

THIS MANUSCRIPT IS CURRENTLY UNDER REVIEW

Supplementary Materials

Mitigating Facial Stereotyping:

Counter-Stereotype Exposure Succeeds Where Explicit Instruction Fails

Public repository of data and analysis code: <https://osf.io/yf4ar>

Table of Contents

Descriptive Statistics	3
Supplementary Data Analysis	4
Target Race	4
Target Gender	4
Sensitivity Analysis	8
Behavioral Descriptions for Instructions Condition	11
Stimuli	11
Results	11
Condition and Trust Game Instructions	13
Exposure Condition	13
Control Condition	13
Instruction Condition	13
Trust Game	15
References	18

Descriptive Statistics**Table S1***Money entrusted by condition and facial trustworthiness (Experiment 1).*

Condition	Trustworthiness	<i>M</i>	<i>SD</i>
Control	Trustworthy	0.60	0.36
Control	Untrustworthy	0.54	0.37
Exposure	Trustworthy	0.56	0.35
Exposure	Untrustworthy	0.54	0.36

Table S2*Money entrusted by condition and facial trustworthiness (Experiment 2).*

Condition	Trustworthiness	<i>M</i>	<i>SD</i>
Control	Trustworthy	0.56	0.38
Control	Untrustworthy	0.50	0.38
Exposure	Trustworthy	0.57	0.38
Exposure	Untrustworthy	0.54	0.38
Instruction	Trustworthy	0.57	0.36
Instruction	Untrustworthy	0.53	0.37

Table S3*Money entrusted by condition and facial trustworthiness (Experiment 3).*

Condition	Trustworthiness	<i>M</i>	<i>SD</i>
Control	Trustworthy	0.63	0.35
Control	Untrustworthy	0.61	0.36
Exposure	Trustworthy	0.58	0.37
Exposure	Untrustworthy	0.56	0.37
Instruction	Trustworthy	0.60	0.36
Instruction	Untrustworthy	0.57	0.37

Supplementary Data Analysis

Target Race

To assess whether the intervention's reduction of facial stereotyping generalized across target race, the linear mixed-effects model was refit including a Trustworthiness $\times$ Intervention $\times$ Race (0.5 = White, -0.5 = non-White) interaction, with random intercepts for participants and stimuli and the PBS and PBS $\times$ Trustworthiness terms retained. The three-way interaction was not significant, $b = 0.00$, $SE = .01$, $t(12,590) = 0.09$, $p = .926$, indicating no evidence that race modulated the trustworthiness $\times$ intervention effect. The trustworthiness $\times$ intervention interaction remained significant and comparable to the primary model, $b = -0.03$, $SE = .01$, $t(12,590) = -3.95$, $p < .001$.

Target Gender

To assess whether the intervention's reduction of facial stereotyping generalized across target gender, the model was refit including a facial trustworthiness $\times$ intervention $\times$ gender (0.5 = female, -0.5 = male) interaction. The three-way interaction was not significant, $b = -0.01$, $SE = .01$, $t(12,590) = -0.56$, $p = .576$, indicating no evidence that gender moderated the trustworthiness $\times$ intervention effect. The trustworthiness $\times$ intervention interaction remained significant and comparable to the primary model, $b = -0.03$, $SE = .01$, $t(12,590) = -4.02$, $p < .001$.

Table S4*Linear Mixed Effects Model Fixed and Random Effects (Experiment 1).*

Predictor	Fixed Effects			
	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	0.56	0.02	36.75	< 0.001
Facial Trustworthiness	0.04	0.01	6.04	< 0.001
Exposure (vs. Control)	-0.03	0.03	-0.87	.385
Physiognomic Beliefs	-0.01	0.00	-3.16	.002
Trustworthiness × Exposure	-0.03	0.01	-4.02	< 0.001
Trustworthiness × Physiognomic Beliefs	0.00	0.00	4.65	< 0.001
Random Effects				
Subject Intercept (τ_{00})	0.30			
Stimulus Intercept (τ_{00})	0.02			
Residual (σ^2)	0.19			

Table S5*Linear Mixed Effects Model Fixed and Random Effects (Experiment 2).*

Predictor	Fixed Effects			
	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	0.54	0.01	37.27	< 0.001
Facial Trustworthiness	0.04	0.01	7.19	< 0.001
Exposure (vs. Control)	0.02	0.04	0.49	.625
Instruction (vs. Control)	0.02	0.04	0.42	.676
Physiognomic Beliefs	-0.01	0.00	-2.42	.016
Trustworthiness × Exposure	-0.02	0.01	-2.53	.012
Trustworthiness × Instruction	-0.01	0.01	-0.60	.548
Trustworthiness × Physiognomic Beliefs	0.00	0.00	4.09	< 0.001
Random Effects				
Subject Intercept (τ_{00})	0.30			
Image Intercept (τ_{00})	0.02			
Biographical Text Intercept (τ_{00})	0.04			
Residual (σ^2)	0.23			

