

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Cognitive Underpinnings to Social Judgments Through Information and Information Processing

Permalink

<https://escholarship.org/uc/item/60x8m3sb>

Author

Klein, Samuel

Publication Date

2024

Peer reviewed|Thesis/dissertation

Cognitive Underpinnings to Social Judgments Through Information and Information Processing

By

SAMUEL A. W. KLEIN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dr. Jeffrey Sherman, Chair

Dr. Andrew Todd

Dr. Jimmy Calanchini

Committee in Charge

2024

Acknowledgements

Portions of this research were supported by National Science Foundation through Grant BCS-1764097 awarded to Dr. Andrew Todd and Grant BCS-2215236 awarded to Dr. Jeffrey Sherman. Portions of this dissertation work were also supported by access to the high-performance computing resources by the High-Performance Computing Facility at the University of California, Davis.

Dr. Jeffrey Sherman has been my PhD advisor and academic mentor for nearly six years. What I have learned from Jeff over the years will stay with me for the rest of my life, including how to think critically about what drives the outcomes we observe and to never stop asking “why?” Jeff has been immensely supportive as a mentor and a friend. In six years, he has never failed to be there when I needed him. But it is the way he pushes me that I am most grateful for in our time together. He never ceases to challenge my ideas, while giving me a stage to defend, reshape, and sharpen my thinking. He encouraged me to seek out and welcome varied and critical feedback early on, and continuously, to make sure my thinking is sharpened as much as it can be with the evidence and viewpoints that are available.

Although not on paper, Dr. Andrew Todd has certainly been another key advisor and mentor to me throughout my time at UC Davis. He offered me a chance to pursue a research proposal I wrote for his Social Cognition class during my first quarter of graduate school, which turned into my first first-author empirical publication. Since then, we have continued to collaborate, with me learning so much from him on every subsequent project. Andy rarely misses an opportunity to think out loud when we work together, offering me many chances to observe and learn his approach to research. Watching his sharp attention to detail, incisive thinking, and clear pragmatism has taught me quite a lot about how to be an effective scientist.

I owe a special thanks to my academic big brother, Ryan Hutchings, whose support during my first couple years of graduate school was integral to my development. I spent a lot (and I mean a lot) of his time with questions about coding, machine learning, face perception, and broader topics about our field. What quickly burgeoned was a friendship, one that I am so grateful to have (it has been “reg” and not “diff”).

I have been so fortunate to work with and learn from many great minds (and even greater hearts) beyond the UC Davis campus. I owe much to Drs. Jimmy Calanchini, Mahzarin Banaji, Tessa Charlesworth, Daniel Heck, and Christoph Klauer. Regarding the latter two, few people have more knowledge and skill in computational cognitive modeling than Drs. Daniel Heck and Christoph Klauer. Fewer still are there two people more generous with their time to guide others in learning these tools. Their support over the years has been invaluable to me. I would not be surprised if the sum of our email conversations took up more pages than this dissertation.

To Aline Da Silva Frost, Chris Coleman, Stephanie Oliinyk, Tommi Mayers, Siuoneh Didarloo, Evan Warfel, Eva Meza, those in the Students of Social Cognition group, and the many more friends I cannot list without adding too many more pages, thank you for the camaraderie and the countless hours of “nerding out” over our shared love of psychological research.

I would also like to thank Northeastern Illinois University. The book stipend offered in my scholarship funded my surveying of a variety of books that caught my eye, including *Heuristics and Biases: The Psychology of Intuitive Judgment*. Without reading that book, my passion for social cognitive psychology may never have been sparked. Yet without Dr. Amanda Dykema-Engblade, that passion would not have taken me very far. Her patience, support, and the opportunities she offered me are some of the most memorable gifts I have received.

Table of Contents

Acknowledgements	ii
List of Tables	v
List of Figures	vi
Abstract	vii
Chapter 1: Revising Mental Representations of Faces Based on New Diagnostic Information ...	1
Abstract	2
Image Generation Experiment	4
Method	4
Results	7
Image Assessment Experiments	10
Experiment 1	10
Method	10
Results	10
Experiments 2A and 2B	12
Method	12
Results	13
Discussion	16
References	18
Chapter 2: Emotion Expression Salience and Racially Biased Weapon Identification: A Diffusion Modeling Approach	22
Abstract	23
Experiment 1	27
Method	27
Results	31
Discussion	35
Experiment 2	35
Method	35
Results	36
Discussion	38
General Discussion	38
References	43
Chapter 3: Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research	47
Abstract	48
Theoretical Background	49
A Multi-Cue Measurement Problem	51
A Multi-Cue Integration Model	53
Demonstrating the MCI	55
Further Applications of the MCI Model	62
Conclusion	66
References	67
Appendix A: Supplemental Materials for Chapter 1	74
Appendix B: Supplemental Materials for Chapter 2	99
Appendix C: Supplemental Materials for Chapter 3	123

List of Tables

Table 1. Chapter 1: Participant Demographics in Each Experiment 5
Table 2. Chapter 2: Parameters of the Diffusion Decision Model in the Weapon Identification
Task 25

List of Figures

Figure 1. Chapter 1: Group Images by Time, Time 1 Induction, and Time 2 Information	7
Figure 2. Chapter 1: Trait Impressions of Robert	9
Figure 3. Chapter 1: Trait Impressions of Group Classification Images	11
Figure 4. Chapter 1: Trustworthiness Impressions of Subgroup Classification Images	14
Figure 5. Chapter 1: Trustworthiness Impressions of Individual Classification Images	15
Figure 6. Chapter 2: Illustration of the Diffusion Decision Process	27
Figure 7. Chapter 2: Behavioral Data Plots by Race Prime and Salience Condition Across Experiments	32
Figure 8. Chapter 2: Relative Start Point (β) Estimates by Race Prime and Salience Conditions Across Experiments	34
Figure 9. Chapter 3: The MCI Model and Its Predicted Responses to Gender Classifications ...	56
Figure 10. Chapter 3: Estimated Use of Sex Cues During Face Classification	60
Figure 11. Chapter 3: Estimated Use of Sex Cues During Face Classification	61

Abstract

Thinking about other people often requires complex cognitive processing to integrate their many and varied features. The three papers I present further our understanding of how we integrate this information into coherent social judgments. In Chapter 1, I tested whether mental representations of others' appearance rapidly update to new information. After learning to ascribe valenced behaviors to a target person and subsequently visualizing his face in a reverse-correlation task, participants learned new information that was (a) counter-attitudinal and diagnostic about his character or (b) neutral and non-diagnostic before generating a second visualization.

Visualizations at Time 2 assimilated to counter-attitudinal information, suggesting that representations of others' appearance may rapidly update to new information. In Chapter 2, I examined whether and how the salience of emotion expression (scowling, smiling) or race (Black, White) cues shapes racially biased weapon identification (gun, tool). Across two manipulations of salience, racially biased weapon identification was weaker when the salience of emotion versus race was heightened. Using diffusion modeling, I tested competing cognitive accounts of this effect. Consistent support emerged for an initial bias account, whereby the decision process initiates closer to "gun" responses upon seeing Black (vs. White) faces, and this racially biased shift in the starting position is weaker when emotion (vs. race) is salient. In Chapter 3, I developed a solution to the inability of conventional measures of social judgment to distinguish the unique contributions of multiple features (e.g., social categories, behaviors). In particular, I introduced a computational model to separately measure the use of multiple features underlying social judgments. Using data from a judgments task in which emotion and sex cues varied in target faces, I initially validate the model's capacity to measure the use of those cues and demonstrated how the model be applied to answer long-standing questions in the field.

Chapter 1

Revising Mental Representations of Faces Based on New Diagnostic Information

Abstract

Extending evidence for the rapid revision of mental representations of what other people are like, we explored whether people also rapidly revise their representations of what others *look* like. After learning to ascribe positive or negative behavioral information to a target person and generating a visualization of their face in a reverse-correlation task, participants learned new information that was (a) counter-attitudinal and diagnostic about the person's character or (b) neutral and non-diagnostic, and then they generated a second visualization. Ratings of these visualizations in separate samples of participants consistently revealed revision effects: Time 2 visualizations assimilated to the counter-attitudinal information. Weaker revision effects also emerged after learning neutral information, suggesting that the evaluative extremity of visualizations may dilute when encountering any additional information. These findings indicate that representations of others' appearance may change upon learning more about them, particularly when this new information is counter-attitudinal and diagnostic.

Keywords: face impressions; mental representations; reverse correlation; social cognition; updating

Revising Mental Representations of Faces Based on New Diagnostic Information

First impressions are lasting impressions (Asch, 1946)—or so it has been assumed, particularly for implicit (i.e., unintentional) impressions. Indeed, some dual-process theorizing has maintained that, although people readily revise their explicit impressions of others when encountering countervailing target information, implicit impression revision occurs more slowly, if at all (e.g., Rydell & McConnell, 2006). This claim was initially supported by research that failed to find implicit impression revision based on countervailing target information (e.g., Gregg et al., 2006). Although the malleability of implicit impressions has long been recognized (Gawronski & Bodenhausen, 2006), an assumption guiding this literature has been that exposure to abundant countervailing information is required for implicit impression revision.

Counter to this assumption, accumulating evidence now indicates that people can rapidly revise their implicit impressions when encountering even a single piece of diagnostic information that contradicts their initial impression (Ferguson et al., 2019). For example, participants who first learned positive behavioral information about a person fully reversed their initially favorable impression after learning about his child molestation conviction (Cone & Ferguson, 2015). Such updating generalizes beyond the context in which the impression was formed (Brannon & Gawronski, 2017) and is evident days later (Cone et al., 2021b), suggesting genuine revision.

Notably, all documented instances of rapid impression revision have emerged in mental representations of the target person's *character*, commonly assessed with sequential-priming tasks (e.g., affect misattribution procedure [AMP]; Payne et al., 2005) and self-report measures. Although tracking impression revision with such measures provides insight into evaluative assumptions about the person's character, it is silent on how the person's physical *appearance* is initially represented or potentially revised.

Here, we explored the revision of facial appearance representations using reverse correlation (Mangini & Biederman, 2004), a data-driven approach for visualizing the features underlying face classifications. This technique imposes no pre-existing assumptions about these features, thereby affording an unconstrained assessment of what another person *looks* like. The measurement outcomes of reverse-correlation tasks, unlike those of the AMP and other indirect measures of what a person *is* like, include conditionally variable features of physical appearance (Brinkman et al., 2017). Exactly how these features relate to measures that serve as proxies of character representations remains an open question (Dotsch et al., 2013).

Our investigation comprised an image-generation experiment and three image-assessment experiments. In the image-generation experiment, participants visualized a target person twice: first after learning to ascribe positive or negative behaviors to him, and again after receiving new information that was (a) diagnostic, extreme, and contradictory to the initial information or (b) neutral and non-diagnostic. This procedure produced classification images for each of 8 experimental conditions. In three image-assessment experiments, with three distinct image-processing procedures, separate samples of participants rated these images on traits that are detectable in faces. Data for all experiments are available here: <https://osf.io/7u6cd/>

Image-Generation Experiment

Method

Participants

Prior research on implicit impression revision has revealed large effects ($\eta_p^2s > .12$; Cone & Ferguson, 2015); however, whether appearance representations shift comparably remains unknown. Rounding up to the nearest number divisible by 50, we thus set our target sample size

near our smallest effect of interest ($\eta_p^2 = .05$; 90% power) in our $2 \times 2 \times 2$ mixed design.^{1,2} After surpassing our target sample (250 participants), we continued data collection until the week’s end. In total, 338 undergraduates participated for course credit. We excluded data from 53 participants who pressed the same key for $\geq 95\%$ of image-generation trials at Time 1 or Time 2. The final sample comprised 285 participants (see Table 1 for participant demographics in all experiments). Across experiments, participants provided informed consent prior to participating.

Table 1

Chapter 1: Participant Demographics in Each Experiment.

Experiment	Age		Gender (%)			Race/Ethnicity (%)				
	<i>M</i>	<i>SD</i>	Male	Female	Nonbinary	W	B	A	L	M
IG	19.9	2.2	20.7	76.8	1.4	18.9	2.5	46.0	20.4	12.3
IA 1	35.3	10.6	63.2	36.1	0.0	38.7	36.8	6.5	7.7	10.3
IA 2A	38.5	12.6	48.2	50.0	0.0	76.3	6.1	6.1	1.8	9.6
IA 2B	20.0	2.7	18.2	79.3	0.0	15.7	0.4	50.8	20.2	12.8

Note. IG = image generation, IA = image assessment. Some participants did not report their gender or race/ethnicity. For race/ethnicity, W = White or European American, B = Black or African American, A = Asian American or Pacific Islander, L = Latinx or Hispanic, and M = reported other or more than one race/ethnicity.

Procedure

Participants first learned to ascribe positive or negative behaviors to a target person, Robert (Cone & Ferguson, 2015). They read (in randomized order) 64 behaviors (32 positive, 32

¹To our knowledge, no formal power analysis procedures exist for the image-generation phase in reverse-correlation paradigms (see also Brown-Iannuzzi et al., 2021).

²Our planned analyses were conducted at the level of participants; however, based on editorial feedback, we shifted to analyses that account for other sources of variance (i.e., traits or stimuli, depending on the experiment). Thus, the reported *a priori* power analyses, conducted with G*Power (Faul et al., 2007), only considered the number of participants, but not the number of traits or stimuli.

negative; Rydell & McConnell, 2006) and indicated whether each was characteristic or uncharacteristic of him, after which condition-specific feedback appeared for 2.5 s. In the *positive-induction* condition, a blue *correct* message appeared after classifying a positive (negative) behavior as characteristic (uncharacteristic), and a red *incorrect* message appeared after classifying a negative (positive) behavior as characteristic (uncharacteristic).

Accompanying each message was a summary statement (e.g., “Giving flowers to his mother is characteristic of Robert”). In the *negative-induction* condition, these feedback contingencies were reversed. Participants then reported their impressions of Robert on 7 traits that are important for person impressions (Oosterhof & Todorov, 2008): trustworthy, attractive, dominant, caring, intelligent, aggressive, and mean (1 = *not at all*, 7 = *extremely*).

Next, participants completed a reverse-correlation task (Brinkman et al., 2017). On each of 350 trials, they selected which of two side-by-side degraded face images looked more like Robert. Each pair of images comprised a random noise pattern³ and its inverse superimposed onto a base face image.⁴ This technique maximizes between-image contrast (Dotsch & Todorov, 2012). Responses <100 ms or >4000 ms after target onset prompted a message to respond more slowly or more quickly, respectively.

Participants then received one new piece of information about Robert. In the *counter-attitudinal* condition, the information was diagnostic about his character and contradicted the valence of their Time 1 induction (“Robert was recently convicted of child molestation” after a positive induction; “Robert donated one of his kidneys to a child in need he had never met

³The noise patterns comprised 4,092 superimposed truncated sinusoid patches in all possible combinations of 2 cycles in 6 orientations (0°, 30°, 60°, 90°, 120°, 150°) × 5 spatial frequencies (1, 2, 4, 8, 16 patches per image) × 2 phases (0, $\pi/2$), with random contrasts.









⁴The base face, which Krosch and Amodio (2014) created by morphing 100 White and 100 Black male faces, has been used in prior reverse-correlation research (e.g., Lei & Bodenhausen, 2017).

before” after a negative induction). These behaviors were rated similarly in diagnosticity and valence extremity (see Cone & Ferguson, 2015). In the *neutral* condition, the information was neutral in valence (“Robert recently bought a soda”). Finally, participants again reported their impressions of Robert and completed the reverse-correlation task in a newly-randomized order.

Using the *rcicr* package (Dotsch, 2014), we created group classification images by superimposing onto the base face the average noise patterns of the selected images across all participants in each condition. Group images reflect the average features visualized of Robert within that condition (see Figure. 1).⁵

Figure 1

Chapter 1: Group Images by Time, Time 1 Induction, and Time 2 Information

	Positive Induction		Negative Induction	
	Time 1	Time 2	Time 1	Time 2
Counter-attitudinal				
Neutral				

Results

All analyses were conducted via linear mixed-effects models (LMEMs), with each model containing fixed effects for Time, Time 1 induction, Time 2 information, and all possible interactions. For each model, we began with its maximal random-effects structure (i.e., random intercepts and all appropriate random slopes for each source of variance; Barr et al., 2013) and

⁵We report in Appendix A additional measures collected in all experiments.

downsized to solve problems of non-convergence and singularity. The sources of variance were participants and traits in the image-generation experiment and image-assessment Experiment 1, and participants and stimuli in image-assessment Experiments 2A and 2B.⁶

We reverse-scored responses for aggressive, dominant, and mean, ensuring that all traits were directionally consistent in valence, and considered the 7 traits as having been sampled from the population of positive traits on which impressions could be formed. Because the trait ratings were highly correlated (see Appendix A) and fitting separate models for each trait can inflate Type-I error (Herzog et al., 2019), we included random effects for traits.⁷

This analysis revealed a significant three-way interaction, $b = -0.37$, $SE = 0.03$, $F(1, 280.97) = 206.33$, $p < .001$ (see Figure 2). To explicate this interaction, we examined contrasts of the model's Time \times Time 2 information interactions separately in the positive-induction and negative-induction conditions. This interaction was significant in both the positive-induction condition, $b = 2.27$, $SE = .15$, $t(281) = 15.68$, $p < .001$, and the negative-induction condition, $b = -0.68$, $SE = .15$, $t(281) = -4.66$, $p < .001$, with significantly greater positive-to-negative than negative-to-positive revision, $b = -1.59$, $SE = .21$, $t(281) = -7.78$, $p < .001$.⁸

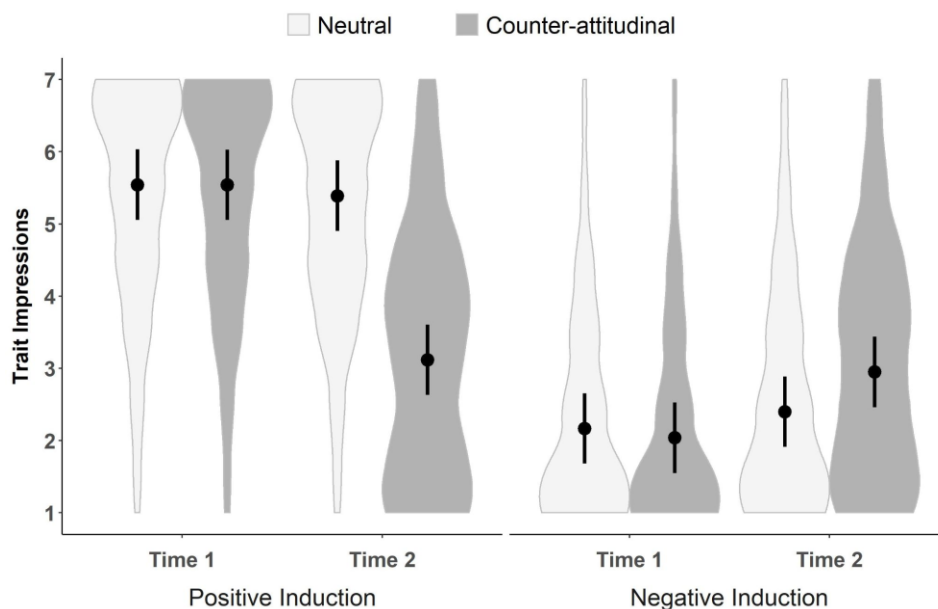
⁶See Appendix A for a detailed description of the random-effects structures for each mixed-effects model reported in the main text, and a discussion of how problems of non-convergence and singularity in the maximal models led to those reported in the main text.

⁷An alternative analytic approach entails using data-reduction techniques (e.g., exploratory factor analysis) to fit these models to latent factor(s). Because our focus was on testing for revision effects in representations of appearance, regardless of the trait, we do not report this approach in the main text (but see Appendix A for results using this alternative approach).

⁸This additional post-hoc test assessed the magnitude of revision, as reflected in the difference between the Time \times Time 2 information contrast in the positive-induction condition and this same contrast in the negative-induction condition. Due to the opposing numerical directions of the contrasts, we first multiplied all ratings in the negative-induction condition by a constant of -1, ensuring that the difference between the two contrasts reflects the *magnitude* of difference. We used this same approach in all three image-assessment experiments.

Figure 2

Chapter 1: Trait Impressions of Robert



Notes. Markers reflect estimated marginal means of trait impressions of Robert by Time, Time 1 induction, and Time 2 information in the image-generation experiment. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image generators' responses after a smoothing function was applied.

Next, we conducted pairwise comparisons in each induction condition. In the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 2.42$, $SE = 0.13$, $t(30.00) = 19.40$, $p < .001$, but learning neutral information did not, $b = 0.15$, $SE = 0.13$, $t(31.60) = 1.21$, $p = .235$. In negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.91$, $SE = 0.13$, $t(32.20) = -7.17$, $p < .001$, but learning neutral information did not, $b = -0.24$, $SE = 0.13$, $t(30.00) = -1.88$, $p = .070$.⁹

These results replicate prior findings indicating that learning diagnostic counter-attitudinal target information prompts character-representation revision (Cone & Ferguson, 2015). As in this

⁹For detailed information on the descriptive statistics for these and all other experiments, see Appendix A.

prior work, we also observed an asymmetry whereby positive-to-negative revision was stronger than negative-to-positive revision.

Image-Assessment Experiments

To examine whether learning countervailing diagnostic information prompts appearance-representation revision, we conducted several image-assessment experiments. In Experiment 1, a new sample of participants rated the 8 group images (see Figure 1) on the same traits from before. Experiments 2A and 2B used two alternative image-processing procedures (detailed below) and two new samples of raters to assess apparent trustworthiness.

Experiment 1

Method

Participants. We again considered the large revision effects (η_p^2 s > .12) in Cone and Ferguson (2015) but allowed for weaker effects. To detect a medium-sized three-way interaction ($\eta_p^2 = .06$) with 80% power in a $2 \times 2 \times 2$ within-participants design, we set our target sample size at 126. Amazon’s Mechanical Turk (MTurk) workers ($N = 155$) participated for pay. No data were excluded; thus, the final sample comprised 155 participants.

Procedure. Participants rated the 8 group images on the same 7 traits from the image-generation experiment.

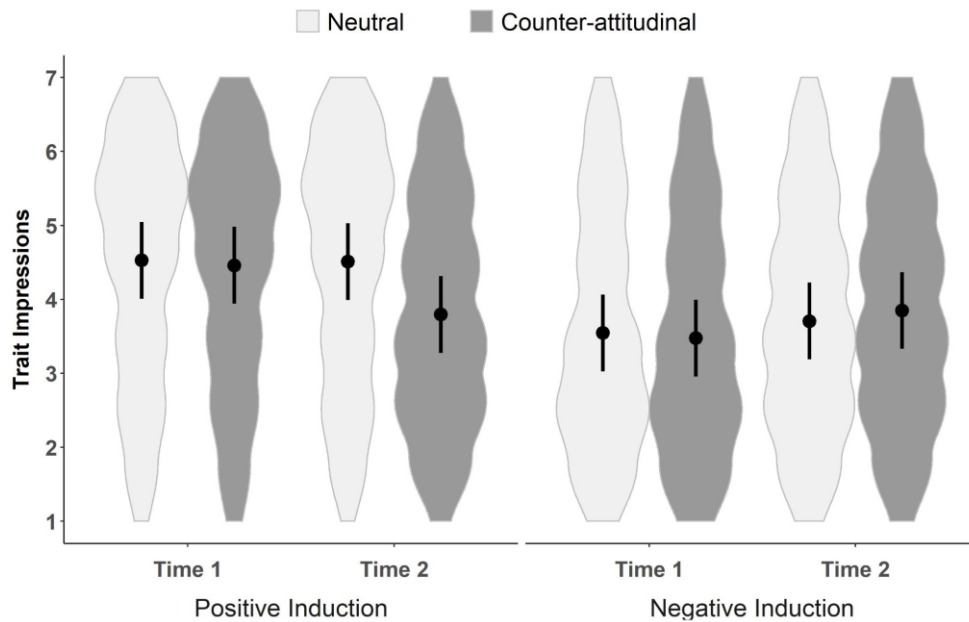
Results

A LMEM revealed a significant three-way interaction, $b = -0.11$, $SE = 0.01$, $F(1, 8358) = 52.32$, $p < .001$ (see Figure 3). We again examined contrasts of the model’s Time \times Time 2 information interactions separately in the two induction conditions. This interaction was significant in both the positive-induction condition, $b = 0.65$, $SE = .08$, $t(8358) = 7.69$, $p < .001$, and the negative-induction condition, $b = -0.21$, $SE = .08$, $t(8358) = -2.54$, $p = .021$, with

significantly greater positive-to-negative than negative-to-positive revision, $b = -0.43$, $SE = .13$, $t(8358) = -3.31$, $p = .001$.

Figure 3

Chapter 1: Trait Impressions of Group Classification Images



Notes. Estimated marginal means of trustworthiness impressions of group classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 1. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.66$, $SE = 0.06$, $t(8358) = 11.16$, $p < .001$, but learning neutral information did not, $b = 0.02$, $SE = 0.06$, $t(8358) = 2.66$, $p = .790$. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.37$, $SE = 0.06$, $t(8358) = -6.63$, $p < .001$, but so did learning neutral information, $b = -0.16$, $SE = 0.06$, $t(8358) = -2.69$, $p = .007$, albeit to a lesser extent.¹⁰

¹⁰A difference between counter-attitudinal and neutral information in the negative-induction condition is evidenced by the significant Time \times Time 2 contrast in the negative-induction condition reported above.

Visualizations of Robert’s appearance grew less favorable upon learning negative counter-attitudinal information and more favorable upon learning positive counter-attitudinal information. Notably, positive-to-negative revision was stronger than negative-to-positive, replicating findings from the image-generation experiment and elsewhere (Cone & Ferguson, 2015). When learning neutral information, participants did not visualize Robert’s appearance differently if their initial visualization was positive; however, they did visualize him more favorably if their initial visualization was negative. Because the neutral information was non-diagnostic, revision here may reflect a dilution effect (Nisbett et al., 1981), whereby highly negative initial visualizations become less extreme upon learning any additional information.

Experiments 2A and 2B

Image-assessment Experiment 1 relied exclusively on group images that were created by aggregating across the responses of all image generators per condition. This practice, though normative in reverse-correlation research (Brinkman et al., 2017), can artificially augment between-condition differences, thereby inflating Type-I error (Cone et al., 2021a). Image-assessment Experiments 2A and 2B used alternative image-processing procedures—subgroup and individual classification images—that avoid this limitation (Cone et al., 2021a; Hutchings et al., 2021). Both experiments assessed apparent trustworthiness, given its centrality in face impressions (Oosterhof & Todorov, 2008).

Method

Participants. We considered the possibility that subgroup and (perhaps especially) individual images are noisier than group images, potentially producing smaller effects. To detect medium-sized three-way interactions (Experiment 2A: $\eta_p^2 = .06$; Experiment 2B: $\eta_p^2 = .03$) with 80% power, we set target sample sizes of 126 (Experiment 2A) and 257 (Experiment 2B). In total, 125

MTurkers (Experiment 2A) and 259 undergraduates (Experiment 2B) participated for pay and course credit, respectively. We excluded data from participants who gave the same response on $\geq 95\%$ of ratings (Experiment 2A: $n = 6$; Experiment 2B: $n = 5$) or who did not finish the entire experiment (Experiment 2A: $n = 5$; Experiment 2B: $n = 12$). The final samples comprised 114 participants in Experiment 2A and 242 participants in Experiment 2B.

Procedure. We used the *rcicr* package (Dotsch, 2014) to create subgroup and individual images. In Experiment 2A, we created subgroup images by aggregating the noise patterns selected by 12 random subsets of image generators in each condition and superimposing them onto the base face (Cone et al., 2021a). The total stimulus set included 96 subgroup images, with each image comprising the average selected noise patterns from 5–7 image generators.

Participants rated all 96 subgroup images (order randomized). In Experiment 2B, we created individual images by aggregating the noise patterns selected by each image generator, separately for each time point, and superimposing them onto the base face. The total stimulus set comprised 570 images. To minimize fatigue, we had participants rate one of three sets of 95 randomized pairs of Time 1 and Time 2 images, totaling 190 images. In both experiments, participants rated how trustworthy the person looked (1 = *extremely untrustworthy*, 7 = *extremely trustworthy*).

