2015) in our Studies 1-3, i.e., a significant relationship between facial trustworthiness and real-world sentencing outcome, and Kramer and Gardner's (2020) reported inability to replicate this effect.

Kramer & Gardner (K&G) conducted three studies, of which their Studies 1 and 3 are most relevant for this discussion. Wilson & Rule's (W&R) original Study 1 analyzed a large set of 742 inmates' faces. In our Studies 1-3, we used a randomized subset of 400 faces from W&R's Study 1. One of K&G's studies was a conceptual replication of W&R's Study 1 using 44 new inmate faces. The other relevant K&G study conducted a direct replication of W&R's original Study 2. This Study 2 involved only a small subset of 37 inmates who had been sentenced to death vs. life in prison and who were subsequently exonerated with DNA evidence. In both of K&G's replication studies, they reported null effects.

One possibility is that, by using small samples of faces, K&G's studies may have suffered from issues of restricted range. There are even some signs of this in W&R's original Study 2. Unlike the significant relationship in W&R's original Study 1 (analyzing 742 faces), which was

robust after adjusting for confounding variables ($p = .005$), this same relationship was only marginally significant after adjusting for confounding variables ($p = .06$) in W&R's original Study 2 (analyzing only 37 faces). In short, K&G did not perform a direct replication of W&R's Study 1 that analyzed a sufficiently large sample of faces and whose stimuli served as the basis for our Studies 1-3. The weak effect in W&R's original Study 2 and the null effects of K&G's direct replication and conceptual replication may have all suffered from issues of restricted range due to the small samples of faces analyzed, leading to Type II errors.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 10.1016/j.jml.2012.11.001.

Chua, K.-W., & Freeman, J. B. (2021). Facial Stereotype Bias Is Mitigated by Training. *Social Psychological and Personality Science*, 1948550620972550.

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.

Kramer, R. S. S., & Gardner, E. M. (2020). Facial Trustworthiness and Criminal Sentencing: A Comment on Wilson and Rule (2015). *Psychological Reports, 123*(5), 1854–1868.