Table S6*Linear Mixed Effects Model Fixed and Random Effects (Experiment 3).*

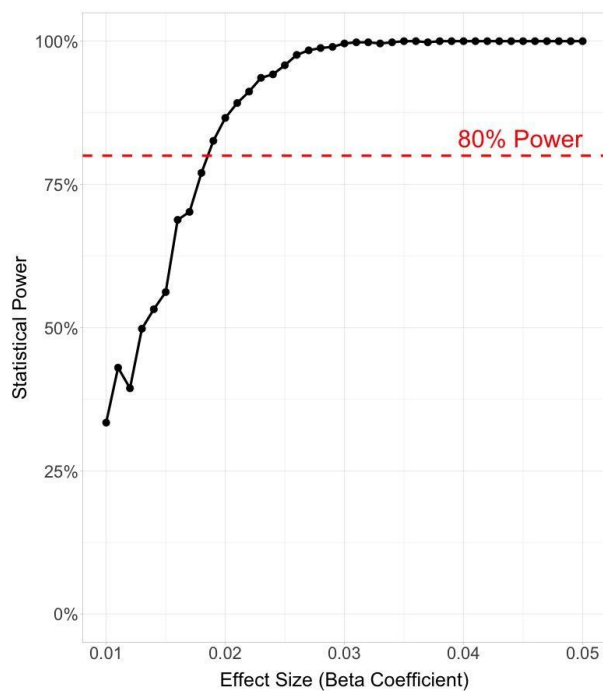
Predictor	Fixed Effects			
	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	0.59	0.01	50.15	< 0.001
Facial Trustworthiness	0.02	0.00	6.27	< 0.001
Exposure (vs. Control)	-0.04	0.03	-1.18	.239
Instruction (vs. Control)	-0.03	0.03	-0.96	.339
Physiognomic Beliefs	-0.01	0.00	-3.12	.002
Trustworthiness × Exposure	-0.01	0.01	-2.17	.030
Trustworthiness × Instruction	0.02	0.01	3.55	< 0.001
Trustworthiness × Physiognomic Beliefs	0.00	0.00	5.27	< 0.001
Random Effects				
Subject Intercept (τ_{00})	0.31			
Image Intercept (τ_{00})	0.01			
Biographical Text Intercept (τ_{00})	0.01			
Residual (σ^2)	0.19			

Sensitivity Analyses

For each experiment, we conducted a parametric sensitivity (power) analysis for the effect of interest. For Experiment 1 the effect of interest was the trait $\times$ time interaction. For Experiments 2–4, the effects of interest were interactive contrasts of trustworthiness $\times$ intervention, with the control condition compared to either the exposure or instruction condition (trust $\times$ control-vs-exposure; trust $\times$ control-vs-instruction). The linear mixed-effects model from each experiment was used as the data-generating model. Fixed-effect coefficients and variance components were taken from the fitted model, and the target interaction coefficient was set to a reasonable grid of effect sizes (Experiments 1: $b = 0.01$ – 0.04 in 0.01 increments; Experiments 2: $b = 0.01$ – 0.05 in 0.001 increments; Experiments 3 & 4: $b = 0.01$ – 0.08 in 0.005 increments). For each effect size and interaction, we generated 500 synthetic datasets by simulating responses on the original design matrix (same sample size, subject/stimulus structure, and covariates), using the estimated random-intercept variances for subjects, stimulus images, and stimulus texts, and the residual SD. Each simulated dataset was re-fit with the same mixed-effects specification, and the p -value for the target interaction was extracted from the model summary. Power at each effect size was computed as the proportion of simulations with $p < .05$. We summarized results as power curves across effect sizes and marked the .80 benchmark for reference.