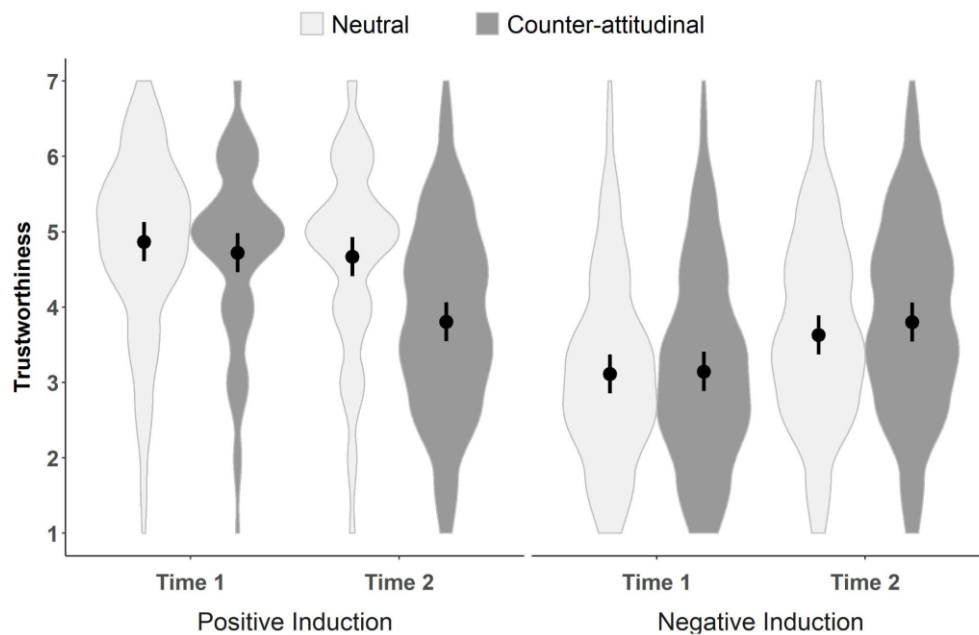
Results

Experiment 2A (Subgroup Images). A LMEM revealed a significant three-way interaction, $b = -0.11$, $SE = 0.03$, $F(1, 45.27) = 10.67$, $p = .002$ (see Figure 4). Next, we examined contrasts of the model's Time \times Time 2 information interactions separately in the two induction conditions. This interaction was significant in the positive-induction condition, $b = 0.72$, $SE = .19$, $t(44.8) = 3.89$, $p < .001$, but not in the negative-induction condition, $b = -0.14$, $SE = .19$,

$t(44.8) = -0.75, p = .456$, with significantly greater positive-to-negative than negative-to-positive revision, $b = -0.58, SE = .26, t(44.3) = -2.23, p = .031$.

Figure 4

Chapter 1: Trustworthiness Impressions of Subgroup Classification Images



Notes. Estimated marginal means of trustworthiness impressions of subgroup classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2A. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

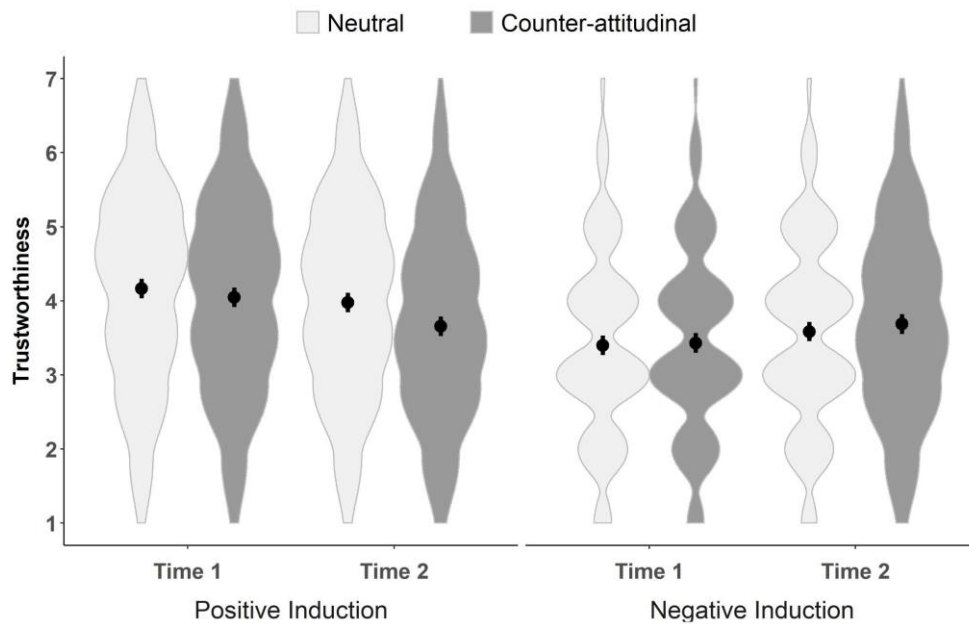
Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.92, SE = 0.13, t(48.4) = 6.87, p < .001$, but learning neutral information did not, $b = 0.20, SE = 0.13, t(48.40) = 1.48, p = .145$. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.65, SE = .13, t(48.40) = -4.90, p < .001$, but so did learning neutral information, $b = -0.52, SE = .13, t(48.40) = -3.86, p < .001$.

Experiment 2B (Individual Images). Once again, the three-way interaction was significant, $b = -0.03, SE = 0.01, F(1, 280.36) = 5.90, p = .016$ (see Figure 5). As before, we examined

contrasts of the model's Time \times Time 2 information interactions separately in the two induction conditions. This interaction was significant in the positive-induction condition, $b = 0.20$, $SE = .08$, $t(281) = 2.55$, $p < .001$, but not in the negative-induction condition, $b = -0.07$, $SE = .08$, $t(282) = -0.89$, $p = .372$, with no significant difference in the magnitude of positive-to-negative versus negative-to-positive revision, $b = -0.13$, $SE = .11$, $t(279) = -1.17$, $p = .244$.

Figure 5

Chapter 1: Trustworthiness Impressions of Individual Classification Images



Notes. Estimated marginal means of trustworthiness impressions of individual classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2B. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.39$, $SE = 0.06$, $t(303) = 6.88$, $p < .001$. Learning neutral information also prompted negative revision, $b = 0.19$, $SE = 0.06$, $t(303) = 3.24$, $p = .001$, albeit to a lesser extent. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -$

0.26, $SE = .06$, $t(293) = -4.43$, $p < .001$, but so did learning neutral information, $b = -0.19$, $SE = .13$, $t(294) = -3.29$, $p < .001$.

Discussion

Using a reverse-correlation paradigm, we found that visualizations of a target person's face consistently assimilated to new information that was extreme, diagnostic, and contradictory to the initial information learned about him. Initially positive visualizations were revised to appear less trustworthy after learning about his child molestation conviction. The opposite pattern emerged when participants with negative initial visualizations learned about his kidney donation, albeit sometimes to no greater extent than the revision prompted by learning neutral information.¹¹ These results complement other evidence of rapid revision in mental representations of what others are like under similar conditions (Ferguson et al., 2019). We also found a valence asymmetry, whereby greater positive-to-negative (vs. negative-to-positive) revision emerged in most cases (cf. Cone & Ferguson, 2015).

Some evidence of revision also emerged, albeit more weakly, when learning new neutral information about the person. Revision here might reflect a dilution effect, whereby new non-diagnostic information diluted the extremity of the initial visualizations (Nisbett et al., 1981). If so, one implication of this finding is that extreme appearance representations may dissipate over time upon learning any additional target information.

A strength of this work is its use of group, subgroup, and individual classification images, with the latter two procedures reducing concerns about Type-I error inflation (Cone et al., 2021a). Because subgroup aggregation is a new technique, future work should explore optimal points at which subgroup images minimize noise but preserve image-generator variability.

¹¹Less conservative analyses that did not account for random effects of stimuli consistently revealed revision effects in both induction conditions in all experiments. We report these analyses in Appendix A.

Furthermore, although some results (e.g., revision after neutral information) varied across experiments, the key effect (i.e., stronger revision after counter-attitudinal vs. neutral information) emerged consistently. Such convergence helps bolster our conclusions. Future work should identify boundary conditions of these effects. For example, if the information initially learned about a person (e.g., broke into his neighbor's house) is reinterpreted based on new information (e.g., the house was on fire and children were inside), do we revise our representations of their appearance accordingly (Mann & Ferguson, 2015)?

Notably, we found a sizable correlation between image generators' trustworthiness ratings of Robert in the image-generation experiment and image raters' trustworthiness ratings of individual images of Robert in image-assessment Experiment 2B, $r(568) = .51, p < .001$, suggesting a correspondence between revision in (explicit) character representations and revision in appearance representations, at least when the new information learned about the person is diagnostic, extreme, and (presumably) believable (see Ferguson et al., 2019). What remains for future research is determining whether similar correspondence emerges if one of these elements is missing or under conditions in which corresponding revisions in implicit and explicit character representations have not materialized in prior work (e.g., Gregg et al., 2006). Future research should also explore whether revisions in character representations precede (and/or cause) changes in appearance representations. Answering these questions promises a richer understanding of how various components of person impressions are integrated.

The current findings indicate that mental representations of others' appearance are far from static. As we learn new information about someone, not only do we revise our representations of what they are like; we also revise our representations of what they *look* like.

References

- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290. <https://doi.org/10.1037/h0055756>
- 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*, 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context-(in)dependent updating of implicit evaluations. *Social Psychological and Personality Science, 8*, 275–283. <https://doi.org/10.1177/1948550616673875>
- Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*, 333–361. <https://doi.org/10.1080/10463283.2017.1381469>
- Brown-Iannuzzi, J. L., Cooley, E., Marshburn, C. K., McKee, S. E., & Lei, R. F. (2021). Investigating the interplay between race, work ethic stereotypes, and attitudes toward welfare recipients and policies. *Social Psychological and Personality Science*. <https://doi.org/10.1177/1948550620983051>
- Cone, J., Brown-Iannuzzi, J., Lei, R., & Dotsch, R. (2021a). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science, 12*, 760–768. <https://doi.org/10.1177/1948550620938616>
- Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*, 37–57. <https://doi.org/10.1037/pspa0000014>

- Cone, J., Flaharty, K., & Ferguson, M. J. (2021b). The long-term effects of new evidence on implicit impressions of other people. *Psychological Science*, *32*, 173–188.
<https://doi.org/10.1177/0956797620963559>
- Dotsch, R. (2014). *rcicr: Reverse correlation image classification toolbox*. R package. Retrieved from <http://ron.dotsch.org/rcicr>
- Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*, 562–571.
<https://doi.org/10.1177/1948550611430272>
- Dotsch, R., Wigboldus, D. H. J., & Van Knippenberg, A. D. (2013). Behavioral information biases the expected facial appearance of members of novel groups. *European Journal of Social Psychology*, *43*, 116–125. <https://doi.org/10.1002/ejsp.1928>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. <https://doi.org/10.3758/bf03193146>
- Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science*, *28*, 331–336.
<https://doi.org/10.1177/0963721419835206>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731. <https://doi.org/10.1037/0033-2909.132.5.692>
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the

malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*, 1–20. <https://doi.org/10.1037/0022-3514.90.1.1>

Herzog, M. H., Francis, G., & Clarke, A. (2019). The Multiple Testing Problem. In *Understanding Statistics and Experimental Design* (pp. 63–66). Springer, Cham.

Hutchings, R. J., Simpson, A. J., Sherman, J. W., & Todd, A. R. (2021). Perspective taking reduces intergroup bias in visual representations of faces. *Cognition*, *214*, 104808. <https://doi.org/10.1016/j.cognition.2021.104808>

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences*, *111*, 9079–9084. <https://doi.org/10.1073/pnas.1404448111>

Lei, R. F., & Bodenhausen, G. V. (2017). Racial assumptions color the mental representation of social class. *Frontiers in Psychology*, *8*, 519. <https://doi.org/10.3389/fpsyg.2017.00519>

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209–226. <https://doi.org/10.1016/j.cogsci.2003.11.004>

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, *108*, 823–849. <https://doi.org/10.1037/pspa0000021>

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*, 248–277. [https://doi.org/10.1016/0010-0285\(81\)90010-4](https://doi.org/10.1016/0010-0285(81)90010-4)

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

<https://doi.org/10.1073/pnas.0805664105>

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes:

Affect misattribution as implicit measurement. *Journal of Personality and Social*

Psychology, *89*, 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*, 995–1008.

<https://doi.org/10.1037/0022-3514.91.6.995>

Chapter 2

Emotion Expression Salience and Racially Biased Weapon Identification: A Diffusion Modeling Approach

Abstract

Racial stereotypes are commonly activated by informational cues that are detectable in people's faces. Here, we used a sequential priming task to examine whether and how the salience of emotion (angry/scowling vs. happy/smiling expressions) or apparent race (Black vs. White) information in male face primes shapes racially biased weapon identification (gun vs. tool) decisions. In two experiments ($N_{\text{total}} = 546$) using two different manipulations of facial information salience, racial bias in weapon identification was weaker when the salience of emotion expression versus race was heightened. Using diffusion modeling, we tested competing accounts of the cognitive mechanism by which the salience of facial information moderates this behavioral effect. Consistent support emerged for an initial bias account, whereby the decision process began closer to the "gun" response upon seeing faces of Black versus White men, and this racially biased shift in the starting position was weaker when emotion versus race information was salient. We discuss these results vis-à-vis prior empirical and theoretical work on how facial information salience moderates racial bias in decision-making.

Emotion Expression Salience and Racially Biased Weapon Identification: A Diffusion Modeling Approach

Racial stereotypes pervade modern thinking, with abundant experimental evidence more strongly linking Black versus White people with weapons (Payne & Correll, 2020). In the weapon identification task (WIT), for example, participants are usually better (i.e., faster and more accurate) at identifying guns and worse at identifying harmless objects (e.g., tools, toys) after seeing Black versus White face primes (Amodio et al., 2004; Payne, 2001; Todd et al., 2016). This typical pattern of racial bias in the WIT is robust (see Rivers, 2017); however, its magnitude may vary by the salience of (i.e., the attention garnered by; Higgins, 1996) information in the face primes. Indeed, racially biased weapon identification is weaker and sometimes eliminated when age versus race information is more salient (Jones & Fazio, 2010; Todd et al., 2021). Granted, age is only one of many sources of social information. In two experiments, we investigated whether attending to another facial cue—emotion expression—likewise weakens weapon-related racial bias, relative to attending to race.

Unlike facial cues pertaining to relatively static social categories (e.g., age, race), emotion expressions can vary moment-to-moment within the same target person. Furthermore, emotion expressions presumably signal affect and intentions (Niedenthal & Brauer, 2012; Todorov et al., 2008) in ways that demographic cues may not, making them informative for basic social judgment (e.g., identifying threats). Accordingly, emotion expressions may garner substantial attention in threat-related contexts like weapon identification, effectively competing against the attention often garnered by race in such contexts (Payne & Correll, 2020). Indeed, the mere availability of scowls and smiles has been found to affect racially biased weapon identification (Kubota & Ito, 2014), whereas the mere availability of other information (e.g., age cues) has not (Todd et al., 2016). Thus, it seems reasonable to propose that directing attention toward emotion

versus race cues moderates racially biased weapon identification. Our experiments tested this proposition.

Besides investigating *whether* the salience of emotion versus race information alters racial bias in the WIT, we examine *how* such an effect emerges using diffusion decision modeling (DDM; Ratcliff et al., 2016). The DDM is a sequential sampling technique designed to disentangle processes underlying behavior in tasks like the WIT. By concurrently modeling both decisions and decision speed, the DDM decomposes decisions into four parameters (see Table 2). We briefly describe two relevant parameters that might explain how information salience moderates racially biased weapon identification.

Table 2

Chapter 2: Parameters of the Diffusion Decision Model in the Weapon Identification Task

Parameter	Interpretation
Relative start point (β)	Initial bias to select <i>gun</i> or <i>tool</i> at the start of evidence accumulation, with $0 < \beta < 1$. Values $>.50$ indicate a bias to select <i>gun</i> ; values $<.50$ indicate a bias to select <i>tool</i> .
Threshold separation (α)	Amount of evidence required to decide, with $0 < \alpha$. Hitting a threshold triggers a decision to select <i>gun</i> or <i>tool</i> .
Drift rate (δ)	Average quality of information extracted at each unit of time, with $-\infty < \delta < \infty$. Higher absolute values indicate stronger evidence. Positive values indicate evidence to select <i>gun</i> ; negative values indicate evidence to select <i>tool</i> .
Non-decision time (τ)	Length of all response components (encoding time, motor response time, and other unknown contaminants) unrelated to decision making, with $0 < \tau$. Measured in milliseconds.

The DDM assumes that evidence is accumulated over time until a decision threshold is reached. It models both the strength of evidence extracted (i.e., drift rate) and the position from

which evidence accumulation begins (i.e., relative start point; see Figure 6). An *evidence accumulation* account of facial information salience moderating racially biased weapon identification posits that seeing a Black versus White face prime strengthens the evidence accumulated for identifying guns (i.e., race-stereotypic objects), but that racially biased evidence accumulation is weaker when emotion versus race information is salient. Alternatively, an *initial bias* account posits that seeing a Black versus White face prime shifts the starting position of the decision process closer to the “gun” response, but that shifts in the start point are less racially biased when emotion versus race is salient.¹²

The initial bias account is more strongly supported in the WIT literature.¹³ Specifically, whereas an evidence accumulation account did not explain racially biased weapon identification, or its moderation by the salience of age versus race information in the face primes, an initial bias account did (Todd et al., 2021). Relative to age cues, the arguably greater relevance of emotion expression in threat-related contexts might undermine racially biased weapon identification by altering the interpretation of object-related content—the process-level pattern predicted by an evidence accumulation account. Thus, testing these accounts when emotion versus race salience varies in the context of weapon identification is instructive. Our experiments provide such a test.

For consistency with prior work (e.g., Todd et al., 2021), we report behavioral analyses of the error rates and correct response times (RTs) along with our analyses of the DDM parameters.

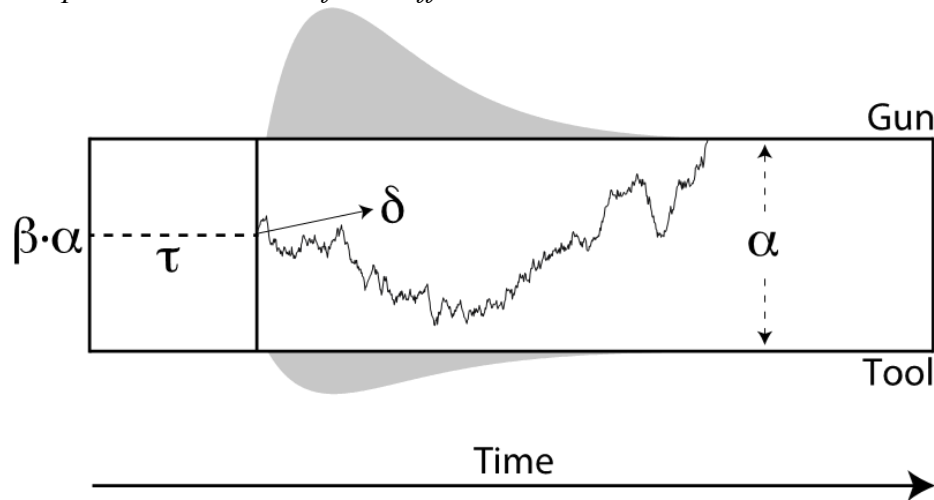
Data and code are available at <https://osf.io/hxywn/>.

¹² We did not derive clear predictions about threshold separation and non-decision time, but we report results pertaining to these parameters for completeness.

¹³ Notably, an evidence accumulation account better explains racial bias in the first-person shooter task (FPST, Correll et al., 2015; Johnson et al., 2018; Pleskac et al., 2018). For a discussion of procedural differences between the FPST and the WIT, see Todd et al. (2021).

Figure 6

Chapter 2: Illustration of the Diffusion Decision Process



Notes. The decision process starts with a bias to select gun or tool, as indicated by the relative start point, β . Evidence is then accumulated (as illustrated by the jagged line) for each decision option, with average strength δ . The distance between the thresholds, α , indicates the amount of evidence needed to decide. Finally, the length of non-decision processes is indicated by τ . The hypothetical distributions (in gray) above and below the decision space indicate that the model predicts the distribution of response times for each decision option.

Experiment 1

Method

Participants

Prior work using a similar design (Todd et al., 2021, Experiment 2) revealed a small-to-medium sized salience effect on racial bias in the WIT (Salience \times Race Prime \times Target Object interaction: $\eta_p^2 = .028$). Thus, we set a target sample size ($N = 280$) affording $\geq 80\%$ power to detect $\eta_p^2 = .028$ (Faul et al., 2007). In total, 311 undergraduates consented to participate for course credit. We decided a priori to exclude data from participants who performed at or below chance (errors on $\geq 50\%$ of trials) on any trial type in the WIT ($n = 21$). Retaining the excluded data did not meaningfully alter any of the conclusions in either experiment. The final sample comprised 290 participants (81% women, 15.4% men, 1.8% non-binary; 15% White, 2.1% Black, 57.3% Asian, 17.1% Latino/a/e/x, 5.2% multiracial; $M_{\text{age}} = 19.3$, $SD = 1.3$).

Procedure

In both experiments, participants arrived at the lab in small groups and were led by an experimenter to an individual computer workstation to complete the experimental tasks. Participants completed a sequential priming task, the WIT (Payne, 2001), wherein two images appeared in quick succession. Instructions urged participants to ignore the first image (face prime) and to classify the second image (target object) quickly and accurately via key press. The face primes were facial images of 48 men varying in apparent race (24 Black, 24 White) and posed emotion expression (24 angry/scowling, 24 happy/smiling) from the Chicago Face Database (Ma et al., 2015).¹⁴ The target objects were 6 gun and 6 tool images from Payne (2001). Each trial comprised the following sequence: fixation cross (500 ms), face prime (200 ms), target object (200 ms), and pattern mask (until participants responded). If participants failed to respond within 500 ms, a message (“Please respond faster!”) appeared (1 s).

We structured the WIT so that apparent race or emotion expression was more distinctive throughout the task (Macrae & Cloutier, 2009; Rees et al., 2022; Todd et al., 2021). Participants were randomly assigned to complete one of two WIT variants, each comprising two blocks of 144 experimental trials (288 total trials that were preceded by 12 practice trials). In the *race-salient* condition, the face primes were scowling Black and White men in one block of trials and smiling Black and White men in the other block. In the *expression-salient* condition, the face primes were smiling and scowling Black men in one block of trials and smiling and scowling White men in the other block. Within a given block of trials, varying only one source of information (e.g., emotion expression) should render it more contextually distinctive, and thus

¹⁴ The emotion expression and apparent race of these face stimuli were likely construed unambiguously. In the face categorization task in Experiment 2, emotion expression and race were both “correctly” classified on $\geq 95\%$ of trials, supporting the assumption that both sources of information were clear and easy to identify (see Tables B7 & B8).

more salient (Taylor & Fiske, 1978), than the other source of information (e.g., apparent race). Block order was counterbalanced and did not moderate racial bias.

Analysis Plan

Prior to all analyses, we excluded trials with RTs <100 ms and >1500 ms (Todd et al., 2021), which eliminated 2.4% of the data in both experiments. We also excluded error trials prior to RT analyses (but see Appendix B for RT analyses of error trials). Below, we report analyses pertinent to our focal hypotheses on information salience effects on racial bias.

Behavioral Data Analyses. All analyses were conducted using linear mixed-effects models (LMEMs), with each model containing fixed effects for Salience, Race Prime, Emotion Prime, Target Object, and all identifiable interactions. Models included a random-effects structure with by-participant and by-stimulus random intercepts.^{15,16} This approach is analogous to fitting the data to a mixed analysis of variance that is adjusted for the cross-classified clustering of responses within participants and within stimuli. We examined our effect of interest (i.e., the Salience \times Race Prime \times Target Object effect) via contrasts of the model's Race Prime \times Target Object interactions across and between salience conditions. Full LMEM tables appear in Appendix B (see Tables B1 and B3).

DDM Parameter Estimation. For each experiment, we estimated the model using a Markov Chain Monte Carlo (MCMC) sampler in JAGS 4.30 (Plummer, 2003) with the Wiener distribution provided by Wabersich and Vandekerckhove (2014) and an estimation approach to make inferences in this framework (Gelman et al., 2003; Kruschke, 2014). Mirroring model

¹⁵ The only exception was for the LMEM on incorrect response times reported in Appendix B. Due to boundary fit conditions, we removed the by-stimulus random intercept from the model.

¹⁶ Although the LMEM on error rates in Experiment 2 afforded the inclusion of by-participant random slopes for Race Prime, we chose to prioritize consistency within and across experiments over that single model's random effects structure. Inclusion versus exclusion of the additional random effect did not meaningfully change the results.

specifications from Todd et al. (2021; see also Pleskac et al., 2018), all parameters were allowed to vary by Race Prime, Emotion Prime, and Salience, and the drift rate and non-decision time parameters also were allowed to vary by Target Object.¹⁷ As in prior work (Todd et al., 2021), posterior predictive checks suggest that the model adequately characterizes the WIT data. The representativeness and accuracy of each model's estimation were assessed both visually and numerically (see Appendix B and the online supplementary materials) and were found to be adequate enough to rely on the parameter estimates for subsequent process analyses.

To compare parameter estimates across conditions, we computed contrasts that included the 95% highest density interval (HDI_{95%}) of the difference between posterior distributions of each parameter across the relevant conditions. Differences with HDI_{95%} excluding 0 are considered credible. For each analysis, we report the most credible estimate of the raw difference, a Cohen's *d*, and the HDI_{95%} around *d*. The effect of Race Prime was compared across levels of Salience for all four parameters. For drift rate and non-decision time, contrasts were further computed to evaluate the effect of Race Prime across levels of Target Object. (Figure 8 displays the relative start point parameter estimates in both experiments; Figures B7–B14 display all other parameter estimates, including the start point parameter estimates varying by emotion expression as well).