Figure S1

Power Curve From Sensitivity Analysis (Experiment 1)

**Figure S2**

Power Curve From Sensitivity Analysis (Experiment 2)

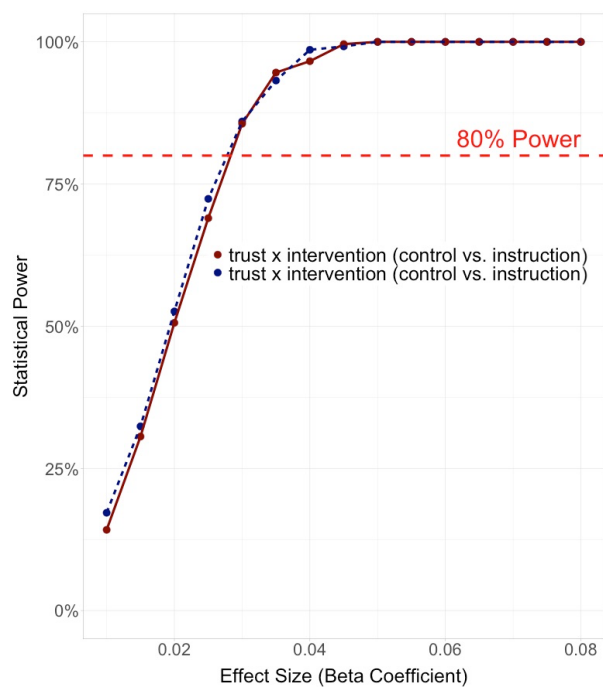
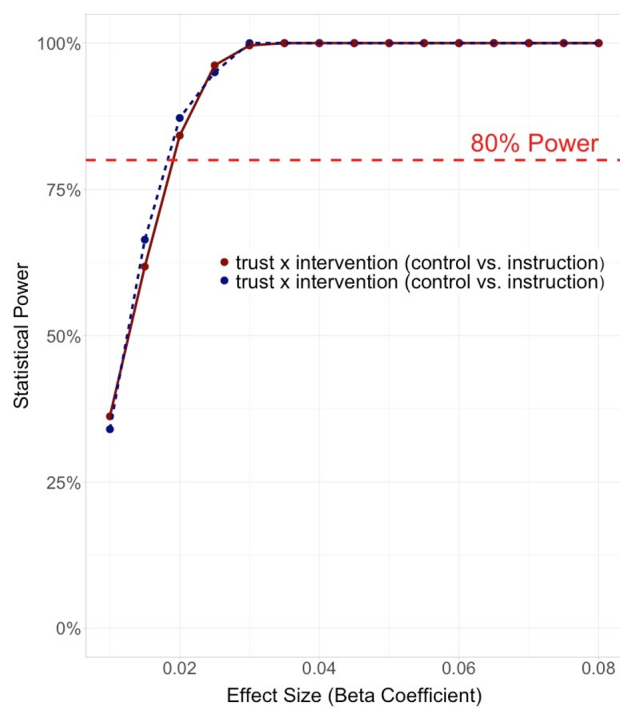


Figure S3

Power Curve From Sensitivity Analysis (Experiment 3)



Behavioral Descriptions for Instructions Condition

Stimuli

Table S7

Selected Behavioral Descriptions From Mickelberg et al (2022)

Selected Behavioral Descriptions	Presented Behavioral Descriptions
Person X solved a crossword puzzle in the newspaper.	Solved a crossword puzzle in the newspaper
Person X did really well at the quiz night.	Did really well at the quiz night
Person X put on a suit and wore their lucky socks in preparation for a job interview.	Put on a suit and wore their lucky socks in preparation for a job interview
Person X learned a secret prize-winning pie recipe from their grandmother before she died.	Learned a secret prize-winning pie recipe from their grandmother before she died
Person X locked their keys in the house and had to call a locksmith.	Locked their keys in the house and had to call a locksmith
Person X laughed at a friend's joke even though it wasn't funny.	Laughed at a friend's joke even though it wasn't funny
Person X received the employee of the month award at their work.	Received employee of the month at work
Person X failed their driver's license test for the fourth time.	Failed their driver's license test for the fourth time
Person X ate their lunch and went back to work with food stuck in their teeth.	Ate their lunch and went back to work with food stuck in their teeth
Person X forgot to turn the stove off before leaving the house.	Forgot to turn the stove off before leaving the house
Person X forgot to put the alarm on when they were the last one to leave the office.	Forgot to put the alarm on when they were the last one to leave the office
Person X sneezed loudly in an important meeting.	Sneezed loudly in an important meeting
Person X forgot it was their wedding anniversary and did not have a gift for their spouse.	Forgot it was their wedding anniversary and did not have a gift for their spouse
Person X drove their car the wrong way down a one-way street.	Drove their car the wrong way down a one-way street