Results

Behavioral Analyses

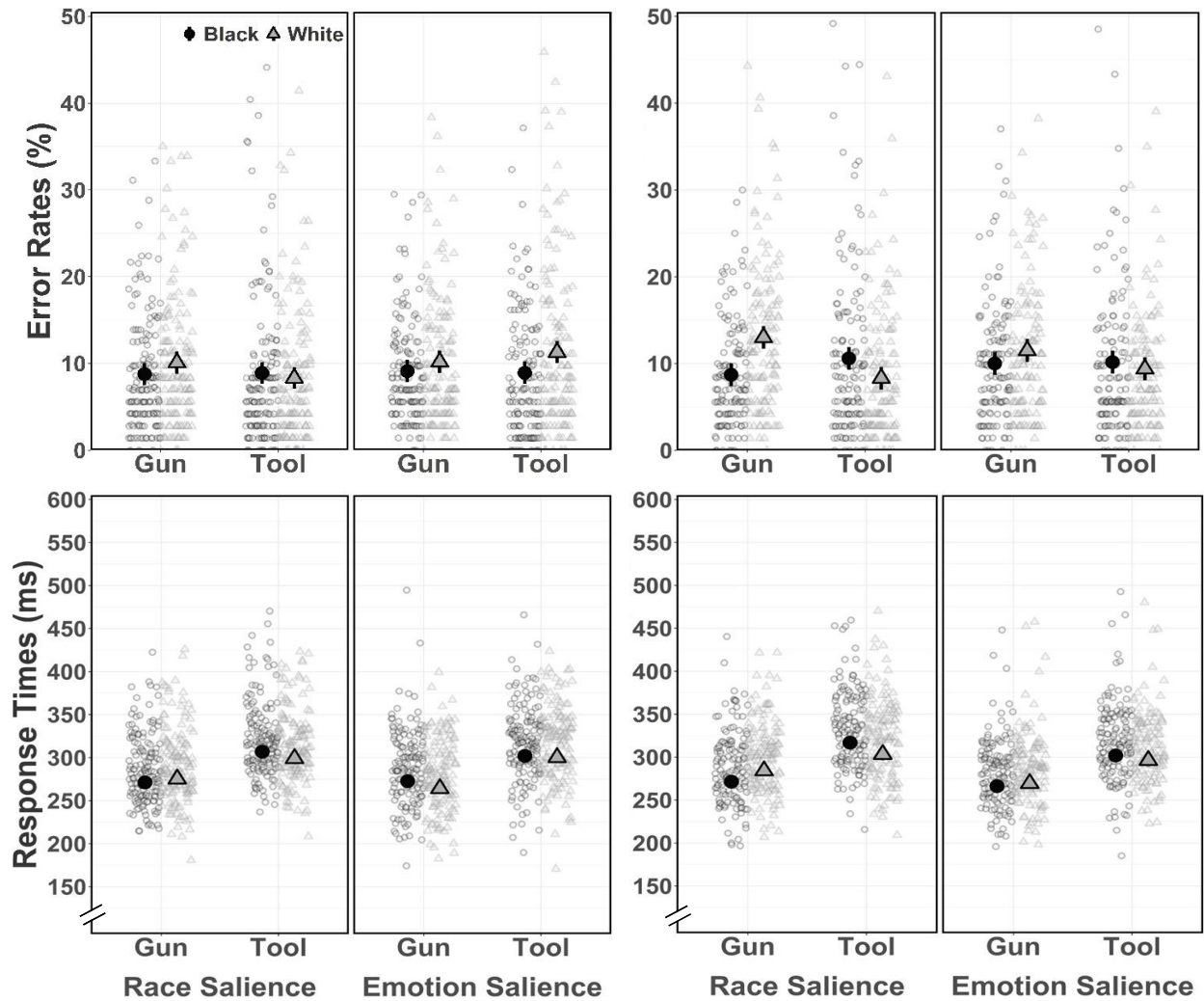
Error Rates. A significant Salience × Race Prime × Target Object interaction, $\beta = 0.03$, $F(1, 82166.6) = 16.92$, $p < .001$, $R^2 < .01$, revealed salience-driven variation in racially biased weapon

¹⁷ As highlighted in Table 2, the threshold separation and relative start point parameters cannot be identified across conditions of Target Object: The relative start point parameter reflects the position at which participants are closer to a gun versus tool decision *at target onset*; the threshold separation parameter reflects the extent to which evidence must be accumulated to reach a gun versus tool decision, presumably determined before target onset. Presumably both the extent of evidence accumulated from the target object (i.e., drift rate) and the processing time prior to a response being recorded (i.e., non-decision time) may vary by Target Object.

identification. When race was salient, the Race Prime \times Target Object interaction (i.e., racial bias) was significant, $b = -0.02$, $z = -2.60$, $p = .009$, though neither underlying simple effect of Race Prime reached significance (see Table B5 for simple effects). When emotion was salient, however, the Race Prime \times Target Object interaction was not significant, $b = 0.01$, $z = 1.79$, $p = .073$ (see Table B6 for descriptive statistics for each experiment).

Figure 7

Chapter 2: Behavioral Data Plots by Race Prime and Salience Condition Across Experiments



Notes. Markers reflect error rates (top row) and correct response times (bottom row) for Black and White prime trials. Empty markers reflect individual-level data and filled shapes and their error bars reflect the estimated marginal means from the linear mixed-effects model applied to those data. The x-axis displays whether the target object was a gun or tool. Shading and shape of markers reflect whether the target object followed a Black or White face prime. Panels vary by salience condition, whereby panels on the left within each plot reflect the race-salient condition and panels on the right within each plot reflect the emotion-salient condition. The plots on the left display data from Experiment 1; the plots on the right display data from Experiment 2.

Correct RTs. A significant Salience \times Race Prime \times Target Object interaction, $\beta = 0.07$, $F(1, 74530.0) = 63.83$, $p < .001$, $R^2 = .03$, again revealed salience-driven variation in racial bias.

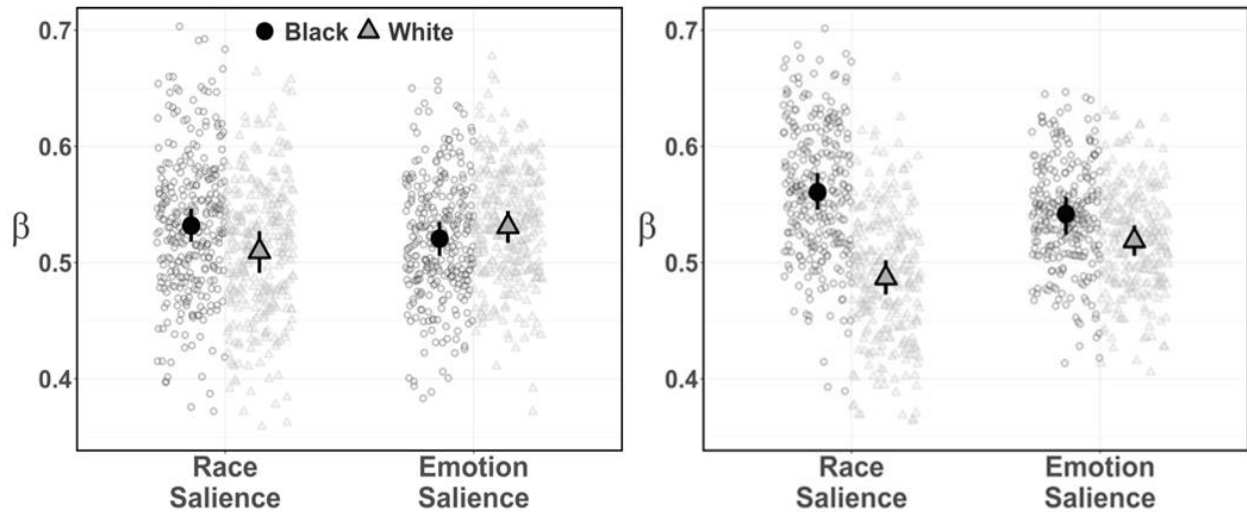
When race was salient, racial bias was evident, $b = -0.04$, $z = -5.11$, $p < .001$. Guns were identified faster ($M_{\text{diff}} = -4$ ms) and tools were identified slower ($M_{\text{diff}} = 6$ ms) after Black versus White primes. Contrary to expectations, when emotion was salient, there was significant racial bias in the opposite direction, $b = 0.03$, $z = 3.16$, $p = .002$. Guns were identified *slower* after Black versus White primes ($M_{\text{diff}} = 7$ ms); the speed of tool identification, by contrast, did not significantly differ between race primes.

Process Analyses

A Salience \times Race Prime contrast on the relative start point (β) was credible, $\mu_{\text{diff}} = 0.02$, $d = 0.25$, HDI_{95%} [0.11, 0.41]. When race was salient, the decision process began closer to “gun” after Black versus White primes, $\mu_{\text{diff}} = -0.02$, $d = -0.33$, HDI_{95%} [-0.54, -0.13]. When emotion was salient, no credible racial bias emerged, $\mu_{\text{diff}} = 0.01$, $d = 0.18$, HDI_{95%} [-0.03, 0.38]. These findings align with an initial bias account: Salience-driven variation in racially biased starting positions in the decision process explain salience-driven moderation of racially biased behavior.

Figure 8

Chapter 2: Relative Start Point (β) Estimates by Race Prime and Salience Conditions Across Experiments



Notes. Markers reflect posterior estimates for Black prime and White prime trials. Empty markers reflect individual-level estimates and filled shapes and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to the data. The x-axis displays information salience condition (race, emotion). Shading and shape of markers reflect the salience condition. The plot on the left displays estimates from Experiment 1; the plot on the right displays estimates from Experiment 2.

A small but credible race prime effect emerged on the drift rate (δ), $\mu_{\text{diff}} = -0.14$, $d = -0.15$, $\text{HDI}_{95\%} [-0.25, -0.06]$, but it did not vary by information salience, $\mu_{\text{diff}} = -0.05$, $d = -0.06$, $\text{HDI}_{95\%} [-0.16, 0.03]$, or target object, $\mu_{\text{diff}} = 0.04$, $d = 0.04$, $\text{HDI}_{95\%} [-0.06, 0.14]$. Accumulated evidence from target objects was stronger after Black versus White primes, regardless of whether emotion or race information was more salient or whether the object was a gun or tool.

The race prime effect on threshold separation (α) was not credible, $\mu_{\text{diff}} = -0.02$, $d = -0.11$, $\text{HDI}_{95\%} [-0.25, 0.02]$. Finally, a small but credible race prime effect emerged on non-decision time (τ), $\mu_{\text{diff}} = -0.004$, $d = -0.10$, $\text{HDI}_{95\%} [-0.20, -0.03]$, but it did not vary by information salience, $\mu_{\text{diff}} = -0.001$, $d = -0.02$, $\text{HDI}_{95\%} [-0.12, 0.05]$, or target object, $\mu_{\text{diff}} < 0.001$, $d = 0.01$, $\text{HDI}_{95\%} [-0.08, 0.09]$.

Discussion

In Experiment 1, racial bias was weaker when emotion versus race was salient. Process analyses failed to support an evidence accumulation account of this effect. Neither target object nor information salience moderated the stronger evidence accumulation occurring after Black versus White primes. Rather, process analyses supported an initial bias account: When race was salient, the decision process began closer to “gun” after Black versus White primes. When emotion was salient, no credible start-point bias emerged. Descriptively, however, start points in the emotion-salient condition were *farther* from “gun” after Black versus White primes, mirroring the atypical pattern of RTs in the emotion-salient condition (e.g., *slower* tool identifications after Black versus White primes). Whether behavior assimilates toward (e.g., typical racial bias) or contrasts from (e.g., atypical racial bias) race stereotypes can vary by context (Bless & Schwarz, 2010), raising questions about whether the atypical pattern in Experiment 1 stems from our blocking design. Experiment 2, therefore, aimed to replicate these results using a different manipulation of information salience.

Experiment 2

Method

Participants

Prior work using a similar design (Todd et al., 2021, Experiment 1) revealed a large effect of information salience on racial bias in the WIT (Salience \times Race Prime \times Target Object interaction: $\eta_p^2 = .139$); however, because smaller effects are of theoretical interest, we set a target sample size ($N = 258$) affording $\geq 80\%$ power to detect $\eta_p^2 = .03$ (Faul et al., 2007). In total, 278 undergraduates consented to participate for course credit. We decided a priori to exclude data from participants who performed at or below chance (errors on $\geq 50\%$ of trials) on

the face categorization task ($n = 1$) or on any trial type in the WIT ($n = 20$). We also excluded data from one participant for whom a computer error caused the WIT to abort early. The final sample comprised 256 participants (73.4% women, 24.2% men, 1.2% non-binary; 12.7% White, 1.9% Black, 61.3% Asian, 15.2% Latino/a/e/x, 4.7% multiracial; $M_{\text{age}} = 19.4$, $SD = 2.0$).

Procedure

Participants first completed a face categorization task (Todd et al., 2021) wherein they viewed one of two stimulus sets of facial images, each containing a randomly selected batch of 24 of the 48 facial images from Experiment 1. Both stimulus sets contained equal numbers of male faces varying in apparent race and posed emotion expression. Depending on information salience condition, participants were randomly assigned to classify the faces by *race* (Black vs. White) or by *emotion expression* (angry vs. happy) via key press. The images appeared one-by-one and remained on screen until participants responded, for a total of 72 trials.

Next, participants completed a WIT that deviated from the WIT in Experiment 1 in two ways. First, the face primes were the other set of 24 facial images not used during the face categorization task. We counterbalanced which stimulus set was used for the face categorization task and the WIT. Using different facial stimuli in the two tasks allowed us to rule out an event coding account (Hommel et al., 2001) whereby memory of specific responses toward specific faces in the face categorization task might affect responses toward those same faces in the WIT. Second, the face prime \times target object combinations were fully integrated within a single block of 288 experimental trials that were preceded by 12 practice trials.

Results

Behavioral Analyses

Error Rates. A significant Salience \times Race Prime \times Target Object interaction, $\beta = 0.04$, $F(1, 74212.8) = 25.35$, $p < .001$, $R^2 < .01$, revealed salience-driven variation in racial bias. Race bias was evident when race was salient, $b = -0.02$, $z = -2.60$, $p = .009$. Guns were misidentified as tools less often after Black versus White primes ($M_{\text{diff}} = -4.3\%$) and tools were misidentified as guns more often after Black versus White primes ($M_{\text{diff}} = 2.2\%$). Racial bias was also evident (albeit more weakly) when emotion was salient, $b = -0.02$, $z = -3.04$, $p = .002$. Guns were misidentified as tools more often after Black versus White primes ($M_{\text{diff}} = -1.5\%$), whereas misidentification of tools did not significantly differ between race primes.

Correct RTs. A significant Salience \times Race Prime \times Target Object interaction, $\beta = 0.59$, $F(1, 66767.8) = 42.52$, $p < .001$, $R^2 = .04$, again revealed salience-driven variation in racial bias. Racial bias emerged when race was salient, $b = -0.07$, $z = -11.67$, $p < .001$. Guns were identified faster after Black versus White primes ($M_{\text{diff}} = -9$ ms) and tools were identified slower after Black versus White primes ($M_{\text{diff}} = 8$ ms). Racial bias also emerged (albeit more weakly) when emotion was salient, $b = -0.03$, $z = -3.98$, $p < .001$. Whereas the speed of gun identification did not significantly differ between race primes, tools were identified slower after Black versus White primes ($M_{\text{diff}} = 5$ ms).

Process Analyses

A Salience \times Race Prime contrast on the relative start point (β) was credible, $\mu_{\text{diff}} = 0.02$, $d = 0.40$, HDI_{95%} [0.25, 0.57]. When race was salient, the decision process began closer to “gun” after Black versus White primes, $\mu_{\text{diff}} = -0.07$, $d = -1.17$, HDI_{95%} [-1.45, -0.94]. Although start-point bias also emerged when emotion was salient, the effect was weaker, $\mu_{\text{diff}} = -0.02$, $d = -0.39$, HDI_{95%} [-0.60, -0.16]. Like Experiment 1, these findings align with an initial bias account.

A small but credible race prime effect emerged on the drift rate (δ), $\mu_{\text{diff}} = -0.13$, $d = -0.17$, HDI_{95%} [-0.27, -0.06], but it did not vary by information salience, $\mu_{\text{diff}} = 0.07$, $d = 0.10$, HDI_{95%} [-0.01, 0.20], or target object, $\mu_{\text{diff}} = 0.07$, $d = 0.10$, HDI_{95%} [-0.02, 0.20]. Stronger evidence was accumulated for the target objects after Black versus White primes, regardless of whether emotion or race information was more salient or whether the object was a gun or tool.

A small but credible race prime effect also emerged on threshold separation (α), $\mu_{\text{diff}} = -0.04$, $d = -0.24$, HDI_{95%} [-0.38, -0.09], but it did not vary by salience, $\mu_{\text{diff}} = 0.01$, $d = 0.07$, HDI_{95%} [-0.07, 0.22]. The amount of evidence required before responding was greater after Black versus White primes, regardless of information salience. No credible effects emerged on non-decision time (τ).

Discussion

In Experiment 2, facial information salience again moderated racial bias in behavior, and these results again were better explained by an initial bias account. The decision process began closer to “gun” following Black versus White primes, but less so when emotion versus race information was more salient. Once again, stronger evidence accumulation following Black versus White primes did not vary by target object or which information was more salient.

General Discussion

In two experiments, we examined if and how the salience of facial information shapes racially biased weapon identification. We manipulated salience either by augmenting the distinctiveness of emotion or race information during the WIT (Experiment 1) or by augmenting participants’ experience in processing emotion or race information prior to the WIT (Experiment 2). Racial bias in behavior was consistently weaker when the salience of emotion versus race information was highlighted. These findings complement a growing body of evidence suggesting

that the salience of facial information other than race (e.g., the age of the face primes) can alter racially biased weapon identification (Jones & Fazio, 2010; Todd et al., 2021; see also Gawronski et al., 2010). Specifically, they suggest that attending to comparatively more dynamic and affect-laden information communicated by facial expressions of emotion can likewise moderate racially biased weapon identification.

Using diffusion modeling, we tested competing cognitive accounts of *how* facial information salience shapes racially biased weapon identification. Our results contradict the evidence accumulation account, which posits that evidence is accumulated from stereotype-congruent (vs. stereotype-irrelevant) target objects more strongly following race primes, and that the salience of information in the face primes shapes this phenomenon. In both experiments, the strength of evidence accumulation did not vary stereotypically (e.g., larger estimates for guns following Black primes) nor by information salience.

Our process-level analyses instead consistently supported an initial bias account, which posits that the weapon identification process begins closer to a race-stereotypic decision after encountering race information in the face primes, and that the salience of facial information shapes the strength of this start-point bias. In both experiments, the decision process began closer to “gun” responses shortly after participants encountered Black versus White face primes. Furthermore, the strength of this effect was shaped by the salience of facial information: Racially biased start points were either eliminated (Experiment 1) or attenuated (Experiment 2) when emotion versus race was salient. Considered alongside previous findings of moderation by age salience (Todd et al., 2021), these results support the initial bias account as a mechanism whereby attending to person information besides race lowers the likelihood of favoring the “gun” response before the object’s appearance, relative to attending to race-related information.

Notably, our experiments failed to replicate prior findings that the mere availability of emotion cues in face primes moderates racially biased weapon identification (see Tables B1 & B3). Whereas Kubota and Ito (2014) found that racial bias emerged for scowling but not smiling face primes (see also Raissi & Steele, 2021), here emotion expressions in the face primes failed to moderate racial bias (despite these emotion expressions being easily detected; see footnote 3 and Tables B7 & B8). Furthermore, emotion expressions weakly moderated weapon identification (i.e., the Emotion Prime \times Target Object interaction) in Experiment 2, but this effect was not moderated by the salience of emotion. This latter point offers further clarity to the question of *how* emotion versus race salience shapes racially biased weapon identification. If racial bias is weaker in the emotion-salient versus race-salient condition because participants attended more to emotion information in the face primes, then the effect of emotion expression on weapon identification (i.e., the Emotion Prime \times Target Object interaction) should be stronger when emotion versus race is salient. That is, if participants are paying more attention to emotion, then the effect of emotion expressions should be more impactful. And yet, we found no evidence that emotion salience moderated the impact of emotion expression on weapon identification.

Our findings suggest that the mere availability of obvious emotion expressions *does not* moderate racially biased weapon identification, but that increasing the salience of emotion expressions *does* moderate racial bias, relative to increasing salience of race. And yet, increasing the salience of emotion expressions failed to moderate the effects of emotion expression on weapon identification. It is unclear, therefore, if attention is simply being drawn *away* from race information without being drawn *toward* emotion information. This pattern of results underscores the importance of directly measuring the processing of prime-related content to clarify when and how salient cues are integrated into object identification. The current

instantiation of the DDM is ill-equipped to answer this question because it measures the decision process from target object onset, treating the influence of face primes as a response bias toward or away from “gun” decisions (i.e., a start-point bias).

In reality, the processing of facial information occurs over a time course rich in nuance (Freeman et al., 2020). Such nuance may be needed to understand where attention is directed at prime onset, and how such attention allocation affects later-stage processing. For example, the length of time spent processing information in the primes might flip the direction of their impact on decisions about targets (Klauer et al., 2009). We see value in future research that uses alternative computational approaches (e.g., Diederich & Trueblood, 2018) to capture the processing of face primes more directly. By dynamically measuring the processing of face primes in the WIT, future work may identify the amount of processing time required for various information in the face primes to have maximal impacts on later-stage processing.¹⁸

Future research should also test the generalizability of the initial bias account across other sources of salient facial information and different social groups. For example, information salience also shapes gender-stereotypic threat impressions (Rees et al., 2022), but it remains unclear where in the decision process these effects emerge. In addition, because we used only male face primes, future research should test whether racially biased weapon identification evoked by Black versus White women (Thiem et al., 2019) is likewise shaped by informational salience (cf. Petsko et al., 2022) and, if so, whether it is best explained by an initial bias account.

Our findings indicate that attending to emotion versus race information can weaken racially biased weapon identification. This phenomenon can be explained by salience-driven changes at the start of the decision process. Racial biases favoring a “gun” response before the object’s

¹⁸ We thank an anonymous reviewer for raising this point.

onset were weaker when emotion versus race was salient, pointing to a mechanism whereby the salience of person information moderates racially biased decision-making.

References

- Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88-93.
- Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 42, pp. 319-373). San Diego, CA: Elsevier Academic Press.
- Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*, 219-233.
- Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review, 125*, 270-292.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.
- Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 61, pp. 237-287). Elsevier.

- Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition and Emotion, 24*, 1008-1025.
- Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133-168). New York: Guilford Press.
- Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-878.
- Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology, 115*, 601-623.
- Jones, C. R., & Fazio, R. H. (2010). Person categorization and automatic racial stereotyping effects on weapon identification. *Personality and Social Psychology Bulletin, 36*, 1073-1085.
- Klauer, K. C., Teige-Mocigemba, S., & Spruyt, A. (2009). Contrast effects in spontaneous evaluations: A psychophysical account. *Journal of Personality and Social Psychology, 96*, 265-287.
- Kubota, J. T., & Ito, T. A. (2014). The role of expression and race in weapons identification. *Emotion, 14*, 1115-1124.
- 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*, 1122-1135.

- Macrae, C. N., & Cloutier, J. (2009). A matter of design: Priming context and person perception. *Journal of Experimental Social Psychology, 45*(4), 1012-1015.
- Niedenthal, P. M., & Brauer, M. (2012). Social functionality of human emotion. *Annual Review of Psychology, 63*, 259-285.
- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181-192.
- Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 62, pp. 1-50). Academic Press.
- Petsko, C. D., Rosette, A. S., & Bodenhausen, G. V. (2022). Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology, 123*, 763-787
- Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review, 25*, 1301-1330.
- Raissi, A., & Steele, J. R. (2021). Does emotional expression moderate implicit racial bias? Examining bias following smiling and angry primes. *Social Cognition, 39*, 570-590.
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*, 260-281.
- Rees, H. R., Sherman, J. W., Klauer, K. C., & Todd, A. R. (2022). On the use of gender categories and emotion categories in threat-based person impressions. *European Journal of Social Psychology, 52*, 597-610.

- Rivers, A. M. (2017). The weapons identification task: Recommendations for adequately powered research. *PLOS ONE*, *12*, e0177857.
- Taylor, S. E., & Fiske, S. T. (1978). Salience, attention and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249-288). New York, NY: Academic Press.
- Thiem, K. C., Neel, R., Simpson, A. J., & Todd, A. R. (2019). Are Black women and girls associated with danger? Implicit racial bias at the intersection of target age and gender. *Personality and Social Psychology Bulletin*, *45*, 1427-1439.
- Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology*, *120*, 672-693.
- Todd, A. R., Thiem, K. C., & Neel, R. (2016). Does seeing faces of young Black boys facilitate the identification of threatening stimuli? *Psychological Science*, *27*, 384-393.
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455-460.

Chapter 3

Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research

Abstract

Forming impressions of others is a fundamental aspect of social life. These impressions necessitate the integration of many and varied sources of information about other people, including social group memberships, apparent personality traits, inferences from observed behaviors, etc. However, methodological limitations have hampered progress in understanding this integration process. In particular, extant approaches have been unable to measure the independent contributions of multiple features to a given impression. In this article, after describing these limitations and their constraints on theory testing and development, we present a multinomial processing tree model as a computational solution to the problem. Specifically, the model distinguishes the contributions of multiple cues to social judgment. We describe an empirical demonstration of how applying the model can resolve long-standing debates among person perception researchers. Finally, we survey a variety of questions to which this approach can be profitably applied.

Keywords: person perception; impression formation; multinomial processing trees; computational modeling; stereotyping

Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research

Since the publication of Asch's seminal work (1946), perhaps the most fundamental objective in the research on person perception has been to understand how people combine the implications of multiple and varied features in judging others (see also Anderson, 1968). Cues relating to social group membership (e.g., racial appearance), personality traits (e.g., trustworthiness), emotions (e.g., anger), witnessed behaviors (e.g., an act of violence), and many other attributes may be relied upon in forming a coherent impression of another person. Though many influential models have been proposed to account for this complex task, testing them has been hindered by a limitation in measurement. In turn, this limitation has significantly slowed theoretical progress. In this paper, we detail the nature of the problem before offering a solution in the form of a computational modeling approach.

Theoretical Background

Models of person perception often posit how multiple features are integrated into a judgment. One of the prevailing claims these models make is that integrating different features occurs through a competitive process, such that relying more on one feature implies relying less on others. We refer to this as the *inverse relativity* assumption. In their initial presentations, both Brewer's (1988; 2014; see also Brewer & Feinstein, 1999) and Fiske and Neuberg's (1990; see also Fiske et al., 1999) influential models propose an inverse relationship between the use of social category (e.g., group stereotypes) and individuating (e.g. individual behavior) information: Increased stereotyping requires decreased individuation and vice versa. So, for example, if cognitive load is predicted to reduce the reliance on individuating behaviors, it should also increase the use of social stereotypes (e.g., Fiske & Neuberg, 1990). More recent models

similarly invoke inverse relativity. Consider Petsko and colleagues' (2022) Lens Model, which proposes that people use a variety of contextually activated lenses in perceiving others. However, according to the model, once one social category lens (e.g., race) has been activated, the use of other categories is necessarily diminished.

Beyond the inverse relativity assumption, another prevailing view in the person perception literature is that certain features *dominate* person perception (cf. Petsko & Bodenhausen, 2020)—that is, some cues are integrated into judgments by default and are highly impactful in determining social judgments. These models generally suppose that social category cues, particularly unambiguous visible cues to sex, race, and age, are processed more efficiently with fewer attentional resources than other cues (e.g., Brewer, 1988; Fiske & Neuberg, 1990). When person information is perceptually disfluent (e.g., inverted face; Cloutier et al., 2005) or a perceiver's cognitive or motivational resources are low (e.g., via a cognitive load task; Wigboldus et al., 2004), social categorization and, by extension, stereotyping is thought to remain active. However, the processing of cues that refer to the personal, individuating attributes of people, such as traits, states, and behaviors, is thought to operate insufficiently under such impoverished circumstances (e.g., Sherman et al., 2000; Swencionis & Fiske, 2013), augmenting the relative impact of social categories.

Of course, inverse relativity and category dominance are not the only perspectives in person perception research. For example, the Social Judgeability Model (SJM; Leyens et al., 1992; Yzerbyt et al., 1994; 1998) predicts that stereotyping is more likely when individuating features are available, if those individuating features provide perceivers with the subjective sense of being fair and decrease concerns with unfairly stereotyping a target (Darley & Gross, 1983; Norton et

al., 2004; Yzerbyt et al., 1994). Thus, this perspective posits that greater individuation may increase categorization (i.e., a direct relationship), contrasting the inverse-relativity perspective.

A class of network models (e.g., Freeman & Ambady, 2011; Kunda & Thagard, 1996) eschews both the inverse relativity and category dominance perspectives, assuming that all available features may be integrated, as in early models of impression formation (e.g., Asch, 1946; Andersen, 1968). They allow for the use of different features to be positively correlated, negatively correlated, or not correlated at all (Freeman et al., 2012). They also suggest that aspects of the perceiver, can affect which features are more or less dominant during the construal process (Freeman et al., 2020; see also Schwarz & Bless, 2010). Altogether, there is great flexibility in the model to account for almost any pattern of feature integration. This is both a strength and weakness of the model, as it does not make sufficiently precise predictions to be falsifiable as a general model of person perception, though some specific hypotheses may be testable (e.g., Freeman et al., 2012; for a more detailed discussion, see Petsko & Bodenhausen, 2020). For example, these models imply that cues processed earlier during person perception have more time and opportunity to influence final judgments.

A Multi-Cue Measurement Problem

Clear tests of the models laid out above require the ability to measure the separate impacts of multiple features on impressions and their theoretically proposed relationships (e.g., race dominating impressions over behavior). For instance, adequately testing whether cognitive load decreases individuation and increases categorization (e.g., Fiske & Neuberg, 1990), or decreases both processes (e.g., Spears & Haslam, 1997), requires that the impacts of social categories and person-specific cues be distinguished from each other. Unfortunately, conventional measurement approaches are unable to do so.