Results

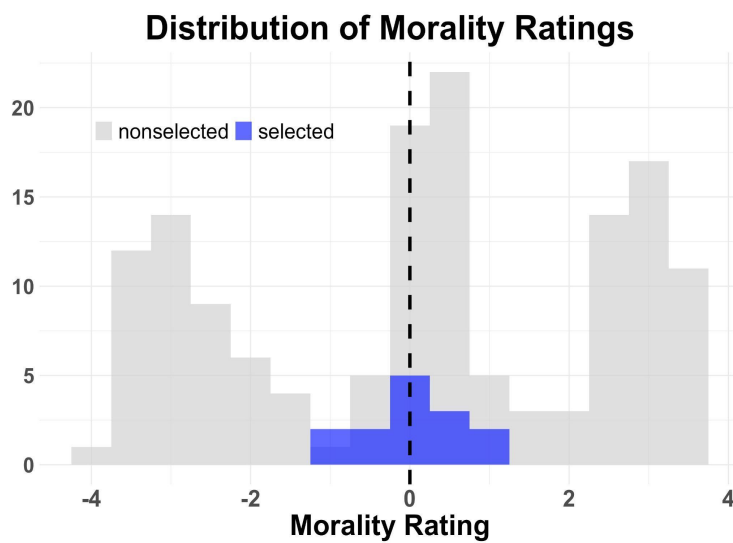
Although the norming data for these behavioral descriptions relies on morality ratings rather than trustworthiness ratings directly (Mickelberg et al., 2022), the two dimensions are often synonymous in person perception (e.g., Brambilla et al., 2011; Jaeger et al., 2022). We therefore used the magnitude of the morality ratings as a proxy for the extent to which each

behavior conveyed trustworthiness-relevant trait information, selecting relatively neutral behaviors to minimize the formation of strong behavior–trustworthiness associations. We analyzed the selected subset of behaviors for the instruction conditions to determine whether their morality ratings were, on average, closer to a neutral value than the remaining behaviors in the sample. The mean absolute morality ratings of the selected behaviors ($M = 0.46$, $SD = 0.36$) was significantly lower than the mean absolute morality ratings for all other behaviors in the stimulus set ($M = 1.94$, $SD = 0.74$). The results of the independent samples t-test confirmed this difference, $t(158) = -4.34$, $p < 0.001$, $d = -1.21$, $CI_{95\%} [-1.78, -0.65]$. This indicates a substantial and meaningful difference in the morality magnitude of the selected versus nonselected behaviors (see also, Figure S11).

The selected moral and immoral behaviors also did not differ significantly in the magnitude of their morality ratings. The absolute morality values for the moral ($M = 0.52$, $SD = 0.36$) and immoral ($M = 0.39$, $SD = 0.36$) behaviors. The analysis revealed a non-significant difference between the two groups, $t(12) = -0.67$, $p = 0.514$, $d = -0.36$, $CI_{95\%} [-1.53, 0.81]$, indicating that the two sets of behaviors were well-matched in the magnitude of their perceived morality.

Figure S4

Distribution of Morality Ratings of Selected and Nonselected Behaviors



Condition and Trust Game Instructions

Below are the instructions participants were given during the initial phase of Studies 1-3, as well as the instructions for the economic trust game that followed the initial phase. For Study 1, the instructions in the exposure condition are identical to what participants were given following the pre-exposure impression formation phase. Instructions are organized by the phase of the study and, where applicable, by the experimental condition to which the participant was assigned.

Exposure Condition

Participants in this condition were instructed to memorize faces and their associated positive or negative behaviors.

The first part of the study involves a face memorization task. In the coming pages, you will see faces of 20 unique individuals and their behavior. Each individual will be presented 3 times. Your job is to remember each face as well as their behavior. You will have as much time as you need to learn their faces as you wish. Once you have spent at least 2 seconds studying a face and memorizing their behavior, press the next button to continue on to the next face. Please pay careful attention, as these behaviors will become important in a later portion of the study.

Control Condition

The first part of the study involves a face memorization task. In the coming pages, you will see faces of 20 unique individuals and their favorite color. Each individual will be presented 3 times. Your job is to remember each face as well as their favorite color. You will have as much time as you need to learn their faces as you wish. Once you have spent at least 2 seconds studying a face and memorizing their favorite color, press the next button to continue on to the next face. Please pay careful attention, as these colors will become important in a later portion of the study.

Instruction Condition (Studies 2 & 3)

The first part of the study involves a face memorization task. In the coming pages, you will see faces of 20 unique individuals. Each individual will be presented 3

times. Your job is to remember each face while simultaneously reporting your impression of the person's behavior. You will have as much time as you need to learn their faces as you wish, and to report how trustworthy you think they are. Once you have spent at least 2 seconds studying a face and reporting your impression of how trustworthy they are, you can continue on to the next face. Please pay careful attention, as these behaviors will become important in a later portion of the study.