To illustrate the problem, consider an archetypal study that attempts to assess the extent to which different types of information influence judgments along some stereotype-relevant dimension (e.g., How threatening is Bob?). Those judgments, in and of themselves, cannot provide independent estimates of the impacts of social stereotypes (Bob is Black and therefore stereotypically threatening), Bob's somewhat threatening behavior, and Bob's smiling facial expression. In this case, a relatively stereotypic judgment of Bob as threatening may result from increased stereotyping, increased influence of his behavior, decreased impact of his facial expression, or all three. In turn, a relatively counter-stereotypic judgment may result from decreased stereotyping, decreased use of the behavior, increased use of the expression, or all three.

Consider also the classic finding that people tend to make more stereotypic judgments of suspects' alleged misbehavior when they are tested at the low point versus high point of their circadian cycles (Bodenhausen, 1990). This is the sort of evidence that has been seen to support prominent dual-process models and their assumptions about inverse-relativity and social category dominance: People make more stereotypic judgments when they have diminished processing capacity and motivation. Although findings like this serve as important illustrations, the extent to which different information contributes to these effects is unclear. Does reducing cognitive resources increase the use of social categories, decrease the use of individuating details about the person, or both? Alternatively, both features may be relied upon more or less, with the change in one being greater than the other. In all cases, the outcome is an increase in stereotypic judgments.

As another example, consider the finding that those with greater implicit bias are quicker to recognize happiness in White faces and anger in Black faces (Hugenburg & Bodenhausen, 2003).

Though an important demonstration of the effects of stereotypes on emotion perception (see also Weisbuch & Ambady, 2008), the extent to which different types of information contribute to the effect is unclear. Does construing Black faces quickly becoming angry reflect relying on race more, relying on facial expressions less, or some combination of changes in both features?

As a final illustration, consider mouse-tracking tasks, which instruct participants to move their cursor from a fixed starting position toward one of two (or more) response options based on the target stimulus provided. The extent to which the cursor initially moves toward one response before being tracked to the other response indicates the extent of conflict between the two response options and that both have been activated in parallel (e.g., Hehman et al. 2015; Stillerman & Freeman, 2019).

However, although mouse-tracking measures are excellent indicators of parallel activation and response conflict, they cannot distinguish the extents to which the two different sources of information influence cursor movement (Stillman et al., 2018). For example, when used to assess race categorization, participants show a stronger initial tendency to move the cursor toward White categorizations when an ambiguously Black target is wearing a suit versus a janitor's uniform (Freeman & Ambady, 2009). This measure of conflict between White and Black response options is interpreted to reflect an initially greater impact of clothing at the expense of race before a transition to a greater impact of race at the expense of clothing. However, the varying influence of each feature cannot be distinguished from the other. The measures are inherently relative and pit the use of each cue against the other in an inverse fashion.

A Multi-Cue Integration Model

Here, we propose a solution to the multi-cue measurement problem in the form of a computational model that we named the multi-cue integration (MCI) model. The MCI model is a

multinomial processing tree (MPT), a class of cognitive models comprised of a set of equations to identify and measure the extent of processes underlying responses in a task (for reviews, see Batchelder & Riefer, 1999; Calanchini et al., 2018; Erdfelder et al., 2009; Hütter & Klauer, 2016; Sherman et al., 2010). Like any MPT, the MCI model is built on a small set of parameters – C_1 , C_2 , and g – with each parameter reflecting the probability of a unique cognitive processing state (e.g., the integration of sex information into an impression). The C_1 and C_2 parameters each reflect the probability of a unique source of information being used to form judgments, whereas g reflects a response bias toward one response over another. If targets in a gender classification task vary in both sex and facial expression, then C_1 could be assigned to reflect the probability of sex cues being used when classifying target faces, whereas C_2 could be assigned to reflect the use of facial expressions for those very same classifications.

Visually, the relationships among these parameters can be depicted as a processing tree, as seen in Figure 9. The MCI model assumes that the probability of using the information assigned to C_2 (e.g., facial expressions) is contingent upon the probability that using the information assigned to C_1 (e.g., sex cues) is insufficient for deriving a particular judgment $[(1 - C_1) \times C_2]$. Although the parameters and their relationships among one another remain the same across judgment tasks, the number of equations used to model the data are determined by the number of unique responses on that task. That is, the MCI model produces an equation for each unique response that can be observed in a judgment task. A task with 12 unique responses would require 12 unique equations, derived from the MCI model's parameters.

It is noteworthy to further highlight what the C_1 and C_2 processing parameters represent. Traditional cognitive models of person perception focus on the various mechanisms (e.g., activated associations, recognition memory, correct response detection) that turn input

information into social judgments. MPTs have been very useful for such investigations (e.g., Heycke & Gawronski, 2020; Klauer & Wegner, 1998; Krieglmeier & Sherman, 2012), as they traditionally quantify how often the various mechanisms work to generate those judgments. The MCI, however, is unique in that it focuses on quantifying the extent to which specific input features are used to form social judgments. The C_1 and C_2 parameter estimates encapsulate the cumulative processing of these features, across whatever mechanisms may be involved. That is, the MCI model offers a quantitative assessment on each feature's impact, summed across all the mechanisms by which they may be used to derive judgments. To illustrate more fully, we describe an experiment designed to test the model's validity for a particular judgment task and its capacity for theory testing and development in person perception research.

Demonstrating the MCI Model

Participants ($N = 593$; Klein & Sherman, 2024) classified faces varying in facial cues to sex (male cues, female cues) and expression (e.g., scowling, smiling). Using morphing techniques described in Appendix C, both cues were manipulated to appear either ambiguous or unambiguous. By assigning participants to classify faces by gender or emotion, the relevance of sex and expression information were manipulated between-participants. As previously stated, the number of equations the MCI model derives depends on the number of unique responses in the task. Here, the MCI model derives 8 unique equations (2 [judgment: man, woman; or angry, happy] $\times 2$ [sex cues: male, female] $\times 2$ [emotion expression: scowling, smiling] equations). For example, separate equations were derived for predicting how often smiling male faces were classified as a man versus woman.

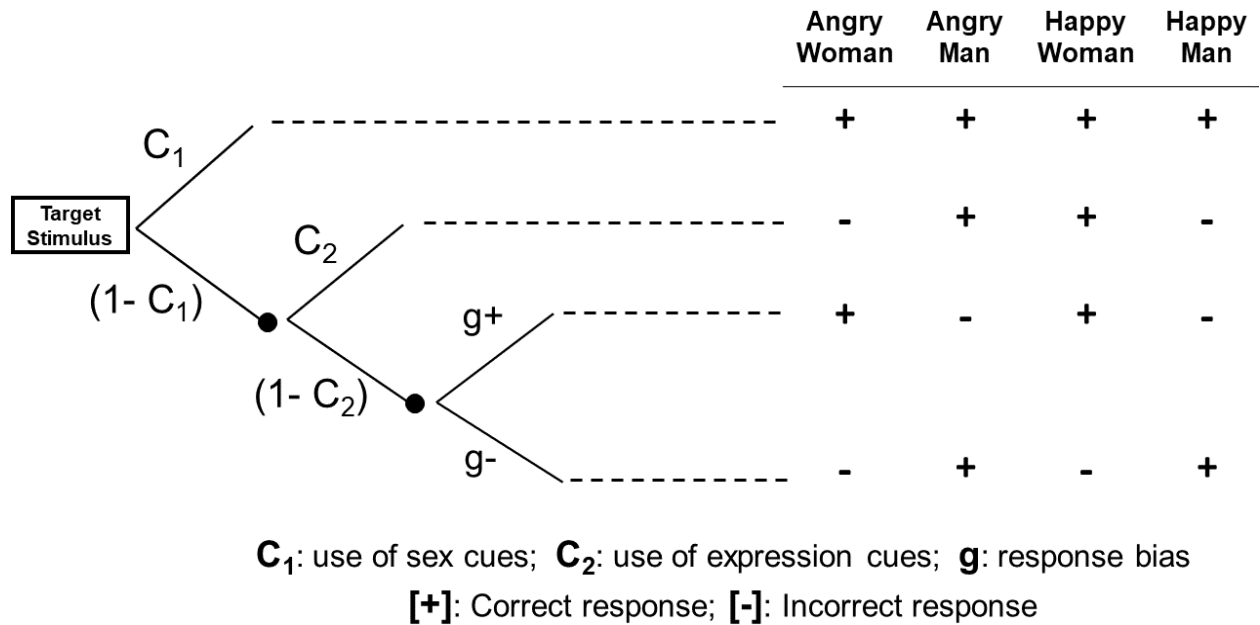
Following along the tree in Figure 9, for the gender classification task, the probability of classifying a happy male face as a man is predicted by the joint contributions of male facial cues

[C_1] and a tendency to categorize faces as men whenever sex and expression cues are insufficient to derive a coherent judgment $[(1 - C_1) \times (1 - C_2) \times (1 - g)]$ – that is, a response bias toward *man*.

The compliment of that probability, the equation for classifying a happy male target as a woman, is predicted by the joint contributions of a smiling facial expression $[(1 - C_1) \times C_2]$ and a tendency to categorize faces as women whenever sex and emotion cues are insufficient to derive a coherent judgment $[(1 - C_1) \times (1 - C_2) \times g]$ – that is, a response bias toward *woman*. Therefore, by simply following the paths along the tree, the equations predicting each unique response can be derived (The full set of equations are displayed in Appendix C).¹⁹

Figure 9

Chapter 3: The MCI Model and Its Predicted Responses to Gender Classifications



Notes. Diagram of the MCI model used to measure person perception data from a paradigm in which judgments were made of targets varying in sex and expression cues. The manifest outcome is represented on the right side of the figure (i.e., binary responses about the person’s gender). The paths along the tree depict the processing paths assumed by the model to explain responses for each trial type.

¹⁹ More detailed discussions of the mechanics of using MPTs is beyond the scope of this text. We recommend general (e.g., Schmidt et al., 2022) and software-specific (e.g., Hartmann et al., 2020; Heck et al., 2018; Moshagen, 2010; Stahl & Klauer, 2007; Singmann & Kellen, 2013) tutorials for instructions on developing and applying MPTs.

MPTs are theoretically derived models, and the MCI model relies on well-established stereotypes we assume participants rely on when forming judgments. Here, the MCI model relies on the stereotype linking men (women) and negative (positive) expressions: the model assumes that expression information is used when smiling faces are classified as *woman* and scowling faces as *man*, but not the other way around. For example, the equation for judging a smiling male face as *woman* $[(1 - C_1) \times C_2 + (1 - C_1) \times (1 - C_2) \times g]$ includes the assumption that smiling expressions are associated with *woman* and not *man* (see Hess et al., 2007). These assumptions are required to identify the model and can be tested by examining whether the model adequately predicts the observed responses (i.e., model fit).

The parameters are estimated by entering the frequencies of participants' actual responses as outcomes in the equations, and their values reflect the probability that their respective processing component contributes toward the observed responses. Each estimated parameter can vary independently of all others, yielding distinct estimates for the relative contributions of each component.

Applying the MCI Model

Model Fit. First and foremost, the MCI model fits well to both the gender classification judgments, Median Individual T_1 p -value = .558, Aggregate T_1 p -value < .001, Aggregate T_2 p -value = .002, $w = .02$, and emotion classification judgments, Median Individual T_1 p -value = .538, Aggregate T_1 p -value = .094, Aggregate T_2 p -value = .192, $w < .01$, albeit far better fitting for emotion classification judgments. Assessment of model fit includes visual examination of the posterior predictions against the observed response frequencies and covariances. Visually, we also plotted expected versus observed mean frequencies (T_1) and covariances (T_2). The observed means and covariances generally fall within the range of box-plotted model expectations, indicating good fit

(see Figures C1-C4). Visual inspection is both inherently Bayesian and a common tool for MPT modeling using the current estimation strategy (e.g., Calanchini et al., 2021; Smith et al., 2019).

Parameter Comparisons. If the MCI model measures the distinct contributions of sex and expression information, we would expect the estimated use of each cue to be greater when it was relevant versus irrelevant to the intended judgment. Indeed, sex cues were used more and expressions were used less during gender versus emotion classification. We would also expect that task-relevant cues (e.g., sex cues during gender classification) would be used less when ambiguous. Aligned with this expectation, introducing ambiguity in sex cues decreased their use during gender classification (Figure 10), whereas introducing ambiguity in expressions decreased their use during emotion classification (Figure 11).

Parameter Correlations. As we previously discussed, a prominent assumption in the person perception literature is that two features are integrated in a competition (e.g., Fiske & Neuberg, 1990). If one feature contributes more, it is at the expense of the other feature's contribution to the judgment. However, alternative relationships have also been proposed, such as positive associations between the two features (e.g., Leyens et al., 1992) – categorization is sometimes thought to increase when individuation does as well. To diagnose these competing accounts, we can examine the correlation between the use of each source of information. Here, we focus on trials when neither feature was ambiguous. For gender classification judgments, the MCI model identified a credible and positive correlation between the use of sex and expression cues $r = .64$, $BCI_{95\%} [.29, .92]$. For emotion classification judgments of the same targets, however, the model failed to identify any association between the use of the two cues, $r = .13$, $BCI_{95\%} [-.92, .90]$.

Model Comparison. As we have noted, extant theory contends with competing predictions about which cues are processed by default. Arguably the most prominent assumption is one in

which social categorization serves as the default process. Regardless of alternative sources of information or the intended judgment, social categories are often thought to be integrated into impressions (Brewer, 1988; Fiske & Neuberg, 1990; Hugenberg et al., 2010). Alternative perspectives suggest that the intended judgment – that is, a perceiver’s goal – and other motives determine which information is more likely to be integrated into an impression by default (Freeman et al., 2020; Petsko et al., 2022; Schwarz & Bless, 2010).

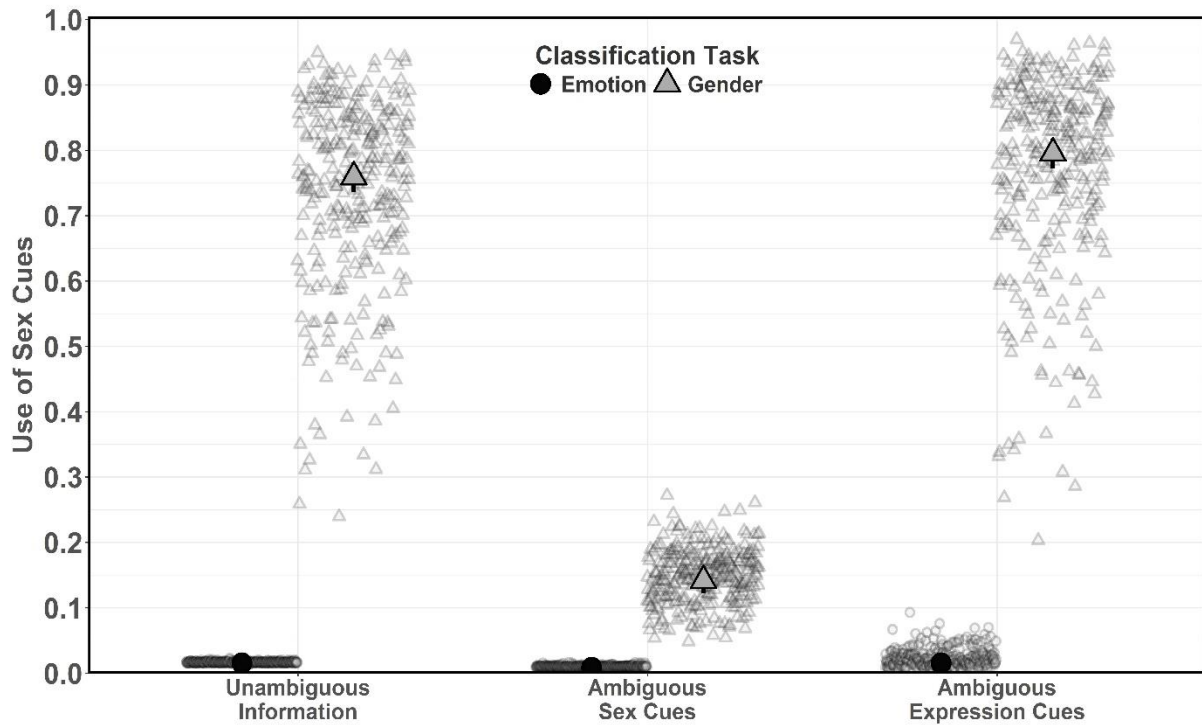
A strength of the MCI model is that it offers a framework within which to formalize and test competing default-processing assumptions. The model’s equations establish conditional relationships among the parameters, assuming that the use of the second feature is contingent upon the first feature being insufficient for producing the judgment $[(1 - C_1) \times C_2]$. By fitting the MCI model both when C_1 is assigned to one cue versus the other, we can identify whichever model variant better characterizes the data (for similar approaches, see Calanchini et al., 2022; and Laukenmann et al., 2023). Here, we demonstrate this procedure by fitting the MCI when C_1 reflects sex processing and again when it reflects expression processing.

After fitting both versions of the model, we compared their Deviance Information Criteria (DIC) to determine which version offers a better characterization of the observed judgments. In both the gender classification task ($\Delta DIC = -2.74$) and emotion classification task ($\Delta DIC = 8.72$), comparison of the two models yielded substantive evidence for a default-sex model. That is, regardless of which cue was more relevant to the intended judgment, sex cues were integrated by default, whereas expression cues were better characterized as being used if sex cues, alone, were insufficient to derive the judgment. It remains an open question as to whether all social categories *dominate* person perception. This procedure, therefore, should be replicated and generalized across various social categories (e.g., race, age) and identity-specific cues (e.g., other

expressions, behaviors), and across various intended judgments (e.g., gender classification, gender-typical versus gender-atypical trait impressions).

Figure 10

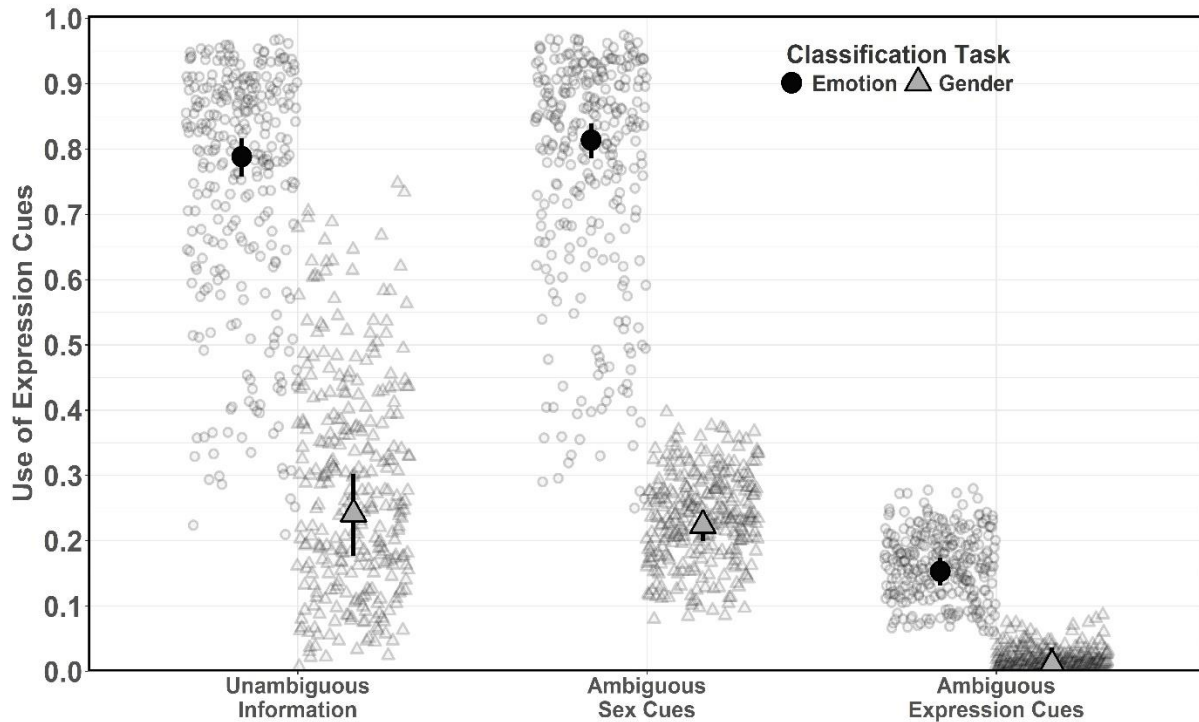
Chapter 3: Estimated Use of Sex Cues During Face Classification



Notes. Markers reflect the estimated use of sex cues during face classification by gender (triangles) or emotion (circles). Solid markers reflect aggregate-level estimates, whereas empty markers reflect individual-level estimates. The x-axis reflects whether target face stimuli were presenting ambiguous sources of information. The y-axis reflects the estimated probability of relying on sex cues when classifying target faces. Error bars signify 95% Bayesian credibility intervals around the aggregate-level estimate.

Figure 11

Chapter 3: Estimated Use of Expression Cues During Face Classification



Notes. Markers reflect the estimated use of expression cues during face classification by gender (triangles) or emotion (circles). Solid markers reflect aggregate-level estimates, whereas empty markers reflect individual-level estimates. The x-axis reflects whether target face stimuli were presenting ambiguous sources of information. The y-axis reflects the estimated probability of relying on expression cues when classifying target faces. Error bars signify 95% Bayesian credibility intervals around the aggregate-level estimate.

Summary

This initial pilot study demonstrates that the MCI provides an accurate account of multi-feature integration in person perception. Further, the results highlight the model's potential for theory testing and development. For instance, the lack of negative correlation between the use of two clear cues challenges the inverse relativity assumption that increases in the use of one feature should coincide with decreases in the other. Obviously, this singular empirical demonstration does not offer a thorough test of inverse-relativity, but it does highlight the need

for research that applies this technique to thoroughly examine *how* various cues are integrated together into social judgments.

We also demonstrated how the MCI model can be applied to test dominance assumptions in person perception research, which generally assume that one feature (usually representing social categories) is used more efficiently, acts as a default, and is more impactful in judgments than other features. Here, we find that sex cues were, indeed, better characterized as a default process, even when expression cues were more relevant to the judgment at hand (i.e., emotion classification). Again, these data are illustrative but preliminary. Considerable further work will be required to draw any broad claims about the kinds of features that tend to dominate and the conditions under which they do so.

Further Applications of the MCI Model

The MCI model offers a flexible solution for testing key questions and theories surrounding person perception that can be applied to most tasks in which judges must select among discrete options. In this paper, we introduce and initially validate the MCI model as one that can capture information processing behind binary classifications of faces by gender and emotion. However, the same framework could be applied to judgments of race, age, or personality traits, or to decision-making given a variety of kinds of available information (e.g., hiring context; Axt et al., 2018), so long as each target belongs to only one level of each dimension measured by the MCI model. MPTs like the MCI model also can be redrawn to accommodate a broader range of data, including data from tasks with three response options (e.g., Klauer & Wegener, 1998). The model also can be extended to include *both* discrete responses and continuous data, such as response times (Heck et al., 2016; Klauer & Kellen, 2018) and mouse-tracking (Heck et al.,

2018), if both are presumed to be integral for explaining the cognitive processing underlying judgments.

Consider the benefits of integrating response times into the MCI model. Doing so could (1) estimate the speeds at which different features lead to judgments and (2) test the temporal order by which two features are processed during person perception. As previously discussed, social categorization is thought to occur prior to the processing of other, more identity-specific information (e.g., Fiske & Neuberg, 1990; Hugenberg et al., 2010). Including response times into the MCI model framework, and subsequently testing the temporal order between social categories and more identity-specific cues, offers a direct test of this assumption. Although we have not yet developed versions of the MCI model to accommodate nonbinary discrete responses or the inclusion of continuous data, it is certainly possible to do so.

Testing Dominance Assumptions of Person Perception Models

As mentioned earlier, another facet of the general assumption that social categories *dominate* person perception is the claim that they are more efficiently processed and applied than other information (e.g., individuating behaviors). As such, these models predict greater impact of social categories and lesser impact of individuating features, especially when perceivers have limited processing capacity (e.g., Brewer, 1988; Fiske & Neuberg, 1990). The supposed efficiency of activation and application of social category stereotypes implies that their processing should be unaffected or even increased when the perceiver is under cognitive load or time pressure, for example. Individuating expressions, traits, and behaviors, on the other hand, are assumed to be applied less fully under those same conditions (e.g., Sherman et al., 2000; Swencionis & Fiske, 2013).

Those same theoretical models of impression formation and social inference also propose that perceivers vary their use of different attributes as a function of their motivation to judge a target accurately (e.g., Fiske, Lin, & Neuberg, 1999; Fiske & Neuberg, 1990). Specifically, according to these models, increased accuracy motivation (via internal motives, interdependence with the target, etc.) should decrease the use of social category information and increase the use of individuating personal information. The MCI model can be applied to directly test these hypotheses by providing a means for estimating the independent contributions of different cues, which, to date, has not been possible.

The MCI model can also be applied to test the extent to which various features are used depending on what other information is also available. Our empirical demonstration measured the use of sex and expression cues to classify faces. However, if those faces varied in sex and race cues instead, would sex cues be used differently than when expression was the alternatively available information? By implementing the MCI model across various information pairings (e.g., sex and expression, sex and race, sex and traits), we can better understand the extent to which the use of specific features is context-general versus context-specific in person perception.

Context Effects on Person Perception

Another central goal of person perception research is to assess the independent contributions of target features (e.g., traits) and situational details in impression formation. Process models designed to account for the supposed under-use of social context on person perception (i.e., the “Fundamental Attribution Error”) propose that inferences about the situation surrounding a person are made less efficiently than inferences about the person’s traits (Gilbert, 1989; Trope, 1986). Accordingly, these models propose that cognitive load reduces the integration of

situational information but does not impair the use of person information (e.g., personality traits) in person perception.

More broadly, a key question in person perception research concerns the joint contributions of person cues and context cues on impression formation. Among many other examples, researchers have investigated the contributions of background imagery (e.g., Brambilla et al., 2018), clothing cues (Freeman et al., 2011; Oh et al., 2020), and accessory items (e.g., tools or guns; Fessler et al., 2012), on person perception. In some cases, researchers have avoided making inferences about the contributions of each cue (e.g., Fessler et al., 2012); in others, cues are assumed to be integrated inversely from one another (e.g., Brambilla et al., 2018; Freeman et al., 2013; Xie et al., 2022). The MCI model provides a means for directly investigating such questions.

Multiply Categorizable Person Perception

All people simultaneously belong to multiple groups based on sex, race, age, etc. In recent years, increasing attention has been paid to how impressions are based not on a single social category, but rather multiple categories (e.g., Kang & Bodenhausen, 2015). This research has revealed considerable nuance in group-based judgments of and behavior toward other people. For example, judgments about a target's sex may vary as a function of target race (Johnson et al., 2012). Judgments of leadership ability may be affected by an interaction between the target's race and sexual orientation (Wilson et al., 2017). Basic intergroup bias favoring ingroups over outgroups may be attenuated if the target and perceiver share a common identity (e.g., Calanchini et al., 2022; Scroggins et al., 2016). However, the literature on judgments of multiply categorizable targets has yet to disentangle the contributions of each category cue. For example, the extents to which each social category plays a role in Black women being mistaken for and

stereotyped as men more frequently than White women (e.g., Kang & Bodenhausen, 2015) is not clear. Do perceivers rely on Black cues more (stereotypically emphasizing masculine qualities), female cues less (stereotypically minimizing feminine qualities), or both? These kinds of questions are can be addressed with the MCI model.

Conclusion

The judgments we make about people are foundational to when, how, and why we treat them the way we do. Theoretical progress in person perception research has been hindered by an inability to distinguish the contributions of multiple available cues to social judgment. Is the processing of social categories highly efficient? Does accuracy motivation reduce the use of social categories and increase the use of identity-specific cues, or both? Is the integration of situational constraints in understanding behavior particularly inefficient? More broadly, to what extent do people integrate personal and contextual features in person perception? Do certain features dominate impressions? If so, are these dominant features processed first, by default, more efficiently, more often, or by some combination of these facets? These questions cannot be addressed effectively without disentangling the contributions of each source of information. The MCI model offers a solution to this multi-cue measurement problem.

References

- Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of personality and social psychology*, 9(3), 272.
- Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3), 258.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of personality and social psychology*, 92(2), 179.
- Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science*, 1(5), 319-322.
- Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology*, 78, 34-42.
- Brewer, M. B. (1988). A dual process model of impression formation. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *A dual-process model of impression formation: Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum
- Brewer, M. B. (2014). A dual process model of impression formation. In *Advances in Social Cognition, Volume I* (pp. 1-36). Psychology Press.
- Brewer, M. B., & Feinstein, A. S. H. (1999). Dual processes in the cognitive representation of persons and social categories.

Burnham, K. P., & Anderson, D. R. (2004). Model selection and multimodel inference. *A practical information-theoretic approach*, 2.

Calanchini, J., Rivers, A. M., Klauer, K. C., & Sherman, J. W. (2018). Multinomial processing trees as theoretical bridges between cognitive and social psychology. In *Psychology of learning and motivation* (Vol. 69, pp. 39-65). Academic Press.

Calanchini, J., Schmidt, K., Sherman, J. W., & Klein, S. A. (2022). The contributions of positive outgroup and negative ingroup evaluation to implicit bias favoring outgroups. *Proceedings of the National Academy of Sciences*, *119*(40), e2116924119.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20.

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108-124.

Fessler, D. M., Holbrook, C., & Snyder, J. K. (2012). Weapons make the man (larger): Formidability is represented as size and strength in humans. *PloS one*, *7*(4), e32751.

Fiske, S. T., Lin, M., & Neuberg, S. L. (2018). The continuum model: Ten years later. *Social cognition*, 41-75.

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 *Advances in experimental social psychology* (Vol. 23, pp. 1-74). Academic Press.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological science*, *20*(10), 1183-1188.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological review*, *118*(2), 247.

Freeman, J. B., Johnson, K. L., Adams Jr, R. B., & Ambady, N. (2012). The social-sensory interface: category interactions in person perception. *Frontiers in integrative neuroscience*, *6*, 81.

Freeman, J. B., Ma, Y., Han, S., & Ambady, N. (2013). Influences of culture and visual context on real-time social categorization. *Journal of experimental social psychology*, *49*(2), 206-210.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PloS one*, *6*(9), e25107.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). The Guilford Press.

Hartmann, R., Johannsen, L., & Klauer, K. C. (2020). rtmpt: An R package for fitting response-time extended multinomial processing tree models. *Behavior Research Methods*, *52*, 1313-1338.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior research methods*, *50*, 264-284.

Helman, E., Stoller, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384-401.

Hess, U., Adams Jr, R. B., & Kleck, R. E. (2004). Facial appearance, gender, and emotion expression. *Emotion*, *4*(4), 378.

Hess, U., Adams, R. B., Grammer, K., & Kleck, R. E. (2009). Face gender and emotion expression: Are angry women more like men?. *Journal of Vision*, 9(12), 19-19.

Hess, U., Thibault, P., Adams Jr, R. B., & Kleck, R. E. (2010). The influence of gender, social roles, and facial appearance on perceived emotionality. *European Journal of Social Psychology*, 40(7), 1310-1317.

Heycke, T., & Gawronski, B. (2020). Co-occurrence and relational information in evaluative learning: A multinomial modeling approach. *Journal of Experimental Psychology: General*, 149(1), 104.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, 14(6), 640-643.

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, 27(1), 116-159.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: how covarying phenotypes and stereotypes bias sex categorization. *Journal of personality and social psychology*, 102(1), 116.

Kang, S. K., & Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: Challenges and opportunities. *Annual review of psychology*, 66, 547-574.

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, 75(5), 1155.

Klein, S. A. W., & Sherman, J. W. (2024). *Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research* [Data set, model, and code]. OSF. osf.io/gxabc5

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of personality and social psychology, 103*(2), 205.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological review, 103*(2), 284.

Leyens, J. P., Yzerbyt, V. Y., & Schadron, G. (1992). The social judgeability approach to stereotypes. *European review of social psychology, 3*(1), 91-120.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*(1), 42-54.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of personality and social psychology, 87*(6), 817.

Oh, D., Shafir, E., & Todorov, A. (2020). Economic status cues from clothes affect perceived competence from faces. *Nature human behaviour, 4*(3), 287-293.

Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A process model of affect misattribution. *Personality and social psychology bulletin, 36*(10), 1397-1408.

Petsko, C. D., & Bodenhausen, G. V. (2020). Multifarious person perception: How social perceivers manage the complexity of intersectional targets. *Social and Personality Psychology Compass, 14*(2), e12518.

Petsko, C. D., Rosette, A. S., & Bodenhausen, G. V. (2022). Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology*.

Schmidt, O., Erdfelder, E., & Heck, D. W. (2022, March 14). How to Develop, Test, and Extend Multinomial Processing Tree Models: A Tutorial. <https://doi.org/10.1037/met0000561>

Scroggins, W. A., Mackie, D. M., Allen, T. J., & Sherman, J. W. (2016). Reducing prejudice with labels: Shared group memberships attenuate implicit bias and expand implicit group boundaries. *Personality and Social Psychology Bulletin*, *42*(2), 219-229.

Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 156–174). The Guilford Press.

Sherman, J. W., Macrae, C. N., & Bodenhausen, G. V. (2000). Attention and stereotyping: Cognitive constraints on the construction of meaningful social impressions. *European review of social psychology*, *11*(1), 145-175.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560-575.

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, *39*, 267-273.

Stillerman, B. S., & Freeman, J. B. (2019). Mouse-Tracking to Understand Real-Time Dynamics of Social Cognition 1. In *A handbook of process tracing methods* (pp. 146-160). Routledge.

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in cognitive sciences*, *22*(6), 531-543.

Swencionis, J. K., & Fiske, S. T. (2013). More human: Individuation in the 21st century. In *Humanness and dehumanization* (pp. 284-301). Psychology Press.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological review*, *93*(3), 239.

Wilson, J. P., Remedios, J. D., & Rule, N. O. (2017). Interactive effects of obvious and ambiguous social categories on perceptions of leadership: When double-minority status may be beneficial. *Personality and Social Psychology Bulletin*, 43(6), 888-900.

Xie, S. Y., Thai, S., & Hehman, E. (2023). Everyday perceiver-context influences on impression formation: No evidence of consistent effects. *Personality and Social Psychology Bulletin*, 49(6), 955-968.

Yzerbyt, V. Y., Dardenne, B., & Leyens, J.-Ph. (1998). Social judgeability concerns in impression formation. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 126-156). London: Sage.

Yzerbyt, V. Y., Schadron, G., Leyens, J. P., & Rocher, S. (1994). Social judgeability: The impact of meta-informational cues on the use of stereotypes. *Journal of Personality and Social psychology*, 66(1), 48.

Appendix A

Revising Mental Representations of Faces Based on New Diagnostic Information

Linear Mixed-Effects Models (LMEMs)

Fixed-Effects Tables

Table A1

Trait Impressions of Robert by Fixed Effects of Time, Time 1 Induction, Time 2 Information, and All Interactions (Image-Generation Experiment)

Fixed Effects	<i>b</i>	<i>SE</i>	<i>t</i>	<i>df</i>
Intercept	3.64***	0.13	28.16	7.85
Time	0.18**	0.04	4.00	8.18
Time 1 Induction	-1.25***	0.18	-7.10	6.91
Time 2 Information	-0.23***	0.06	-3.90	35.07
Time × Time 1 Induction	-0.46***	0.03	-18.15	280.97
Time × Time 2 Information	0.20***	0.03	7.77	280.97
Time 1 Induction × Time 2 Information	0.34**	0.08	4.21	13.28
Time × Time 1 Induction × Time 2 Information	-0.37***	0.03	-14.36	280.97

Note. * $p < .05$ ** $p < .01$ *** $p < .001$; degrees of freedom (*df*) were calculated using Satterthwaite approximation.

Table A2

Trait Ratings of Group Classification Images by Fixed Effects of Time, Time 1 Induction, Time 2 Information, and All Interactions (Image-Assessment Experiment 1)

Fixed Effects	<i>b</i>	<i>SE</i>	<i>t</i>	<i>df</i>
Intercept	3.98***	0.26	15.37	6.17
Time	0.02	0.01	1.23	8358.00
Time 1 Induction	-0.34***	0.04	-8.87	154.00
Time 2 Information	-0.09***	0.01	-5.93	8358.00
Time × Time 1 Induction	-.015***	0.01	-10.20	8358.00
Time × Time 2 Information	0.05***	0.01	3.64	8358.00
Time 1 Induction × Time 2 Information	0.11***	0.01	7.17	8358.00
Time × Time 1 Induction × Time 2 Information	-0.11***	0.01	-7.23	8358.00

Note. * $p < .05$ ** $p < .01$ *** $p < .001$; degrees of freedom (*df*) were calculated using Satterthwaite approximation.

Table A3

Trait Ratings of Subgroup Classification Images by Fixed Effects of Time, Time 1 Induction, Time 2 Information, and All Interactions (Image-Assessment Experiment 2a)

Fixed Effects	<i>b</i>	<i>SE</i>	<i>t</i>	<i>df</i>
Intercept	3.97***	0.08	49.43	153.49
Time	-0.01	0.03	-0.21	45.04
Time 1 Induction	-0.55***	0.05	-10.61	81.29
Time 2 Information	-0.10*	0.04	-2.31	45.95
Time × Time 1 Induction	-0.29***	0.04	-8.12	58.67
Time × Time 2 Information	0.07*	0.03	2.23	44.26
Time 1 Induction × Time 2 Information	0.15***	0.04	3.41	48.40
Time × Time 1 Induction × Time 2 Information	-0.11**	0.03	-3.27	45.27

Note. * $p < .05$ ** $p < .01$ *** $p < .001$; degrees of freedom (*df*) were calculated using Satterthwaite approximation.

Table A4

Trait Ratings of Individual Classification Images by Fixed Effects of Time, Time 1 Induction, Time 2 Information, and All Interactions (Image-Assessment Experiment 2b)

Fixed Effects	<i>b</i>	<i>SE</i>	<i>t</i>	<i>df</i>
Intercept	3.74***	0.05	74.76	315.60
Time	0.02	0.01	1.18	314.38
Time 1 Induction	-0.22***	0.02	-10.19	395.34
Time 2 Information	-0.04	0.02	-1.97	281.41
Time × Time 1 Induction	-0.13***	0.01	-8.78	314.36
Time × Time 2 Information	0.02	0.01	1.17	278.53
Time 1 Induction × Time 2 Information	0.07***	0.02	3.65	291.52
Time × Time 1 Induction × Time 2 Information	-0.03*	0.01	-2.43	280.36

Note. * $p < .05$ ** $p < .01$ *** $p < .001$; degrees of freedom (*df*) were calculated using Satterthwaite approximation.

Random-Effects Structures

Image-Generation Experiment

In this and all other linear mixed-effects models (LMEMs) reported below, we aimed to maintain the largest possible random-effects structure without convergence failure or singularity. The two sources of variance accounted for in the image-generation experiment were participants (i.e., image generators) and traits (i.e., image generators' trait impressions of Robert). We reverse-scored the negatively-valenced traits (*mean*, *dominant*, and *aggressive*), ensuring that the traits can be interpreted as a sample drawn from a population of positive traits. The maximal random-effects structure included intercepts for participants and traits, a by-participant slope for Time, and by-trait slopes for Time, Time 1 induction, Time 2 information, and all possible interactions among those variables. Time is the only within-participant variable. Time, Time 1 induction, and Time 2 information are all within-trait variables.

This maximal model failed to converge. Thus, we downsized the random-effects structure until the model converged. The higher-order random-effects interactions involving traits proved most problematic, so we removed the three-way interaction, along with the Time \times Time 1 and Time \times Time 2 interactions. The final random-effects structure included intercepts for participants and traits, a by-participant slope for Time, and by-trait slopes for Time, Time 1 induction, Time 2 information, and the Time 1 induction \times Time 2 information interaction.

Image-Assessment Experiment 1

The two sources of variance accounted for in image-assessment Experiment 1 were participants (i.e., image raters) and traits (i.e., image raters' trait impressions of the group classification images of Robert). As before, we reverse-scored the negative-valenced traits. Because each image rater rated all 8 group images on each of the 7 traits, Time, Time 1

induction, and Time 2 information are within-participant and within-trait variables. Therefore, the maximal random-effects structure includes intercepts for participants and traits, and both by-participant and by-trait slopes for Time, Time 1 induction, Time 2 information, as well as all possible interactions involving those variables.

This maximal model failed to converge without singular fit. Therefore, we downsized the random-effects structure until singular fit was eliminated. All by-trait slopes riddled the model with singular fit. So too did all by-participant slopes, except a slope for Time 1 induction. Therefore, the final random-effects structure included intercepts for participants and traits, a by-participant slope for Time, and by-trait slopes for Time, Time 1 induction, Time 2 information, and the Time 1 induction \times Time 2 information interaction.

Image-Assessment Experiment 2a

The two sources of variance accounted for in image-assessment Experiment 2a were participants (i.e., image raters) and stimuli. Here, “stimuli” refers to the subgroup of image generators whose data were used to generate subgroup classification images at Time 1 and Time 2. With 96 subgroup images, there were 48 subgroup stimuli IDs, with each ID corresponding to the pair of Time 1 and Time 2 subgroup images generated by the same subgroup of image generators. Because each image rater rated all 96 subgroup images, Time, Time 1 induction, and Time 2 information were all within-participant variables. Because each subgroup image ID contained a Time 1 and Time 2 image, Time was also a within-stimuli variable. Therefore, the maximal random-effects structure included intercepts for participants and stimuli, by-participant slopes for Time, Time 1 induction, and Time 2 information, and all possible interactions involving those variables, as well as a by-stimuli slope for Time. This maximal model converged without any concerns over singular fit. Therefore, we reported the maximal model.

Image-Assessment Experiment 2b

The two sources of variance accounted for in image-assessment Experiment 2b were participants (i.e., image raters) and stimuli. Here, “stimuli” refers to the image generators of the individual classification images. These image generator IDs correspond to the 285 pairs of images (Time 1 and Time 2) created by each image generator. Because each image rater was randomly assigned images to rate from all possible conditions, Time, Time 1 induction, and Time 2 information were within-participant variables. With two time points of image generations per image generator, Time was a within-stimuli variable. Therefore, the maximal random-effects structure included intercepts for participants and stimuli, by-participant slopes for Time, Time 1 induction, and Time 2 information, and all possible interactions involving those variables, as well as a by-stimuli slope for Time.

This maximal model failed to converge without singular fit. Therefore, we downsized the random-effects structure until singular fit was no longer a concern. After removing the by-participant three-way interaction slope, singular fit was eliminated. Therefore, the final random-effects structure included intercepts for participants and traits, by-participant slopes for Time, Time 1 induction, Time 2 information, and all two-way interactions between the variables, as well as a by-stimuli slope for Time.

Analyses of Variance (ANOVAs)

Based on recommendations received during the editorial process, we reported LMEMs that account for additional sources of variance (traits in both the image-generation experiment and image-assessment Experiment 1; stimuli in both image-assessment Experiments 2a and 2b) in the main text. Here, we report analyses of variance (ANOVAs) that do not account for those additional sources of variance.

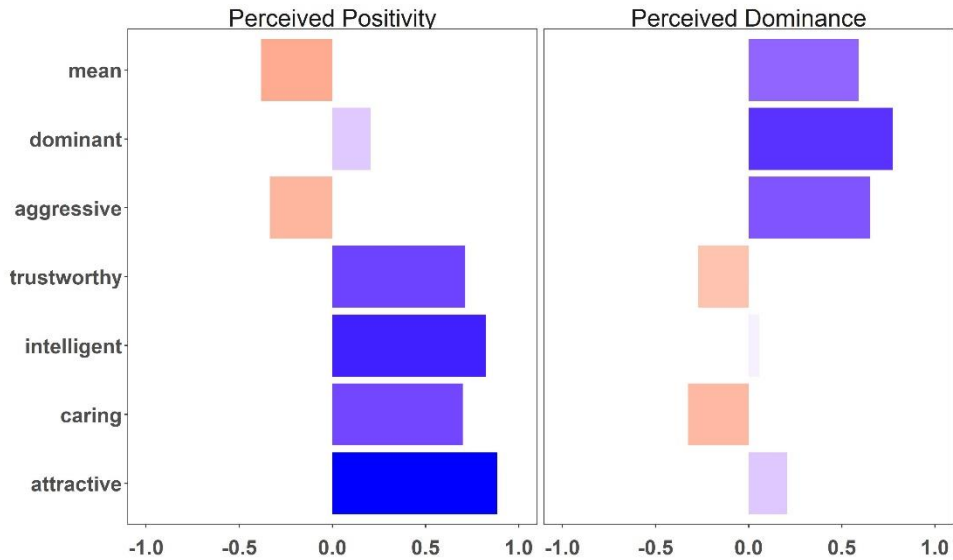
Image-Generation Experiment

Data Reduction

To limit the dimensionality of our data, we used the *psych* 1.8.12 package (Revelle, 2018) to conduct an exploratory factor analysis (EFA) with a *promax* rotation. We arrived at a two-factor solution, $\chi^2(570) = 6.69, p < .001$, accounting for 63% of the total variance (Factor 1 = 44%; Factor 2 = 27%; Figure A1). Although a three-factor solution also fit the data, the two-factor solution made more theoretical sense. Similar to other two-factor solutions in the face-impression literature (e.g., Oosterhof & Todorov, 2008), four traits loaded onto a *positivity* factor (attractive, caring, intelligent, and trustworthy), and three traits loaded onto a *dominance* factor (aggressive, dominant, and mean). Each item loaded onto its primary factor at $|\lambda| > .58$, with all but *mean* loading at $|\lambda| > .65$. All items loaded onto the other factor at $|\lambda| < .40$. Composite indices were generated for each factor with its primary loadings; higher scores reflect greater positivity ($\alpha = .90$) and dominance ($\alpha = .83$), respectively. Although the positivity and dominance indices were highly correlated ($r = -.77, p < .001$), the EFA suggests they should be treated as two distinct variables.

Figure A1

Factor Loadings From an EFA on Trait Ratings of Robert



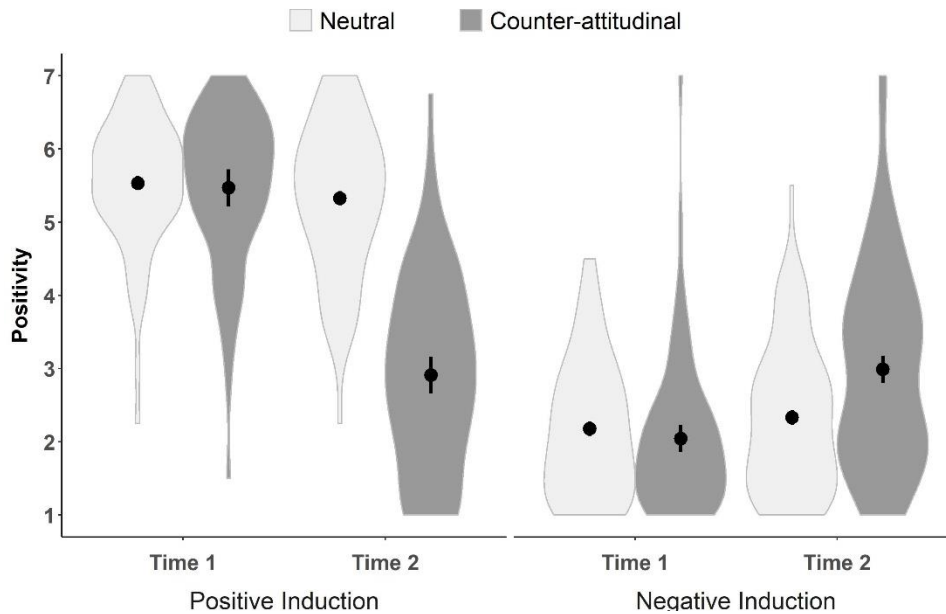
Notes. The *x*-axis depicts the positive loading strength for items on each factor. Factors are separated by panel and labeled by panel heading. Horizontal bars range from blue to red, with the greater appearance of blue representing higher positive load strength.

Positivity

A 2 (Time: 1 vs. 2; within-participants) \times 2 (Time 1 induction: positive vs. negative; between-participants) \times 2 (Time 2 information: control vs. counter-attitudinal; between-participants) mixed ANOVA on the positivity of the group images revealed a significant three-way interaction, $F(1, 560) = 73.20, p < .001, \eta_p^2 = .12$ (Figure A2). We decomposed this interaction by conducting separate Time \times Time 2 information ANOVAs in the positive-induction and negative-induction conditions.

Figure A2

Positivity Factor Impressions of Robert



Notes. Markers reflect mean positivity factor scores of Robert by Time, Time 1 induction, and Time 2 information in the image-generation experiment. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of the composite indices for the positivity factor (i.e., composite scores of trustworthy, intelligent, caring, and attractive) after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 280) = 83.77, p < .001, \eta_p^2 = .24$. Simple-effects tests indicated that learning about Robert’s child molestation conviction prompted negative revision (Time 1: $M = 5.47, SD = 1.07$; Time 2: $M = 2.91, SD = 1.30$), $t(72) = 14.37, p < .001, d = 1.68$. Learning neutral information also prompted negative revision (Time 1: $M = 5.53, SD = 0.90$; Time 2: $M = 5.33, SD = 1.04$), $t(69) = 3.04, p = .003, d = 0.36$; however, this effect was smaller than that in the counter-attitudinal condition.

Evidence of revision also emerged in the negative-induction condition—Time \times Time 2 information interaction, $F(1, 278) = 9.35, p = .002, \eta_p^2 = .03$. Learning about Robert’s kidney donation prompted positive revision (Time 1: $M = 2.04, SD = 1.04$; Time 2: $M = 2.99, SD =$

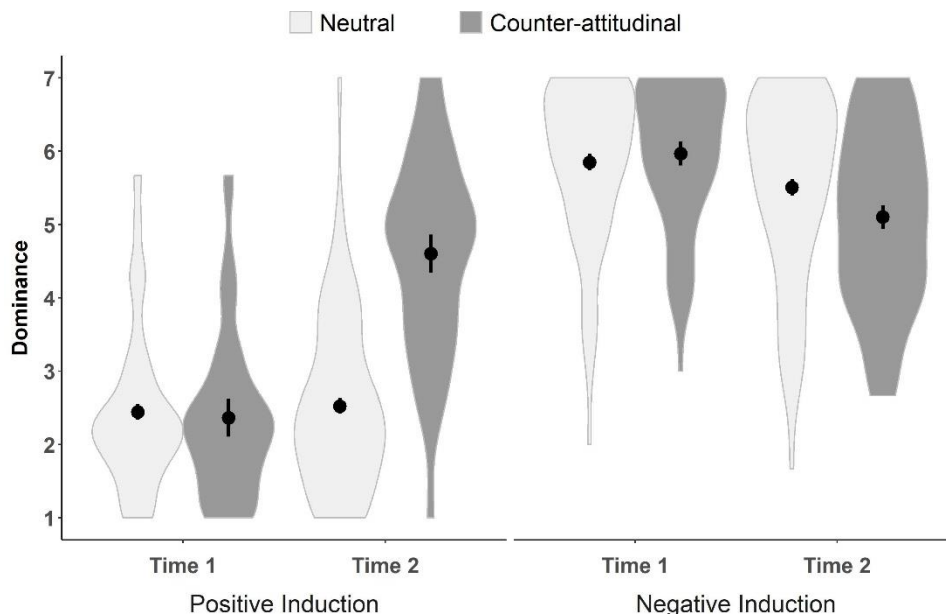
1.37), $t(68) = -7.22, p < .001, d = -0.87$. Learning neutral information also prompted positive revision (Time 1: $M = 2.18, SD = 0.93$; Time 2: $M = 2.33, SD = 1.02$), $t(72) = -2.16, p = .034, d = -0.25$; however, this effect was smaller than that in the counter-attitudinal condition.

Dominance

An identical 2 (Time) \times 2 (Time 1 condition) \times 2 (Time 2 condition) mixed ANOVA on the dominance of the group images also revealed a significant three-way interaction, $F(1, 560) = 46.38, p < .001, \eta_p^2 = .08$ (Figure A3). We again decomposed this interaction by conducting separate 2 (Time) \times 2 (Time 2 information) ANOVAs in the positive-induction and negative-induction conditions.

Figure A3

Dominance Factor Impressions of Robert



Notes. Marketers reflect mean dominance factor scores of Robert images by Time, Time 1 induction, and Time 2 information in the image-generation experiment. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of the composite indices for the dominance factor (i.e., composite scores of aggressive, dominant, and mean) after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 280) = 60.27, p < .001, \eta_p^2 = .18$. Learning about Robert’s child molestation conviction prompted revision in the more dominant direction (Time 1: $M = 2.37, SD = 1.10$; Time 2: $M = 4.60, SD = 1.35$), $t(154) = -12.72, p < .001, d = -1.43$, whereas learning neutral information did not (Time 1: $M = 2.44, SD = 1.09$; Time 2: $M = 2.52, SD = 1.18$), $t(69) = -1.07, p = .287, d = -0.13$.

Evidence of revision also emerged in the negative-induction condition; however, it was not restricted to the counter-attitudinal condition—Time \times Time 2 information interaction, $F(1, 278) = 3.78, p = .053, \eta_p^2 = .03$. Learning about Robert’s kidney donation prompted revision in the less dominant direction (Time 1: $M = 5.97, SD = 0.99$; Time 2: $M = 5.10, SD = 1.22$), $t(68) = 7.59, p < .001, d = 0.91$, as did learning neutral information (Time 1: $M = 5.85, SD = 1.11$; Time 2: $M = 5.51, SD = 1.28$), $t(72) = 4.32, p < .001, d = 0.51$.

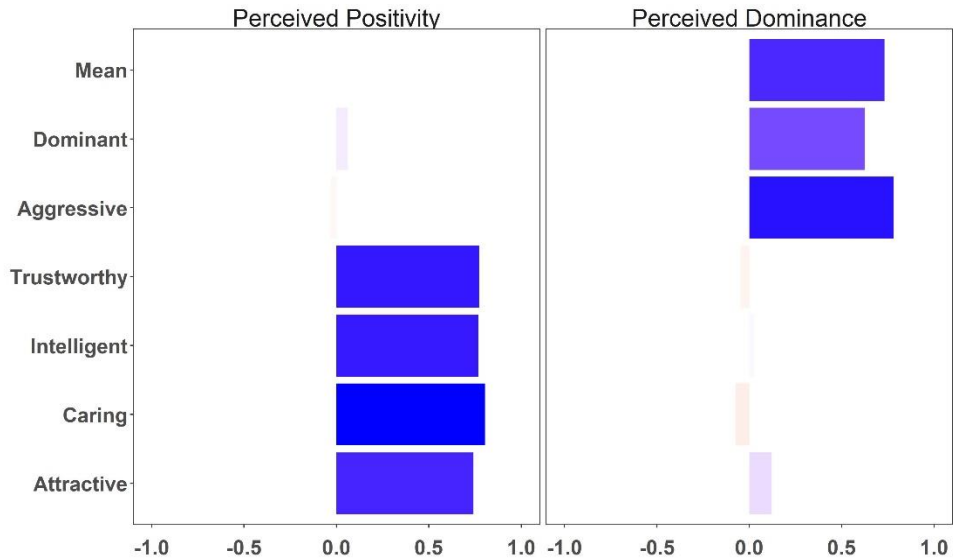
Image-Assessment Experiment 1

Data Reduction

An EFA with *promax* rotation again revealed a two-factor solution, $\chi^2(1240) = 4.69, p = .086$, that accounted for 56% of the total variance (Factor 1 = 34%, Factor 2 = 22%; Figure A4). Although a three-factor solution also fit the data, the two-factor solution made more theoretical sense. As before, four traits loaded onto a *positivity* factor (attractive, caring, intelligent, and trustworthy), and three traits loaded onto a *dominance* factor (aggressive, dominant, and mean). Each item loaded onto its primary factor at $\lambda > .60$ and the other factor at $\lambda < .15$. Composite indices were generated for each factor, with higher scores reflecting greater positivity ($\alpha = .85$) and dominance ($\alpha = .76$), respectively. Although the positivity and dominance indices were weakly correlated ($r = .09, p = .002$), the EFA suggested they should be treated separately.