Following the initial block, participants received the following three-part intervention.

We are stopping this task to inform you that your responses suggest a bias in how you form impressions. They indicate that you are relying on the person's face to form impressions. In particular, it seems that you see faces on the left as less trustworthy than faces on the right, and that this perception is influencing your impressions.

[NEXT PAGE]

When our judgments of how trustworthy another person is rely on their facial appearance, we are rarely accurate. People are generally not able to predict whether someone will behave in a trustworthy manner in the future by looking at their face. Inaccurately judging someone's trustworthiness by their face can lead us to treat them unfairly. Social scientists have discovered that using how trustworthy a face looks is associated with biases in hiring, criminal sentencing, voting, and more.

[NEXT PAGE]

Because of how inaccurate and biasing the appearance of trustworthiness in the face can be, it is important to avoid using it to make judgments and decisions. For the remainder of this study, avoid letting your responses be influenced by how trustworthy a person's face looks to you. Remember, this perception is a bias that is inaccurate and should be avoided.

Trust Game

Following the learning task, all participants received instructions about the trust game.

Now we'll move on to the second part of today's study, where you will be playing a money game with other participants. Please read the instructions as carefully as possible, because this game is probably different than others you have played in the past.

This is also where you may see some attention check questions, so please pay attention, and try your best to answer each question thoughtfully.

This particular game is played in pairs. Each pair will have one person designated as Player 1 and one person designated as Player 2.

*You have been randomly assigned to be **Player 1** in all the rounds of today's game. So that the other players have some idea of who they're playing against, we'll have you select an avatar to represent you during the game, as well as answer a few questions about yourself. We realize that these avatars might not capture your appearance perfectly. Please simply choose one of the faces that you feel best represents you.*

Finally, all participants received instructions explaining the rules and structure of the trust game.

Great! Now that you've selected your avatar and answered those questions, we'll explain the rules of the game.

[NEXT PAGE]

On each turn, Player 1 will receive \$1. Player 1 then has the opportunity to give a portion of their money to Player 2. For example, they could give \$1.00, \$0.75, \$0.25, or \$0.00. Whatever amount Player 1 decides to give Player 2 will then be tripled by the researcher. Player 2 has the opportunity to return any portion of that amount to Player 1. Then the turn is over.

Player 1 goes home with whatever he or she kept from the original \$1 plus anything returned to them by Player 2. Player 2 goes home with whatever amount they are left with after splitting the money tripled by the researcher.

Note that the larger the amount that Player 1 gives Player 2, the greater the amount the two players can take home. However, it is entirely up to Player 2 to decide how much to give back to Player 1. As a result, Player 1 could end up with more than their original dollar amount or less than their original dollar amount.

[NEXT PAGE]

Now, we will show you some examples to help you understand the game.

Example #1. *Imagine that Player 1 gives \$1.00 to Player 2. The researcher triples this amount, so Player 2 now has \$3.00 ($\$1 \times 3 = \3.00). Then Player 2 has to decide whether they wish to give anything back to Player 1, and if so, how much. Suppose Player 2 decides to return \$2.00 to Player 1. At the end of the turn, Player 1 will go home with \$2.00 and Player 2 will go home with \$1.00.*

Example #2. *Now let's try another example. Imagine that Player 1 gives \$0.75 to Player 2. The researcher triples this amount, so Player 2 gets \$2.25 ($\$0.75 \times 3 = \2.25). Then Player 2 has to decide whether they wish to give anything back to Player 1, and if so, how much. Suppose Player 2 decides to return \$0 to Player 1. At the end of the turn, Player 1 will go home with \$0.25 and Player 2 will go home with \$2.25.*

Example #3. *Now let's try one last example. Imagine that Player 1 gives \$0.25 to Player 2. The researcher triples this amount, so Player 2 gets \$0.75 ($\$0.25 \times 3 =$*

\$0.75). Then Player 2 has to decide whether they wish to give anything back to Player 1, and if so, how much. Suppose Player 2 decides to return \$0.75 to Player 1. At the end of the turn, Player 1 will go home with \$1.50 and Player 2 will go home with \$0.

[NEXT PAGE]

Each round of the game will pair you with a new participant. You will have as long as necessary to make your decision in each round, but you will not learn how much money you earned until the end of the study. This is intentional so that your earnings don't affect your performance in the remaining parts of the study. There may be a slight delay while we find another participant to match you with for each round, but it shouldn't take too long.

References

- 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.
- 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.
- 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.