Figure A4

Factor Loadings From an EFA on Trait Ratings of Group Classification Images



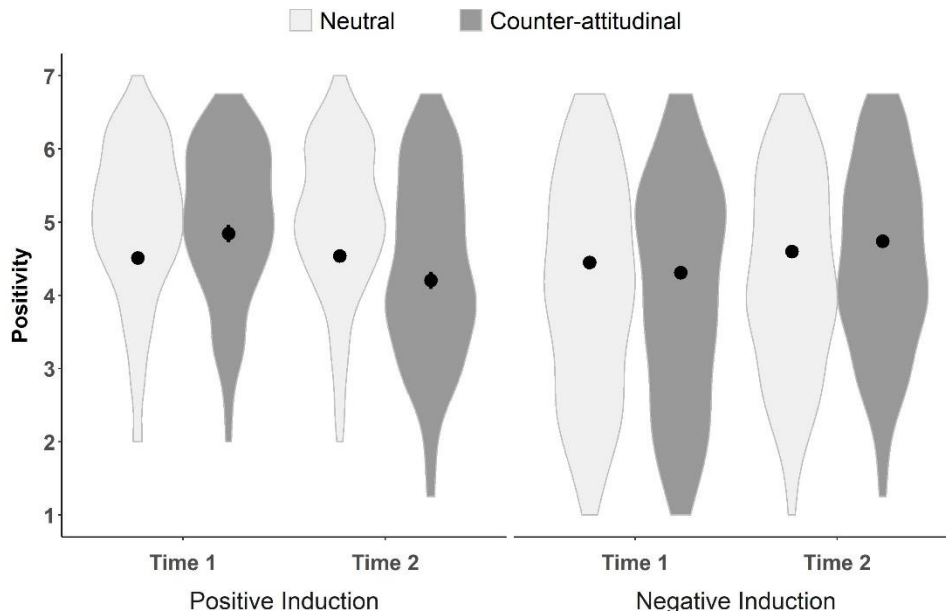
Notes. The *x*-axis depicts the positive loading strength for items on each factor. Factors are separated by panel and labeled by panel heading. Horizontal bars range from blue to red, with the greater appearance of blue representing higher positive load strength.

Positivity

A 2 (Time: 1 vs. 2) × 2 (Time 1 induction: positive vs. negative) × 2 (Time 2 information: control vs. counter-attitudinal) repeated-measures ANOVA on the positivity of the group images revealed a three-way interaction, $F(1, 154) = 58.62, p < .001, \eta_p^2 = .28$ (Figure A5). We decomposed this interaction by conducting separate Time × Time 2 information ANOVAs in the positive-induction and negative-induction conditions.

Figure A5

Positivity Factor Impressions of Group Classification Images



Notes. Markers reflect mean positivity factor scores of group classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 1. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of the composite indices for the positivity factor (i.e., composite scores of trustworthy, intelligent, caring, and attractive) after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 154) = 48.96, p < .001, \eta_p^2 = .24$. Simple-effects tests indicated that learning about Robert’s child molestation conviction prompted negative revision (Time 1: $M = 5.00, SD = 1.04$; Time 2: $M = 4.36, SD = 1.25$), $t(154) = 7.58, p < .001, d = 0.61$, whereas learning neutral information did not (Time 1: $M = 5.00, SD = 1.07$; Time 2: $M = 5.02, SD = 1.04$), $t(154) = -0.61, p = .546, d = -0.05$.

Evidence of revision also emerged in the negative-induction condition—Time \times Time 2 information interaction, $F(1, 154) = 13.39, p < .001, \eta_p^2 = .08$. Learning about Robert’s kidney donation prompted positive revision (Time 1: $M = 4.03, SD = 1.49$; Time 2: $M = 4.46, SD = 1.27$), $t(154) = -6.47, p < .001, d = -0.52$. Learning neutral information also prompted positive

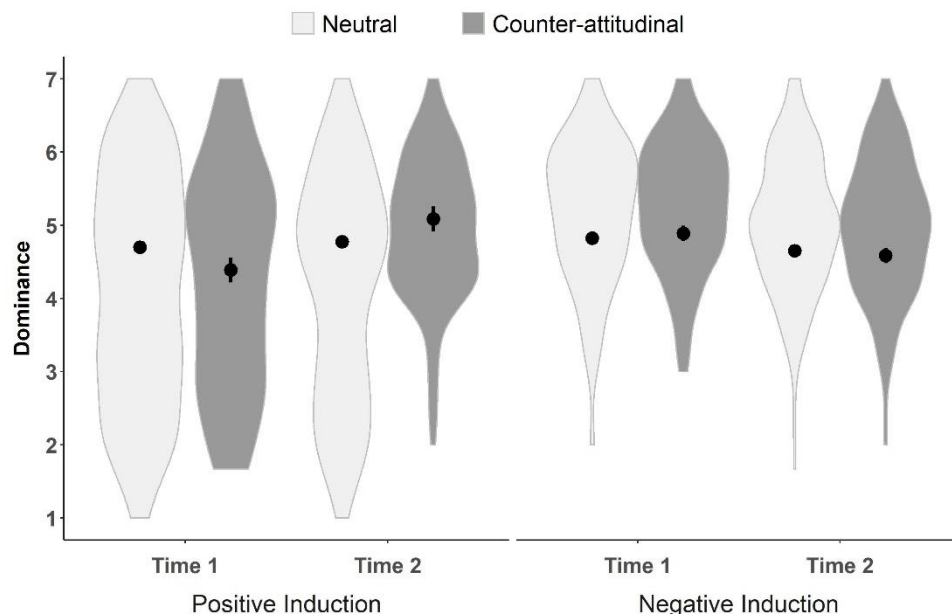
revision (Time 1: $M = 4.10$, $SD = 1.46$; Time 2: $M = 4.25$, $SD = 1.33$), $t(154) = -2.76$, $p = .007$, $d = -0.22$; however, this effect was smaller than that in the counter-attitudinal condition.

Dominance

A 2 (Time) \times 2 (Time 1 condition) \times 2 (Time 2 condition) repeated-measures ANOVA on the dominance of the group images also revealed a significant three-way interaction, $F(1, 154) = 24.55$, $p < .001$, $\eta_p^2 = .14$ (Figure A6). We again decomposed this interaction by conducting separate 2 (Time) \times 2 (Time 2 information) ANOVAs in the positive-induction and negative-induction conditions.

Figure A6

Dominance Factor Impressions of Group Classification Images



Notes. Markers reflect dominance factor scores of group classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 1. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of the composite indices for the dominance factor (i.e., composite scores of aggressive, dominant, and mean) after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 154) = 22.58$, $p < .001$, $\eta_p^2 = .13$. Learning about Robert’s child

molestation conviction prompted revision in the more dominant direction (Time 1: $M = 4.25$, $SD = 1.46$; Time 2: $M = 4.95$, $SD = 0.96$), $t(154) = -5.79$, $p < .001$, $d = -0.47$, whereas learning neutral information did not (Time 1: $M = 4.10$, $SD = 1.54$; Time 2: $M = 4.17$, $SD = 1.50$), $t(154) = -1.38$, $p = .169$, $d = -0.11$.

Evidence of revision also emerged in the negative-induction condition; however, it was not restricted to the counter-attitudinal condition—Time \times Time 2 information interaction, $F(1, 154) = 2.53$, $p = .114$, $\eta_p^2 = .02$. Learning about Robert’s kidney donation prompted revision in the less dominant direction (Time 1: $M = 5.26$, $SD = 0.87$; Time 2: $M = 4.96$, $SD = 0.92$), $t(154) = 4.03$, $p < .001$, $d = 0.32$, as did learning neutral information (Time 1: $M = 5.19$, $SD = 0.97$; Time 2: $M = 5.02$, $SD = 0.94$), $t(154) = 2.71$, $p = .008$, $d = 0.22$.

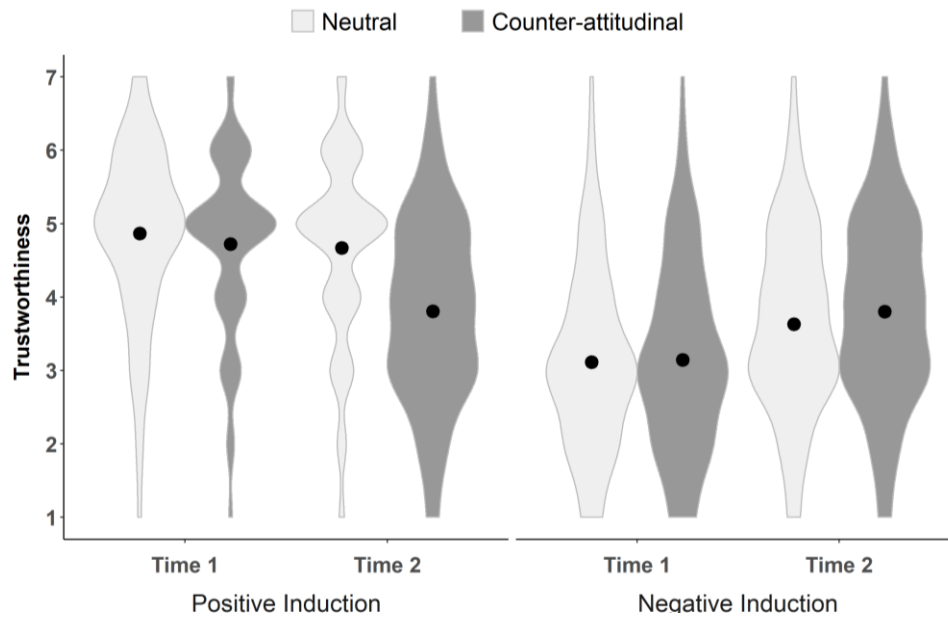
Image-Assessment Experiment 2a

A 2 (Time) \times 2 (Time 1 induction) \times 2 (Time 2 information) repeated-measures ANOVA on the trustworthiness impressions of the subgroup images revealed the expected three-way interaction, $F(1, 113) = 129.20$, $p < .001$, $\eta_p^2 = .53$ (Figure A7).²⁰ We decomposed this interaction by conducting separate 2 (Time) \times 2 (Time 2 information) ANOVAs in the positive-induction and negative-induction conditions.

²⁰An unexpected difference between Time 2 information conditions at Time 1 emerged in the positive-induction condition in Experiment 2a, $t(113) = -4.73$, $p < .001$, $d = -0.44$, and Experiment 2b, $t(241) = -5.75$, $p < .001$, $d = -0.37$. This difference did not emerge in the negative-induction condition in Experiment 2a, $t(113) = 0.96$, $p = .338$, $d = 0.09$, or Experiment 2b, $t(241) = 0.88$, $p = .382$, $d = 0.06$; nor did it emerge in either Time 1 induction condition in Experiment 1.

Figure A7

Trustworthiness Impressions of Subgroup Classification Images



Notes. Markers reflect mean trustworthiness ratings of subgroup classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2a. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of image raters' responses after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 113) = 158.64, p < .001, \eta_p^2 = .58$. Learning about Robert's child molestation conviction prompted negative revision (see Table A7 for descriptive statistics in Experiments 2a and 2b), $t(113) = 15.94, p < .001, d = 1.49$. Although learning neutral information about Robert also prompted negative revision, $t(113) = 6.02, p < .001, d = 0.56$, this effect was smaller than that in the counter-attitudinal condition.

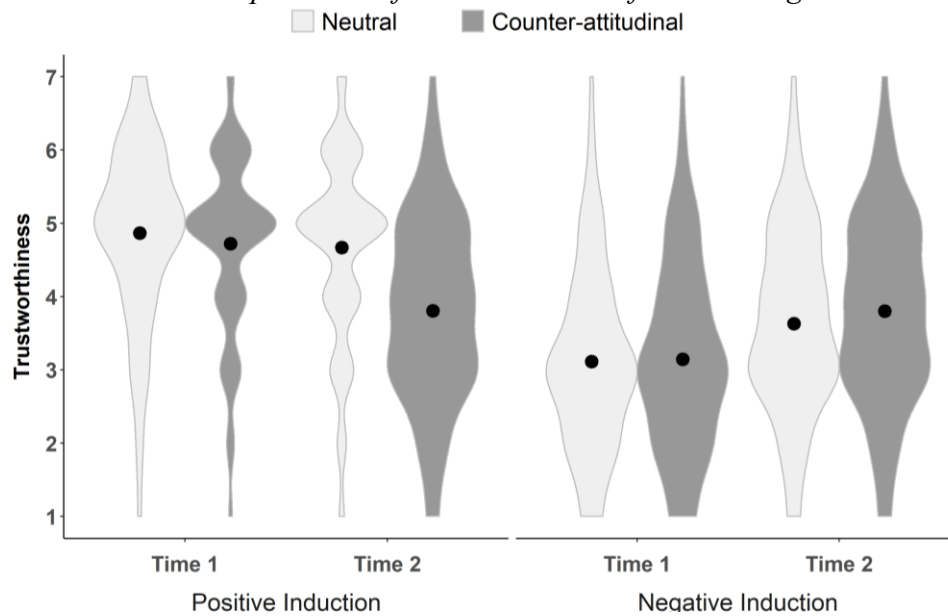
Revision was also evident in the negative-induction condition—Time \times Time 2 information interaction, $F(1, 113) = 9.40, p = .003, \eta_p^2 = .08$. Learning about Robert's kidney donation prompted positive revision, $t(113) = -14.20, p < .001, d = -1.33$. Learning neutral information also prompted positive revision, $t(113) = -13.31, p < .001, d = -1.25$, but again this effect was smaller than that in the counter-attitudinal condition.

Image-Assessment Experiment 2b

A 2 (Time) \times 2 (Time 1 induction) \times 2 (Time 2 information) repeated-measures ANOVA on the trustworthiness impressions of individual images revealed the expected three-way interaction, $F(1, 241) = 60.00, p < .001, \eta_p^2 = .19$ (Figure A8). We again decomposed this interaction by examining the underlying patterns separately in the positive-induction and negative-induction conditions.

Figure A8

Trustworthiness Impressions of Individual Classification Images



Notes. Markers reflect trustworthiness ratings of individual classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2b. Error bars represent 95% confidence intervals. The surrounding violin plots are mirrored density distributions of image raters' responses after a smoothing function was applied.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 241) = 56.72, p < .001, \eta_p^2 = .19$. Learning about Robert's child molestation conviction prompted negative revision, $t(241) = 16.17, p < .001, d = 1.04$. Learning

neutral information also prompted negative revision, $t(241) = 9.31, p < .001, d = 0.60$; however, this effect was smaller than that in the counter-attitudinal condition.

Revision was also evident in the negative-induction condition—Time \times Time 2 information interaction, $F(1, 241) = 11.68, p < .001, \eta_p^2 = .05$. Learning about Robert's kidney donation prompted positive revision, $t(241) = -12.55, p < .001, d = -0.81$. Learning neutral information also prompted positive revision, $t(241) = -8.75, p < .001, d = -0.56$, but again this effect was smaller than that in the counter-attitudinal condition.

Descriptive Statistics Tables

Table A5

Descriptive Statistics for Trait Ratings of Robert (Image-Generation Experiment)

<i>Positive-induction Condition</i>				
Trait	Neutral		Counter-attitudinal	
	Time 1	Time 2	Time 1	Time 2
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Aggressive	1.93 (1.47)	2.21 (1.42)	1.74 (1.11)	4.84 (1.68)
Attractive	4.27 (1.46)	4.10 (1.58)	4.42 (1.59)	2.64 (1.66)
Caring	6.39 (1.09)	6.17 (1.26)	6.30 (1.14)	2.95 (1.50)
Dominant	3.83 (1.51)	3.43 (1.56)	3.53 (1.60)	4.63 (1.74)
Intelligent	5.49 (1.13)	5.31 (1.20)	5.26 (1.35)	3.85 (1.66)
Mean	1.57 (1.04)	1.93 (1.24)	1.82 (1.38)	4.34 (1.91)
Trustworthy	5.99 (1.19)	5.71 (1.34)	5.89 (1.37)	2.21 (1.56)

<i>Negative-induction Condition</i>				
Trait	Neutral		Counter-attitudinal	
	Time 1	Time 2	Time 1	Time 2
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Aggressive	6.03 (1.38)	5.74 (1.55)	6.28 (1.21)	5.33 (1.56)
Attractive	2.14 (1.28)	2.34 (1.45)	2.26 (1.54)	2.75 (1.71)
Caring	1.67 (1.08)	1.85 (1.05)	1.51 (1.13)	3.26 (1.63)
Dominant	5.37 (1.66)	5.04 (1.80)	5.39 (1.61)	5.09 (1.56)
Intelligent	3.18 (1.41)	3.18 (1.47)	2.84 (1.45)	3.35 (1.60)
Mean	6.15 (1.43)	5.74 (1.61)	6.23 (1.15)	4.88 (1.71)
Trustworthy	1.73 (1.15)	1.96 (1.26)	1.57 (1.22)	2.59 (1.68)

Table A6

Descriptive Statistics for Trait Ratings of Group Classification Images (Image-Assessment Experiment 1)

<i>Positive-induction Condition</i>				
Trait	Neutral		Counter-attitudinal	
	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>
Aggressive	3.97 (1.81)	4.09 (1.77)	4.18 (1.78)	4.97 (1.21)
Attractive	4.62 (1.52)	4.68 (1.57)	4.86 (1.47)	4.38 (1.62)
Caring	5.32 (1.27)	5.28 (1.38)	5.19 (1.31)	4.29 (1.53)
Dominant	4.23 (1.56)	4.33 (1.64)	4.40 (1.64)	4.90 (1.33)
Intelligent	5.07 (1.36)	5.06 (1.25)	4.97 (1.31)	4.50 (1.39)
Mean	4.09 (1.90)	4.10 (1.83)	4.17 (1.73)	4.98 (1.36)
Trustworthy	4.97 (1.33)	5.07 (1.30)	4.95 (1.26)	4.25 (1.54)

<i>Negative-induction Condition</i>				
Trait	Neutral		Counter-attitudinal	
	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>
Aggressive	5.29 (1.21)	5.14 (1.19)	5.30 (1.17)	4.90 (1.25)
Attractive	4.02 (1.76)	4.26 (1.65)	3.84 (1.71)	4.34 (1.53)
Caring	4.10 (1.86)	4.14 (1.67)	3.93 (1.81)	4.47 (1.62)
Dominant	5.06 (1.35)	5.00 (1.32)	5.28 (1.27)	5.05 (1.34)
Intelligent	4.25 (1.54)	4.41 (1.44)	4.28 (1.54)	4.66 (1.37)
Mean	5.21 (1.27)	4.91 (1.34)	5.21 (1.24)	4.93 (1.30)
Trustworthy	4.02 (1.68)	4.19 (1.62)	4.06 (1.69)	4.36 (1.65)

Table A7

Descriptive Statistics for Trustworthiness Impressions of Subgroup Classification Images (Image-Assessment 2a) and Individual Classification Images (Image-Assessment Experiment 2b)

Time 1 Induction	Time 2 Information			
	Neutral		Counter-attitudinal	
	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>
<i>Experiment 2a</i>				
Positive	4.87 (0.92)	4.67 (0.95)	4.72 (0.88)	3.80 (0.75)
Negative	3.11 (0.85)	3.63 (0.79)	3.15 (0.81)	3.80 (0.78)
<i>Experiment 2b</i>				
Positive	4.16 (0.81)	3.98 (0.76)	4.05 (0.76)	3.65 (0.76)
Negative	3.41 (0.78)	3.58 (0.75)	3.42 (0.79)	3.69 (0.77)

Additional Dependent Measures

Feeling Thermometer

Alongside the seven traits on which participants in the image-generation experiment rated Robert (discussed in the main text), they also indicated their feelings about him by entering a number between 0 and 100. Higher numbers reflect warmer impressions.

A mixed ANOVA on the feeling thermometer revealed a three-way interaction, $F(1, 278) = 149.31, p < .001, \eta_p^2 = .35$.²¹ We decomposed this interaction by conducting separate 2 (Time) \times 2 (Time 2 information) mixed ANOVAs in the positive-induction and negative-induction conditions.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $F(1, 139) = 126.50, p < .001, \eta_p^2 = .48$. Learning about Robert's child molestation conviction prompted negative revision (Time 1: $M = 82.22, SD = 19.30$; Time 2: $M = 27.53, SD = 27.57$), $t(71) = 14.46, p < .001, d = 1.70$. Learning neutral information also prompted negative revision (Time 1: $M = 82.38, SD = 21.68$; Time 2: $M = 76.71, SD = 20.64$), $t(68) = 2.62, p = .011, d = 0.32$; however, this effect was smaller than that in the counter-attitudinal condition.

Evidence of revision also emerged in the negative-induction condition—Time \times Time 2 information interaction, $F(1, 139) = 25.00, p < .001, \eta_p^2 = .15$. Learning about Robert's kidney donation prompted positive revision (Time 1: $M = 18.54, SD = 22.59$; Time 2: $M = 33.84, SD = 25.74$), $t(67) = -6.66, p < .001, d = -0.81$, but learning neutral information did not (Time 1: $M = 22.78, SD = 19.13$; Time 2: $M = 24.96, SD = 20.01$), $t(72) = -1.60, p = .113, d = -0.19$.

²¹Either due to technical errors in data collection or blank responses on the feeling thermometer measure, we excluded three participants' data, leaving a final sample of 282 participants for these analyses.

Race/Ethnicity Categorizations

In the image-generation experiment, participants also made race/ethnicity categorizations of Robert at Time 1 and Time 2. They selected from six response options the one that best matched their impression of Robert's race/ethnicity: Asian, Black, Latinx, Native American, White, or other. 'Other' responses account for approximately 3% of total race/ethnicity categorizations. Given that all but 2 of those responses were irrelevant, ambiguous, or left uncategorized, 'other' responses were excluded from the analyses below.

We conducted generalized linear model (GLM) analyses with logit link functions on the race/ethnicity categorizations. Time (1 vs. 2), Time 1 induction (positive vs. negative), Time 2 information (neutral vs. counter-attitudinal), and all interactions were included as predictors. To account for repeated-measures (Time) in the model, we also included random intercepts for each participant. Due to low frequencies in Latinx and Native American categorizations (frequencies <10 for both), we do not include analyses for these racial/ethnic categorizations.

Neither Asian nor Black categorizations produced significant three-way interactions ($Bs < -2.00$, $ps > .136$; Table A8).²² White categorizations, however, revealed a significant three-way interaction, $B = 0.33$, $SE = 0.13$, $z = 2.46$, $p = .014$. We decomposed this interaction by conducting separate 2 (Time) \times 2 (Time 2) analyses in the two induction conditions.

Evidence of revision emerged in the positive-induction condition—Time \times Time 2 information interaction, $B = -0.38$, $SE = 0.17$, $z = -2.26$, $p = .024$. Whereas the odds of White categorization were not revised after learning about Robert's child molestation conviction, $OR = 0.74$, $p = .503$, they did decrease after learning neutral information, $OR = 3.37$, $p = .012$. No

²²Either due to technical errors in data collection, blank responses, or other categorizations, we excluded 31 participants' data, leaving a final sample of 254 participants for these analyses.

significant evidence of revision emerged for White categorizations in the negative-induction condition—Time \times Time 2 information interaction, $B = 0.21$, $SE = 0.21$, $z = 0.99$, $p = .321$.

Table A8

Descriptive Statistics (Proportions) for Race/Ethnicity Categorizations (Image-Generation Experiment)

<i>Positive-induction Condition</i>				
Race/Ethnicity	Neutral		Counter-attitudinal	
	Time 1	Time 2	Time 1	Time 2
Asian	0.02	0.00	0.05	0.02
Black	0.38	0.60	0.49	0.48
Latinx	0.11	0.09	0.08	0.06
Native American	0.02	0.02	0.00	0.02
White	0.48	0.29	0.38	0.43

<i>Negative-induction Condition</i>				
Race/Ethnicity	Neutral		Counter-attitudinal	
	Time 1	Time 2	Time 1	Time 2
Asian	0.05	0.00	0.00	0.02
Black	0.23	0.36	0.35	0.46
Latinx	0.11	0.07	0.14	0.09
Native American	0.03	0.00	0.02	0.02
White	0.57	0.57	0.49	0.42

Race/Ethnicity Prototypicality Ratings

In image-assessment Experiment 1, participants also rated the group images of Robert on how prototypical they were for each of four races/ethnicities: Asian, Black, Latinx, and White (0 = *not at all prototypical*, 100 = *extremely prototypical*).

We conducted a 2 (Time: 1 vs. 2) \times 2 (Time 1 induction: positive vs. negative) \times 2 (Time 2 information: neutral vs. counter-attitudinal) ANOVA on the prototypicality ratings separately for each race/ethnicity. These analyses revealed a three-way interaction only for ratings of Black

prototypicality, $F(1, 154) = 19.52, p < .001, \eta_p^2 = .11$. Ratings of Asian, Latinx, and White prototypicality revealed no significant three-way interactions ($F_s < 3.43, p_s > .07, \eta_p^2_s < .02$). We decomposed the interaction for Black prototypicality by inspecting the underlying patterns separately in the positive-induction and negative-induction conditions (see Table A9).

Revision was evident in the positive-induction condition—Time \times Time 2 information interaction—for ratings of Black prototypicality. Visualizations of Robert were less prototypically Black after learning about his child molestation conviction, but not after learning neutral information. This pattern of revision did not emerge in the negative-induction condition.

Table A9

Descriptive Statistics and Univariate ANOVA Results for Race/Ethnicity Prototypicality Ratings of Group Images (Image-Assessment Experiment 1)

<i>Positive-induction Condition</i>									
Race/ethnicity	Neutral			Counter-attitudinal			F values		
	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	<i>d</i>	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	<i>d</i>	Time	Time 2 Info	Time × Time 2 Info
Asian	51.91 (32.42)	50.65 (32.49)	0.08	51.85 (32.06)	49.15 (32.93)	0.15	3.61	0.59	0.62
Black	67.17 (26.87)	67.55 (27.54)	0.03	68.97 (25.06)	56.23 (30.30)	0.43	21.24***	12.45***	24.74***
Latinx	58.45 (25.37)	59.03 (27.28)	0.04	59.27 (25.77)	60.75 (26.96)	0.07	1.08	1.62	0.18
White	56.09 (29.61)	56.34 (29.53)	0.01	58.32 (29.00)	67.23 (23.70)	0.29	11.58***	19.44***	7.96**

<i>Negative-induction Condition</i>									
Race/ethnicity	Neutral			Counter-attitudinal			F values		
	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	<i>d</i>	Time 1 <i>M (SD)</i>	Time 2 <i>M (SD)</i>	<i>d</i>	Time	Time 2 Info	Time × Time 2 Info
Asian	47.89 (33.99)	48.20 (33.85)	0.02	47.33 (33.79)	49.77 (33.70)	0.16	2.80	0.36	1.48
Black	59.84 (28.60)	53.75 (31.68)	0.27	60.01 (26.75)	54.05 (32.03)	0.25	12.75***	0.05	0.01
Latinx	57.90 (26.98)	58.66 (27.48)	0.04	59.95 (26.78)	57.48 (29.76)	0.13	0.67	0.20	2.45
White	65.53 (25.36)	71.13 (23.67)	0.25	63.10 (25.67)	70.49 (23.04)	0.32	19.27***	2.03	0.72

Note. Info = Information. * $p < .05$ ** $p < .01$ *** $p < .001$

Appendix B

Chapter 2: Emotion Expression Saliency and Racially Biased Weapon Identification:

A Diffusion Modeling Approach

Weapon Identification Task

Results (Incorrect Response Times)

In the main text, we reported analyses of response times (RTs) only on trials with correct responses, as is common in research using the Weapon Identification Task (see Rivers, 2017). For completeness, we report RT analyses on trials with incorrect responses. Because error rates were low overall, these results should be interpreted cautiously.

Experiment 1

The Race Prime \times Target Object interaction indicative of racial bias was not significant, $\beta = 7.79$, $F(1, 489.6) = 1.40$, $p = .237$, $R^2 < .01$. Nor was this interaction moderated by Saliency, $\beta = -2.48$, $F(1, 7489.8) = 0.04$, $p = .845$, $R^2 < .01$.

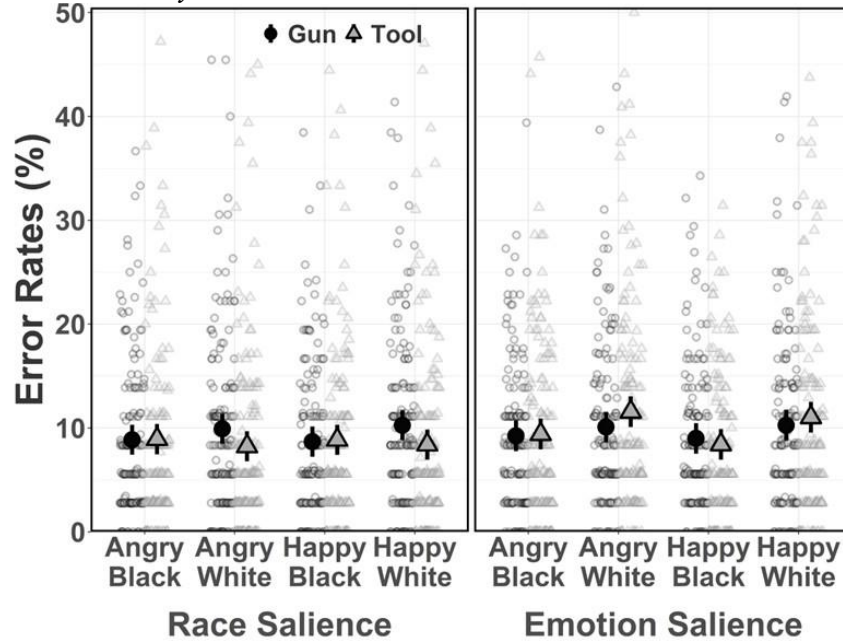
Experiment 2

The model had to be re-fit after removing the by-stimulus random intercept from the model, as boundary fit emerged. The Race Prime \times Target Object interaction was significant, $\beta = 23.49$, $F(1, 7354.2) = 11.56$, $p = .001$, $R^2 = .01$, but it was not moderated by Saliency, $\beta = -26.45$, $F(1, 7354.2) = 3.66$, $p = .056$, $R^2 = .01$.

Behavioral Data Plots

Figure B1

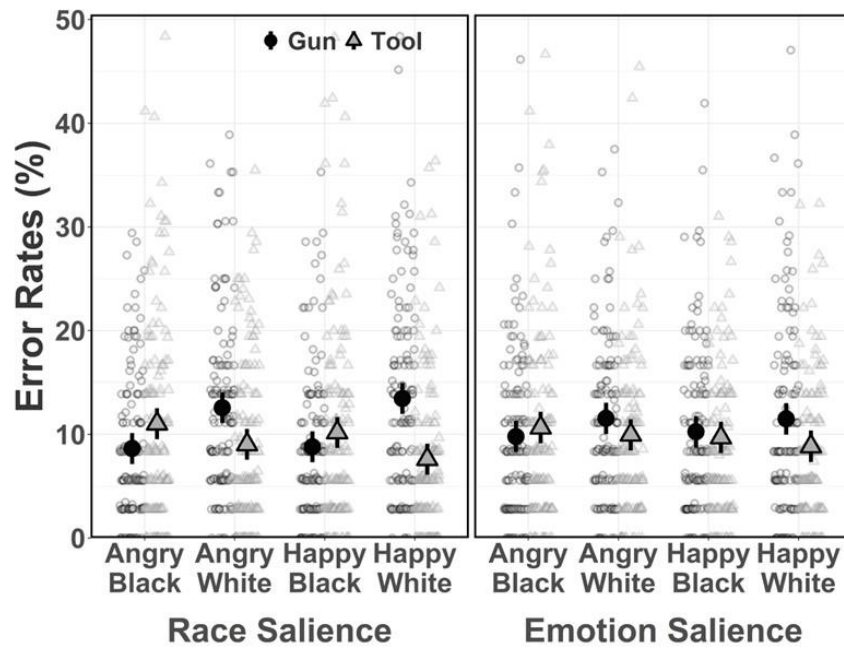
Error Rates by Race Prime, Emotion Prime, and Salience Conditions (Experiment 1)



Notes. Empty markers reflect individual-level error rates and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to error rates in Experiment 1.

Figure B2

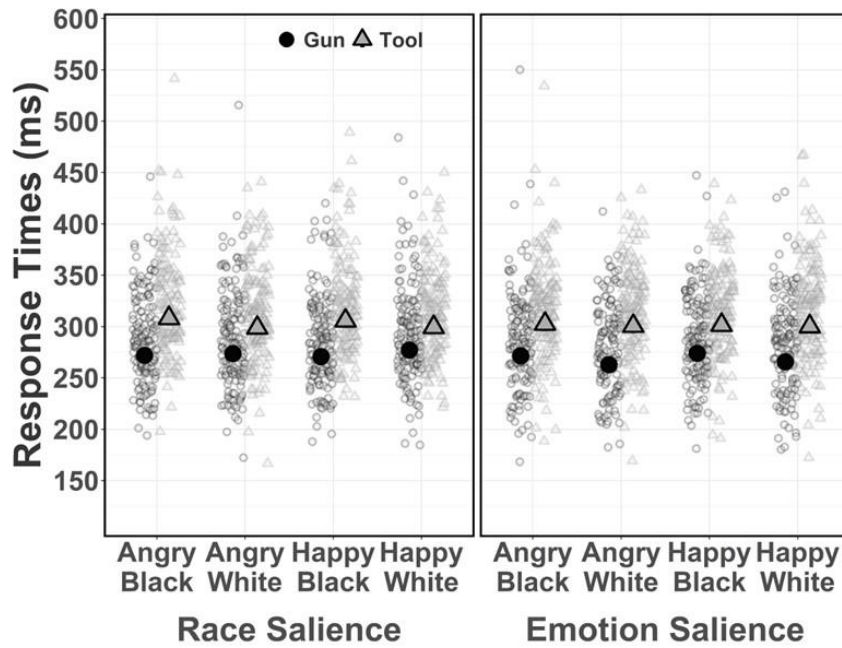
Error Rates by Race Prime, Emotion Prime, and Salience Conditions (Experiment 2)



Notes. Empty markers reflect individual-level error rates and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to error rates in Experiment 2.

Figure B3

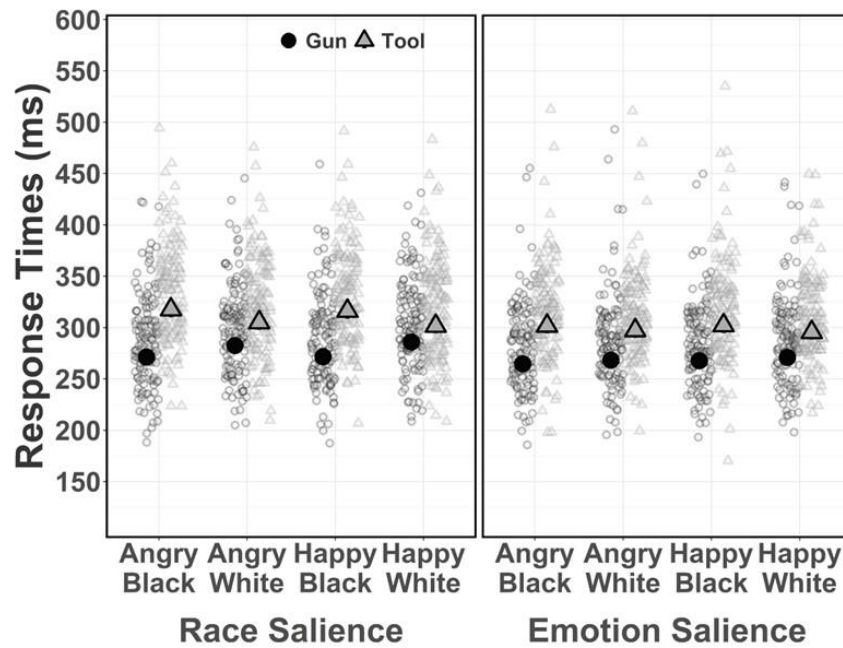
Correct RTs by Race Prime, Emotion Prime, and Salience Conditions (Experiment 1)



Notes. RTs = response times. Empty markers reflect individual-level correct RTs and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to correct RTs in Experiment 1.

Figure B4

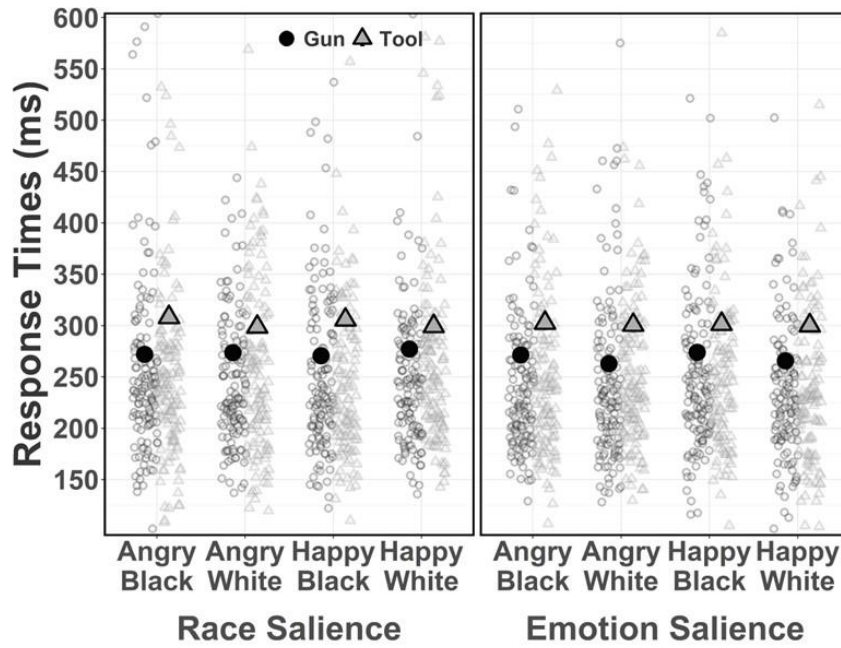
Correct RTs by Race Prime, Emotion Prime, and Salience Conditions (Experiment 2)



Notes. RTs = response times. Empty markers reflect individual-level correct RTs and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to correct RTs in Experiment 2.

Figure B5

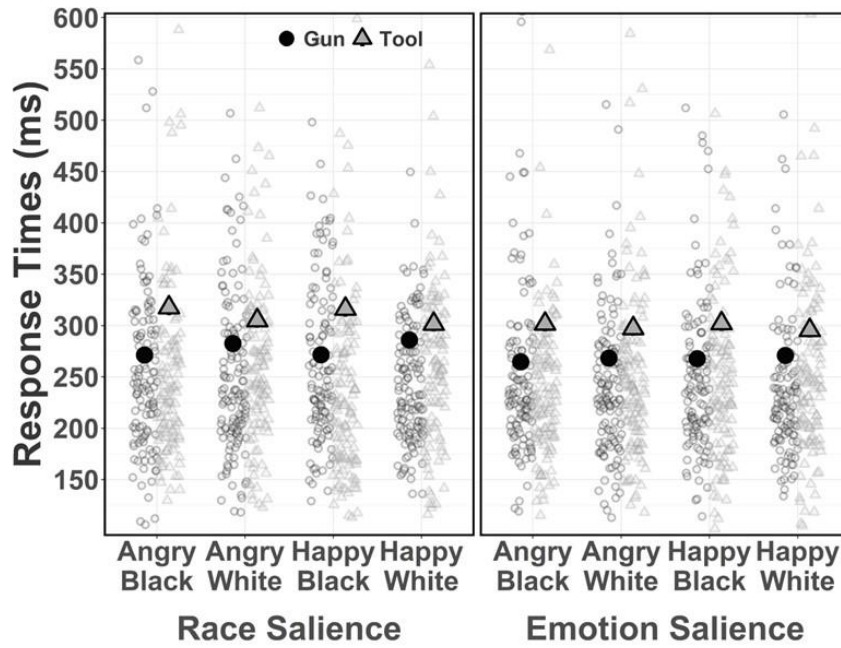
Incorrect RTs by Race Prime, Emotion Prime, and Salience Conditions (Experiment 1)



Notes. RTs = response times. Empty markers reflect individual-level incorrect RTs and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to incorrect RTs in Experiment 1.

Figure B6

Incorrect RTs by Race Prime, Emotion Prime, and Salience Conditions (Experiment 2)



Notes. RTs = response times. Empty markers reflect individual-level incorrect RTs and filled markers and their error bars reflect estimated marginal means and confidence intervals, respectively, from the LMEM modeled to incorrect RTs in Experiment 2.

Behavioral Data Tables

Table B1

LMEM of Error Rates and Correct Response Times (Experiment 1)

Effect	β	SE	df	t	p
<i>Error rates</i>					
(Intercept)	.09	< .01	343.49	23.11	< .001
Race Prime	.01	< .01	563.22	3.36	.001
Target Object	< .01	< .01	563.08	-0.6	.551
Saliency	.01	.01	289.97	1.1	.274
Emotion Prime	< .01	< .01	563.27	-0.54	.587
Race Prime \times Target Object	< .01	.01	563.08	-0.46	.645
Race Prime \times Saliency	.01	< .01	82169.63	3.45	.001
Target Object \times Saliency	.01	< .01	82167.17	3.28	.001
Race Prime \times Emotion Prime	< .01	.01	563.08	0.66	.512
Target Object \times Emotion Prime	< .01	.01	563.08	-0.62	.537
Saliency \times Emotion Prime	< .01	< .01	82172.11	-1.18	.236
Race Prime \times Target Object \times Saliency	.03	.01	82166.56	4.11	< .001
Race Prime \times Target Object \times Emotion Prime	< .01	.01	563.09	-0.11	.915
Race Prime \times Saliency \times Emotion Prime	< .01	.01	82166.54	0.09	.932
Target Object \times Saliency \times Emotion Prime	-.01	.01	82166.63	-0.89	.373
Race Prime \times Target Object \times Saliency \times Emotion Prime	< .01	.02	82166.72	0.22	.826
<i>Correct response times</i>					
(Intercept)	5.66	.01	311.86	742.22	< .001
Race Prime	-0.01	< .01	553.93	-3.48	.001
Target Object	0.11	< .01	554.02	31.7	< .001
Saliency	-0.01	.01	292.36	-0.8	.424
Emotion Prime	< 0.01	< .01	553.98	0.6	.552
Race Prime \times Target Object	-0.01	.01	553.66	-1.1	.271
Race Prime \times Saliency	-0.01	< .01	74529.95	-3.13	.002
Target Object \times Saliency	0.01	< .01	74533.29	2.9	.004
Race Prime \times Emotion Prime	0.01	.01	553.56	1.14	.256
Target Object \times Emotion Prime	-0.01	.01	553.62	-1.38	.167
Saliency \times Emotion Prime	< 0.01	< .01	74533.57	0.76	.444
Race Prime \times Target Object \times Saliency	0.07	.01	74529.96	7.99	< .001
Race Prime \times Target Object \times Emotion Prime	< 0.01	.01	553.6	-0.25	.807
Race Prime \times Saliency \times Emotion Prime	-0.01	.01	74528.38	-1.42	.156
Target Object \times Saliency \times Emotion Prime	-0.01	.01	74528.64	-0.68	.499
Race Prime \times Target Object \times Saliency \times Emotion Prime	0.01	.02	74529.22	0.46	.644

Table B2*LMEM of Incorrect Response Times (Experiment 1)*

Effect	β	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
(Intercept)	252.35	2.88	304.07	87.64	< .001
Race Prime	3.89	3.28	483.49	1.19	.237
Target Object	6.28	3.34	518.89	1.88	.061
Saliency	-8.09	5.69	311.99	-1.42	.156
Emotion Prime	-0.97	3.3	491.74	-0.29	.770
Race Prime \times Target Object	7.79	6.58	489.58	1.18	.237
Race Prime \times Saliency	-3.47	6.35	7482.64	-0.55	.584
Target Object \times Saliency	0.86	6.47	7574.77	0.13	.895
Race Prime \times Emotion Prime	-0.33	6.55	478.56	-0.05	.960
Target Object \times Emotion Prime	7.18	6.58	488.52	1.09	.275
Saliency \times Emotion Prime	-1.66	6.37	7513.26	-0.26	.795
Race Prime \times Target Object \times Saliency	-2.48	12.72	7489.82	-0.2	.845
Race Prime \times Target Object \times Emotion Prime	2.30	13.13	483.35	0.18	.861
Race Prime \times Saliency \times Emotion Prime	-24.08	12.66	7455.82	-1.9	.057
Target Object \times Saliency \times Emotion Prime	2.36	12.71	7488.3	0.19	.853
Race Prime \times Target Object \times Saliency \times Emotion Prime	4.07	25.38	7480.25	0.16	.873

Table B3*LMEM of Error Rates and Correct Response Times (Experiment 2)*

Effect	β	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
<i>Error rates</i>					
(Intercept)	0.1	< .01	294.32	24.14	< .001
Race Prime	0.01	< .01	565.36	2.27	.024
Target Object	-0.01	< .01	565.35	-3.93	< .001
Saliency	< .01	0.01	259.99	0.14	.889
Emotion Prime	< .01	< .01	565.36	-1.19	.235
Race Prime × Target Object	-0.04	0.01	565.36	-7.31	< .001
Race Prime × Saliency	-0.01	< .01	74213.34	-1.5	.135
Target Object × Saliency	< .01	< .01	74212.81	0.99	.323
Race Prime × Emotion Prime	< .01	0.01	565.37	-0.22	.825
Target Object × Emotion Prime	-0.01	0.01	565.35	-2.4	.017
Saliency × Emotion Prime	< .01	< .01	74212.87	-0.24	.808
Race Prime × Target Object × Saliency	0.04	0.01	74212.84	5.03	< .001
Race Prime × Target Object × Emotion Prime	< .01	0.01	565.36	-0.4	.693
Race Prime × Saliency × Emotion Prime	< .01	0.01	74213.48	-0.47	.639
Target Object × Saliency × Emotion Prime	< .01	0.01	74212.81	0.5	.618
Race Prime × Target Object × Saliency × Emotion Prime	0.02	0.02	74213.49	0.97	.334
<i>Correct response times</i>					
(Intercept)	5.66	.01	274.21	705.40	< .001
Race Prime	< .01	< .01	551.44	-0.32	.747
Target Object	0.11	< .01	552.09	34.99	< .001
Saliency	-0.04	.02	264.14	-2.27	.024
Emotion Prime	< .01	< .01	551.42	0.52	.601
Race Prime × Target Object	-0.06	.01	551.73	-9.59	< .001
Race Prime × Saliency	< .01	< .01	66766.01	-0.92	.359
Target Object × Saliency	< .01	< .01	66769.88	0.15	.880
Race Prime × Emotion Prime	< .01	.01	551.47	-0.22	.829
Target Object × Emotion Prime	-0.01	.01	551.49	-2.21	.028
Saliency × Emotion Prime	< .01	< .01	66767.17	1.07	.283
Race Prime × Target Object × Saliency	0.06	.01	66767.76	6.52	< .001
Race Prime × Target Object × Emotion Prime	-0.01	.01	551.43	-0.96	.336
Race Prime × Saliency × Emotion Prime	-0.01	.01	66766.64	-0.73	.465
Target Object × Saliency × Emotion Prime	< .01	.01	66767.15	0.17	.865
Race Prime × Target Object × Saliency × Emotion Prime	0.01	.02	66767.39	0.73	.468

Table B4*LMEM of Incorrect Response Times (Experiment 2)*

Effect	β	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
(Intercept)	256.42	3.29	257.01	78.00	< .001
Race Prime	-2.41	3.43	7305.24	-0.70	.482
Target Object	8.22	3.50	7417.28	2.35	.019
Saliency	-10.13	6.57	257.01	-1.54	.125
Emotion Prime	-0.92	3.42	7291.95	-0.27	.788
Race Prime \times Target Object	23.49	6.91	7354.22	3.40	.001
Race Prime \times Saliency	-10.12	6.86	7305.24	-1.48	.140
Target Object \times Saliency	5.93	7.01	7417.28	0.85	.398
Race Prime \times Emotion Prime	-0.91	6.87	7314.43	-0.13	.894
Target Object \times Emotion Prime	0.34	6.87	7321.63	0.05	.961
Saliency \times Emotion Prime	12.72	6.85	7291.95	1.86	.063
Race Prime \times Target Object \times Saliency	-26.45	13.82	7354.22	-1.91	.056
Race Prime \times Target Object \times Emotion Prime	14.41	13.72	7304.46	1.05	.293
Race Prime \times Saliency \times Emotion Prime	0.60	13.73	7314.43	0.04	.965
Target Object \times Saliency \times Emotion Prime	24.32	13.74	7321.63	1.77	.077
Race Prime \times Target Object \times Saliency \times Emotion Prime	4.41	27.43	7304.46	0.16	.872

Table B5*Race Prime Contrasts (Black – White) Across Conditions of Target Object and Saliency*

Effect	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Experiment 1				
<i>Error Rates</i>				
Gun Trial, Race Saliency	-0.01	.01	-2.55	.053
Tool Trial, Race Saliency	0.01	.01	1.13	.674
Gun Trial, Emotion Saliency	-0.01	.01	-2.02	.179
Tool Trial, Emotion Saliency	-0.02	.01	-4.55	< .001
<i>Correct Response Times</i>				
Gun Trial, Race Saliency	-0.01	.01	-2.65	.041
Tool Trial, Race Saliency	0.03	.01	4.59	< .001
Gun Trial, Emotion Saliency	0.03	.01	5.47	< .001
Tool Trial, Emotion Saliency	0.01	.01	0.99	.757
Experiment 2				
<i>Error Rates</i>				
Gun Trial, Race Saliency	-0.04	.01	-8.25	< .001
Tool Trial, Race Saliency	0.02	.01	4.39	< .001
Gun Trial, Emotion Saliency	-0.02	.01	-2.27	.023
Tool Trial, Emotion Saliency	.01	.01	0.52	.463
<i>Correct Response Times</i>				
Gun Trial, Race Saliency	< .01	.01	0.15	< .001
Tool Trial, Race Saliency	< .01	.01	-0.22	< .001
Gun Trial, Emotion Saliency	-0.01	.01	-2.21	.109
Tool Trial, Emotion Saliency	0.01	.01	0.73	.004

Note. Contrasts were computed to compare the estimated marginal mean error rates or correct response times following Black primes minus the estimated marginal mean error rates or correct response times following White primes.

Table B6*Descriptive Statistics by Race Prime, Emotion Prime, and Salience Conditions (Experiments 1 & 2).*

Variable	Emotion prime, race prime, and target object							
	Angry prime				Happy prime			
	Black prime		White prime		Black prime		White prime	
	Gun	Tool	Gun	Tool	Gun	Tools	Gun	Tool
Experiment 1								
Error rate (%)								
Race-salient	8.6 (28.1)	8.7 (28.2)	9.7 (29.6)	8.0 (27.2)	8.3 (27.6)	8.6 (28.0)	10.0 (30.0)	8.2 (27.4)
Emotion-salient	9.1 (28.7)	9.3 (29.0)	9.8 (29.8)	11.4 (31.7)	8.8 (28.4)	8.2 (27.4)	10.0 (30.0)	10.8 (31.0)
Correct RT (ms)								
Race-salient	288 (103)	323 (108)	289 (102)	314 (102)	286 (99)	320 (104)	293 (105)	314 (103)
Emotion-salient	288 (106)	318 (105)	279 (104)	317 (110)	290 (103)	316 (101)	283 (107)	317 (110)
Incorrect RT (ms)								
Race-salient	254 (140)	253 (145)	251 (137)	260 (156)	244 (123)	250 (138)	251 (152)	265 (145)
Emotion-salient	240 (128)	246 (152)	249 (151)	255 (160)	243 (123)	255 (130)	236 (117)	252 (159)
Experiment 2								
Error rate (%)								
Race-salient	8.5 (27.9)	10.8 (31.0)	12.4 (33.0)	8.9 (28.4)	8.6 (28.1)	10.0 (30.1)	13.3 (33.9)	7.4 (26.3)
Emotion-salient	9.4 (29.3)	10.4 (30.5)	11.2 (31.6)	9.6 (29.5)	9.9 (29.9)	9.3 (29.1)	11.2 (31.5)	8.5 (27.9)
Correct RT (ms)								
Race-salient	288 (108)	333 (114)	299 (106)	321 (110)	288 (106)	332 (112)	304 (114)	319 (116)
Emotion-salient	282 (106)	317 (106)	286 (113)	313 (103)	285 (107)	318 (110)	289 (113)	310 (103)
Incorrect RT (ms)								
Race-salient	269 (146)	258 (159)	251 (150)	282 (182)	270 (148)	242 (147)	244 (112)	271 (167)
Emotion-salient	249 (145)	257 (172)	243 (128)	251 (160)	253 (155)	265 (160)	239 (127)	266 (176)

Note. Values in parentheses are standard deviations. RT = response time.

Face Categorization Task

Table B7

Analyses of Error Rates and Correct Response Times by Condition (Experiment 2)

Effect	<i>F</i> (1, 253)	<i>p</i>	η_p^2
<i>Error rates</i>			
Saliency (Race vs. Emotion)	11.01	< .001	.042
Stimulus Set	0.16	.689	< .001
Saliency × Stimulus Set	1.02	.313	.004
<i>Correct response times</i>			
Saliency (Race vs. Emotion)	4.26	< .001	.239
Stimulus Set	0.19	.762	< .001
Saliency × Stimulus Set	0.19	.321	.004

Note. *Stimulus Set* refers to the two randomly selected sets of 24 faces (Race: 12 Black, 12 White; emotion expression: 12 angry/scowling, 12 happy/smiling).

Table B8

Descriptive Statistics for Face Categorization Task (Experiment 2)

Effect	Stimulus Set 1		Stimulus Set 2	
	Emotion Classification	Race Classification	Emotion Classification	Race Classification
Error rate (%)	0.05 (0.22)	0.03 (0.16)	0.04 (0.20)	0.03 (0.17)
Correct RT (ms)	718.7 (481.4)	517.5 (242.2)	710.8 (497.6)	542.9 (355.5)

Note. Values in parentheses are standard deviations. RT = response time.

Hierarchical Diffusion Decision Model

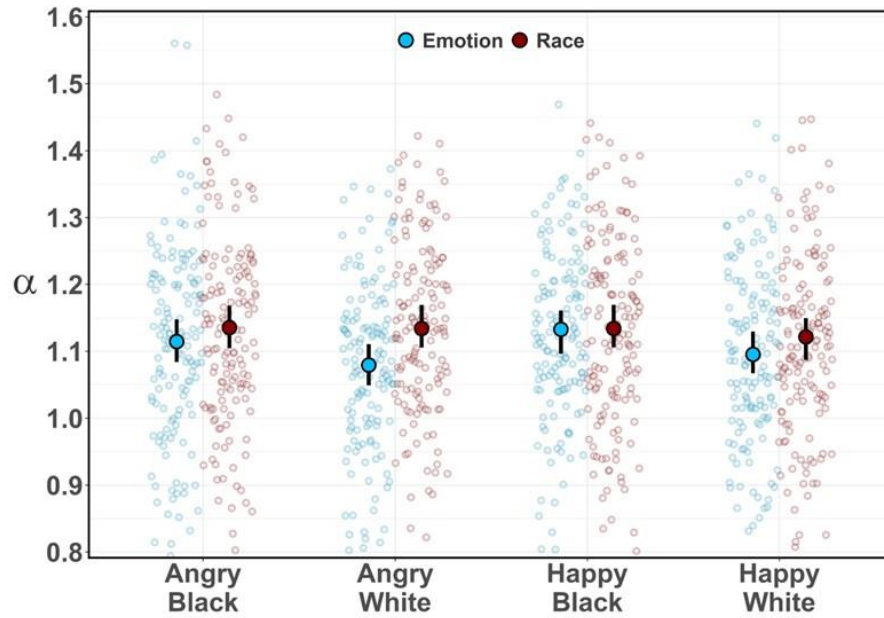
As stated in the main text, we estimated the model using an MCMC sampler in JAGS 4.30 (Plummer, 2003), with the Wiener distribution provided by Wabersich and Vandekerckhove (2014) and an estimation approach to make inferences in this framework (Gelman et al., 2003; Kruschke, 2014). We collected 12,000 samples using 10 chains with 1,200 samples per chain. We also included a burn-in of 500 samples and recorded only every tenth sample.

Representativeness of the posterior distribution was evaluated visually by inspecting caterpillar plots for overlap in MCMC chains, and it was evaluated numerically by checking whether the Gelman-Rubin convergence statistic (R-hat) was < 1.05 . Overall, the chains met these standards, suggesting representativeness of the posterior distributions. Accuracy was evaluated by examining autocorrelations, the effective sample size, and the Monte-Carlo standard error (MCSE). The ESS estimates the sample size of MCMC chains after accounting for autocorrelations. We sought an effective sample size $> 10,000$. Group-level distributions were $> 8,500$, with most $> 10,000$. The MCSE estimates the noise in the sampled estimates (Kruschke, 2015), and is computed by dividing the standard deviation of the chains by their effective sample size. All MSCE values were $< .001$, suggesting extremely low levels of noise in the model's parameter estimation. For brevity, we exclude tables and figures demonstrating the model's representativeness and accuracy from this document. However, that information is readily accessible within the online supplementary materials (<https://osf.io/hxywn/>), including tables which offer the ESS, MSCE, and R-hat values per condition, per experiment.

DDM Parameter Plots

Figure B7

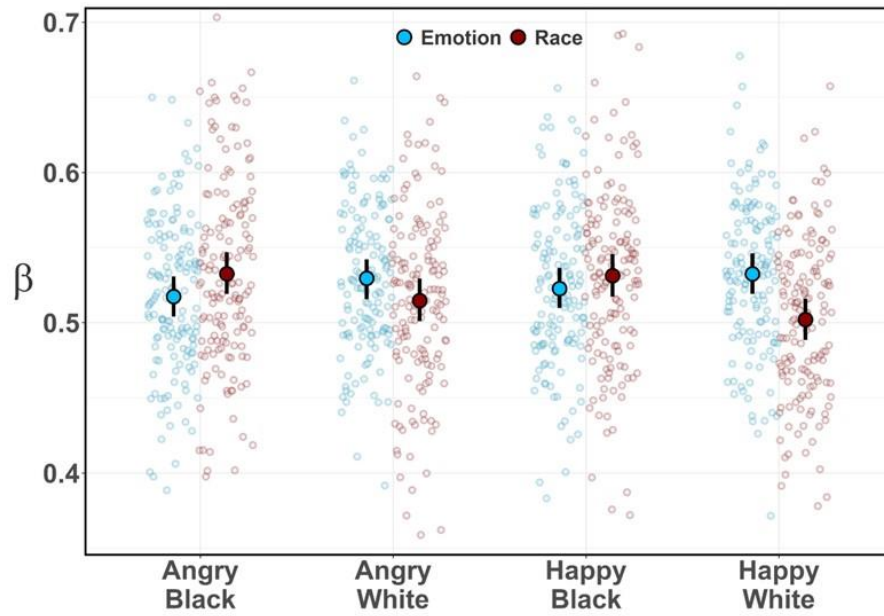
Alpha Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 1)



Notes. Alpha (threshold separation) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 1.

Figure B8

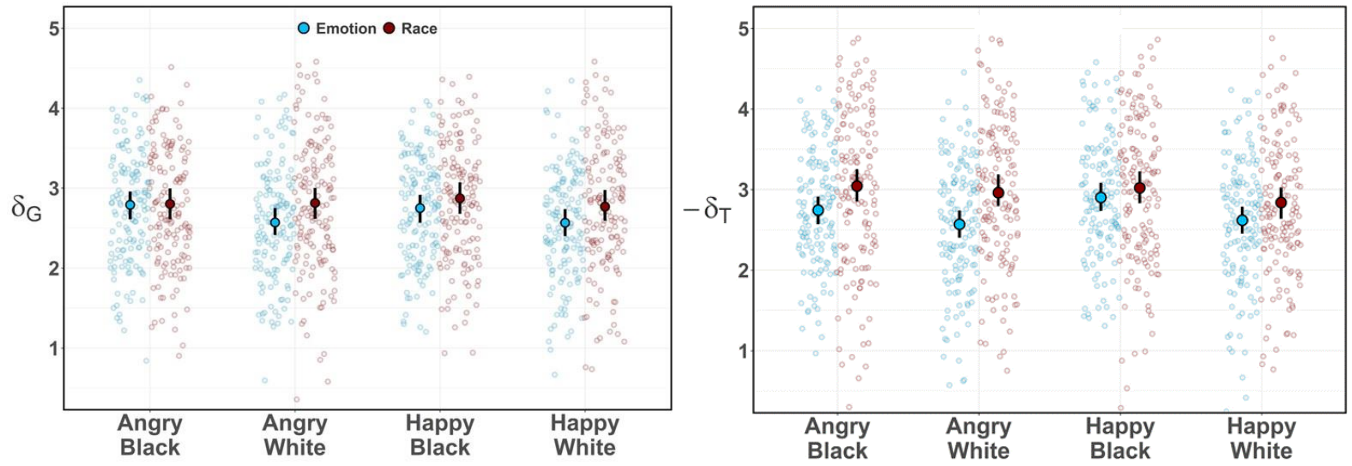
Beta Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 1)



Notes. Beta (starting point) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 1.

Figure B9

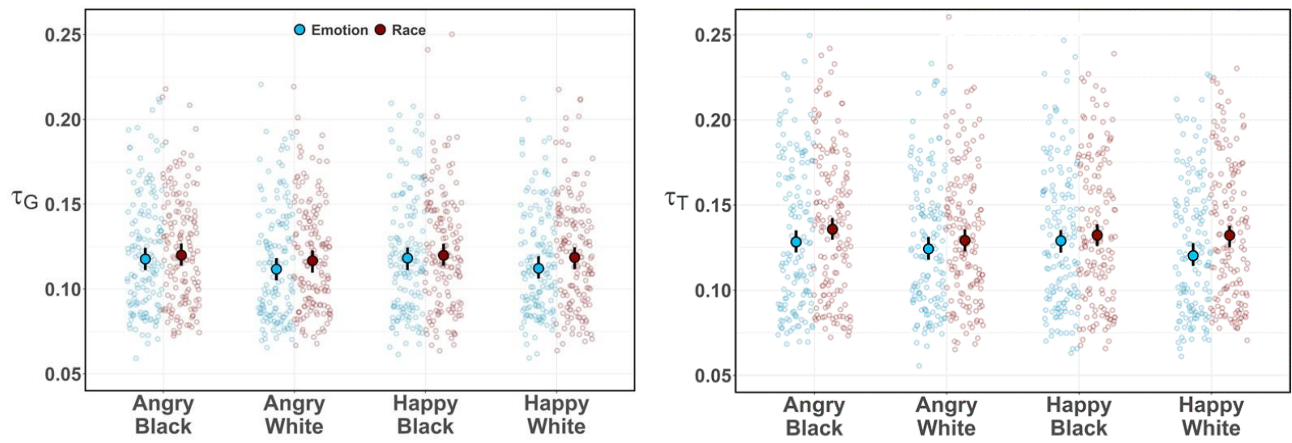
Delta Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 1)



Notes. Delta (drift rate) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 1. The plot on the left reflects delta estimates for tool trials. The plot on the right reflects delta estimates for gun trials.

Figure B10

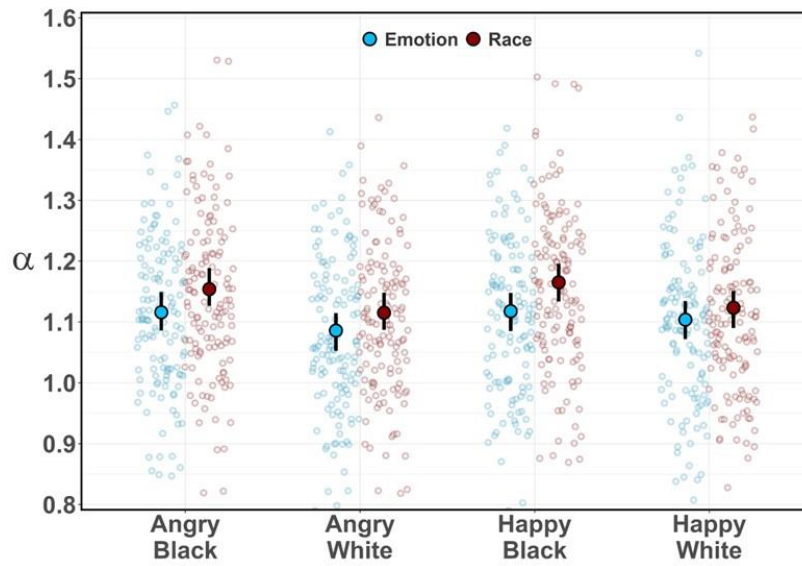
Tau Estimates by Race Prime, Emotion Prime, and Saliency Conditions (Experiments 1)



Notes. Tau (nondecision time) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 1. The plot on the left reflects tau estimates for tool trials. The plot on the right reflects tau estimates for gun trials.

Figure B11

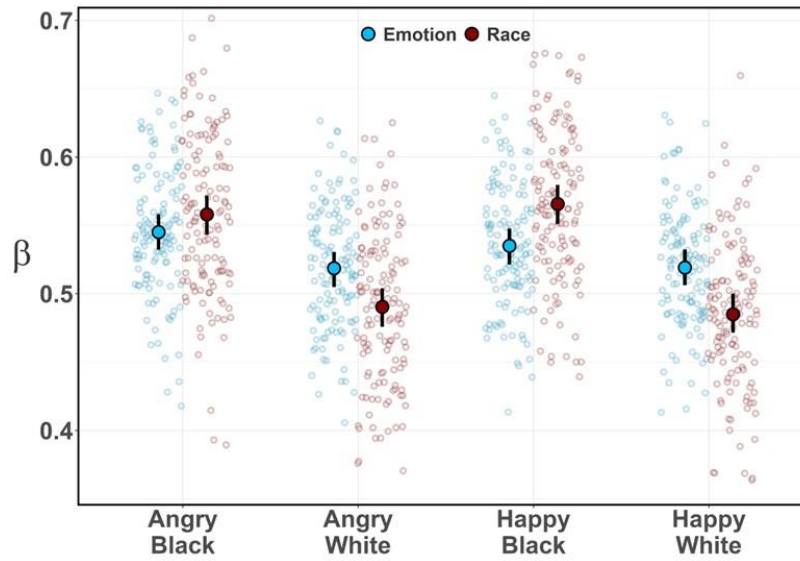
Alpha Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 2)



Notes. Alpha (threshold separation) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 2.

Figure B12

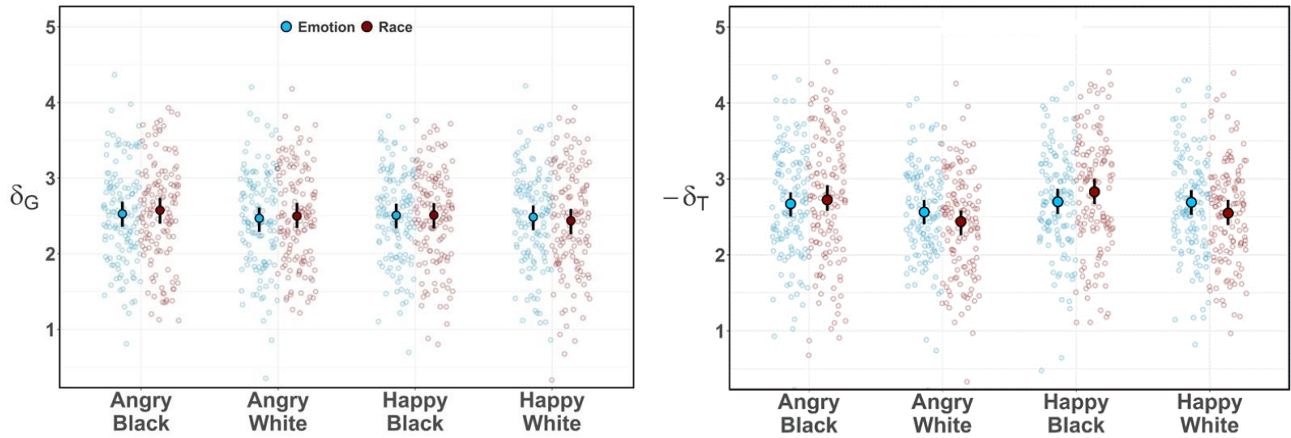
Beta Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 2)



Notes. Beta (starting point) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 2.

Figure B13

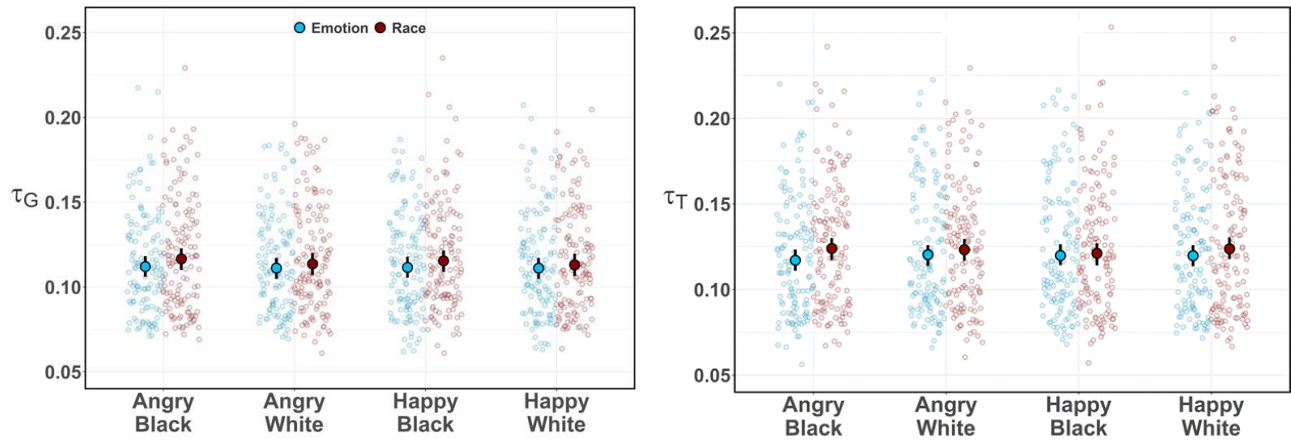
Delta Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 2)



Notes. Delta (drift rate) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 2. The plot on the left reflects delta estimates for tool trials. The plot on the right reflects delta estimates for gun trials.

Figure B14

Tau Estimates by Race Prime, Emotion Prime, and Salience Conditions (Experiments 2)



Notes. Tau (nondecision time) estimates. Empty dots reflect individual-level estimates and filled dots and their error bars reflect the most credible values and 95% highest density intervals, respectively, from the DDM modeled to data in Experiment 2. The plot on the left reflects tau estimates for tool trials. The plot on the right reflects tau estimates for gun trials.

Posterior Predictive Checks

Response Proportions

Within the online supplementary materials (<https://osf.io/hxywn/>), figures can be found which display the observed proportion of gun responses and the posterior predicted proportion of gun responses from the DDM. Overall, the model adequately characterized the proportion of times participants identified objects as guns. However, it should be noted that it somewhat overestimates the number of *gun* responses on trials with scowling face primes and Black face primes.

Response Times

Also within the online supplementary materials, figures can be found which display the observed and posterior predicted response times (RTs). Overall, the model adequately characterized the proportion of times participants identified objects as guns. The model overestimates incorrect response times, but does so consistently across trial types and conditions, and it varies to a highly similar pattern across quantiles of response times as seen in the observed data.

Appendix C

Chapter 3: Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research

Method

Participants

In total, 619 undergraduates consented to participate for course credit. We decided a priori to exclude data from participants who performed at or below chance (errors on $\geq 50\%$ of trials; $n = 22$) or who responded with the same key for more than 90% of trials ($n = 2$). Trial-level exclusions included the removal of responses recorded in < 100 ms or $> 2,500$ ms. Two participants' data were fully excluded due to having < 80 responses left after trial-level exclusions. The final sample comprised 593 participants (70.5% women; 45.5% Asian, 2.7% Black, 19.3% Latino, 18.2% White; $M_{\text{age}} = 19.90$, $SD = 2.45$).

Stimuli

This preregistered experiment relied on a stimulus set of faces selected from the Chicago Face Database (CFD; Ma et al., 2015). First, we sampled 10 female actors and 10 male actors from the CFD. When displaying neutral expressions, norming data from the CFD indicates that these actors' faces are correctly classified by gender $> 95\%$ of the time. We then selected two face images per actor: one image in which they display a smiling expression and another in which they display a scowling expression. In total, we were left with 40 face images.

We then applied facial morphing techniques, using the online morphing tool *WebMorph* (Debruine, 2018), to manipulate ambiguity within the original stimulus set of 40 faces. Smiling and scowling expressions, per actor, were blended together such that each actor was associated with two additional face images, besides their original smiling and scowling images: one image in which 55% of their smiling image and 45% of their scowling image remained (i.e., an

ambiguous smiling expression) and another in which 45% of their smiling image and 55% of their scowling image remained (i.e., an ambiguous scowling expression). At that point, we were left with double the original set of images – that is, 80 images in total. We then randomly paired male and female actors together, and morphed together their associated faces images. This secondary morphing procedure matched faces by their expression. If Female Actor A was randomly paired with Male Actor B, then their two unambiguous smiling expressions were morphed together, as were each of their other associated images (i.e., unambiguous scowling expressions, ambiguous smiling expressions, ambiguous scowling expressions). Morphs along sex generated ambiguous male (i.e., 45% male, 55% female) and ambiguous female (55% female, 45% male) face images. Because all image stimuli retained at least 55% properties of one level of sex and expression, each stimulus was associated with a correct response.

Procedure

In each experiment, participants were instructed to “quickly and accurately” classify a series of target faces. As previously mentioned, those targets varied along dimensions of sex and expression. Therefore, these two dimensions varied in relevance, depending on the task to which participants were assigned. Sex cues were considered task-relevant for those assigned to classify the faces by gender and task-irrelevant for those assigned to classify the faces by emotion. In contrast, expression cues were considered task-relevant for those assigned to classify the faces by emotion and task-irrelevant for those assigned to classify the faces by gender.

On each trial, stimuli were presented in the following sequence: a fixation cross for 500 ms, a target face for 200 ms, and, finally, a pattern mask until a response was recorded. Following a block of 16 practice trials, participants completed three critical test blocks of 60, 60, and 40 trials, with self-paced breaks provided between blocks. Each face was presented just once during

the critical test phase. Participants then provided demographics information before being debriefed on the purpose of the experiment. This experiment was approved by the University of California Institutional Review Board (protocol number: 223029; experiment name: *SHER806*).

Analytic Approach

Heterogeneity of Variance

We first tested the heterogeneity of participant variance with an asymptotic chi-squared test which holds a null-hypothesis that response distributions are identical between participants. Indeed, there was a significant level of heterogeneity between participants' responses in both the gender classification condition, $\chi^2(3564) = 6471.32, p < .001$, and emotion classification condition, $\chi^2(3528) = 6664.29, p < .001$. Therefore, a hierarchical Bayesian approach (Klauer, 2010) was implemented using the *TreeBUGS* package (Heck et al., 2018) to estimate the MCI model parameters. This approach estimates individual-level parameter values under the assumption that each individual set of parameters follows the same hierarchical distribution, essentially treating participants as random factors.

Assessing Model Fit

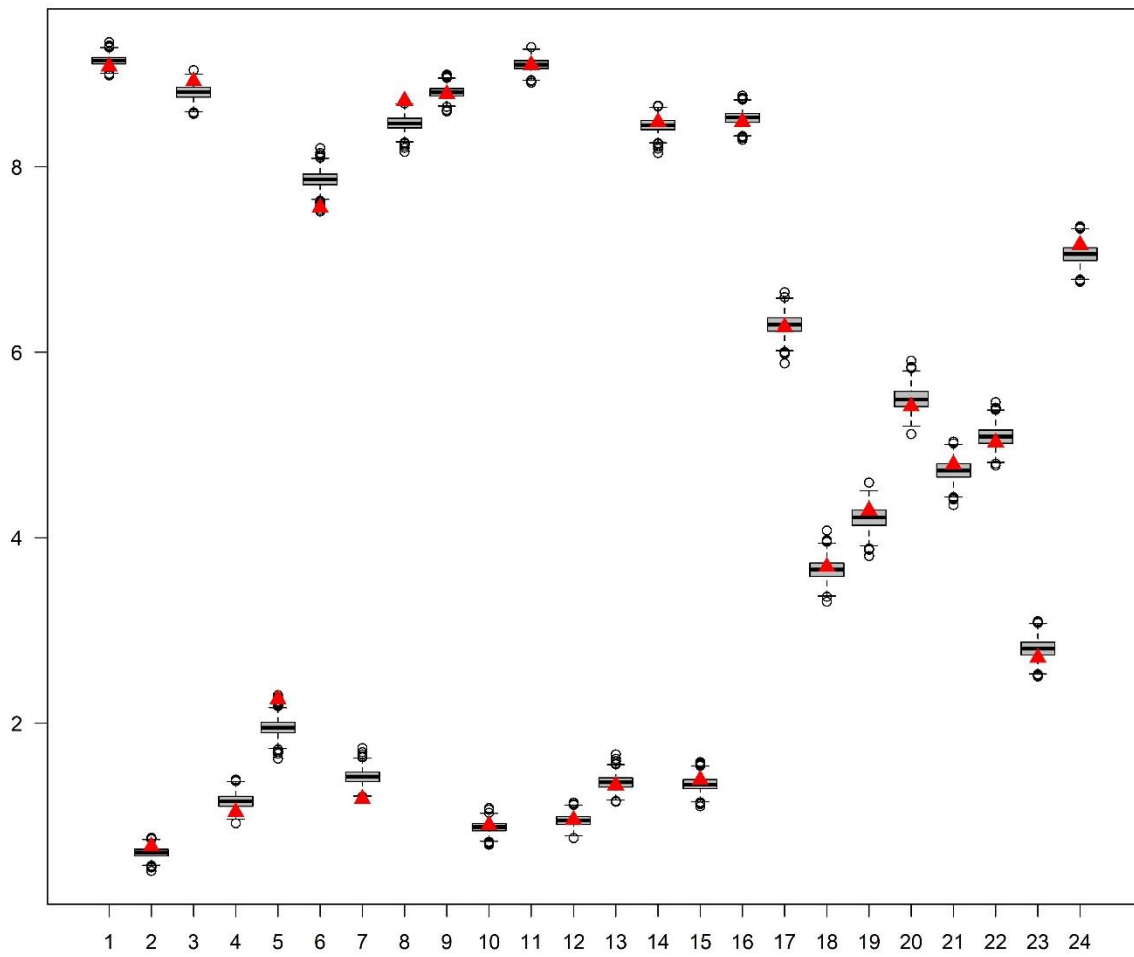
Model fit, the ability for the model to predict observed response data, was mainly assessed via posterior predictive checks (PPCs). That is, we assessed the divergence between the observed response data and those predicted by the model's posterior distribution. Analytically, we conducted PPCs via two fit statistics. First, T_1 indicates the model's ability to account for discrepancies between observed and expected responses across all participants. That is, T_1 indicates mean-level fit. Second, T_2 indicates the model's ability to account for discrepancies between observed and expected covariances, providing an assessment of individual-level fit. We collected one thousand samples from the model's posterior distribution, and the means and

covariances of the posterior were then compared to those in the observed data to produce T_1 (mean) and T_2 (covariance) statistics. Both T_1 and T_2 p-values reflect the probability that discrepancies between the expected and posterior-predicted data are greater than those between the expected and observed data (Heck et al., 2018, p. 273). The T_1 statistic is defined by a chi-square distribution and, therefore, must be considered in regards to the size of our dataset. The larger the dataset, the more sensitive the statistic will be at identifying even minor deviations of the model from the data. With $N_{\text{obs}} \sim 70,000$ observations, we expect a high level of sensitivity. Therefore, using the w statistic, we describe the level of misfit between the model's expectations and the observed aggregate responses. We rely on $w < .10$ as a descriptive benchmark for small levels of misfit.²³

²³ Due to poor model fit when both sources of information were ambiguous, we removed the eight response categories from this condition before modeling the data. Although we our preregistration states that we would model 32 response categories, and not 24, the model poorly characterizes those data. Alternatively, we could include those response categories and simply avoid analyzing the parameters across those conditions. The results of the parameter analyses do not meaningfully change between these selection criteria.

Figure C1

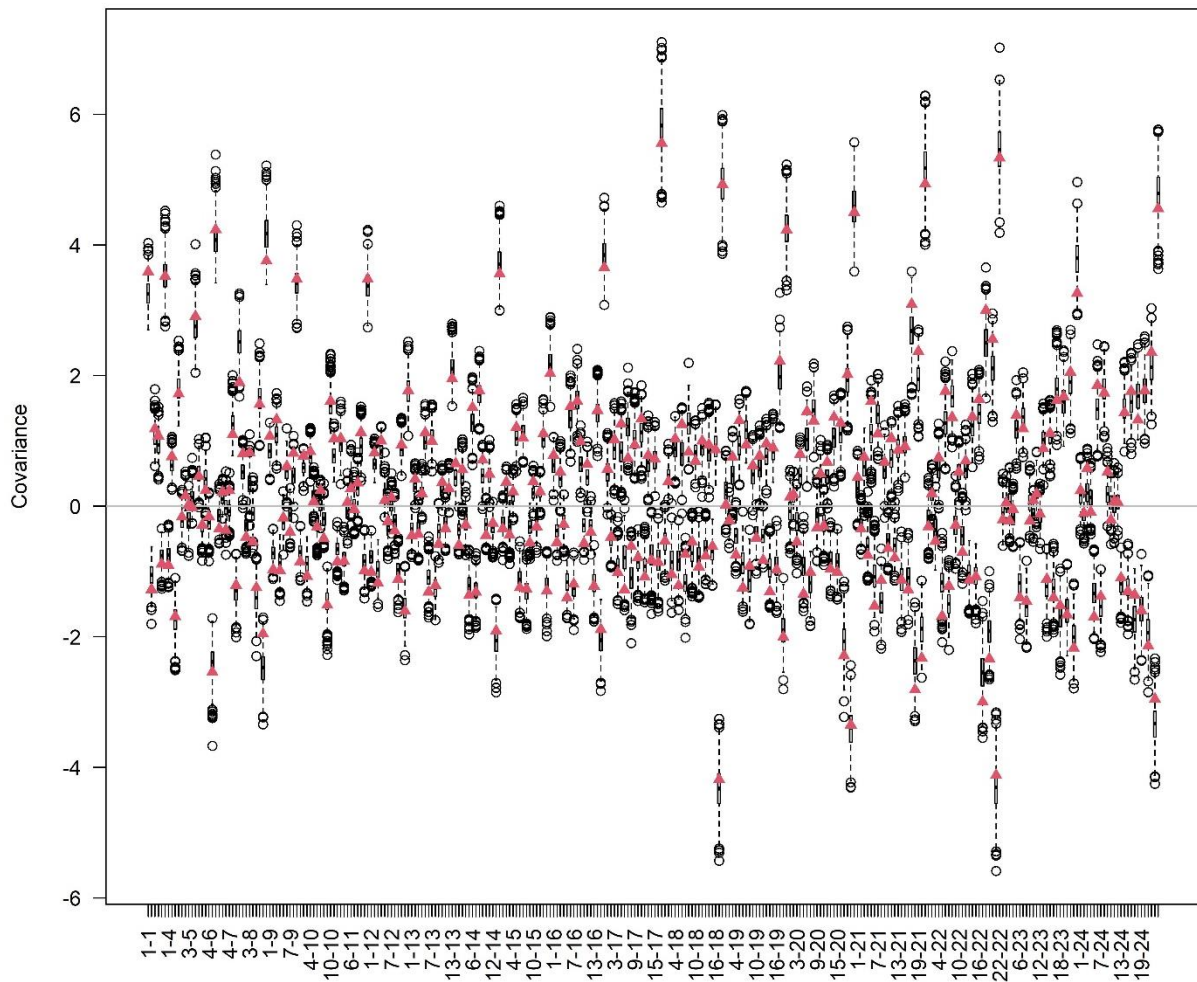
Plot of Posterior Predictive and Observed Mean Frequencies (Gender Classification)



Notes. Red triangle markers reflect observed mean frequencies. Boxplots reflect summarized frequencies from the fitted model's posterior distribution. Even-numbered response categories ("Woman" judgments) are the complement of the odd-numbered response categories ("Man" judgments) and are, therefore, excluded from the plot.

Figure C2

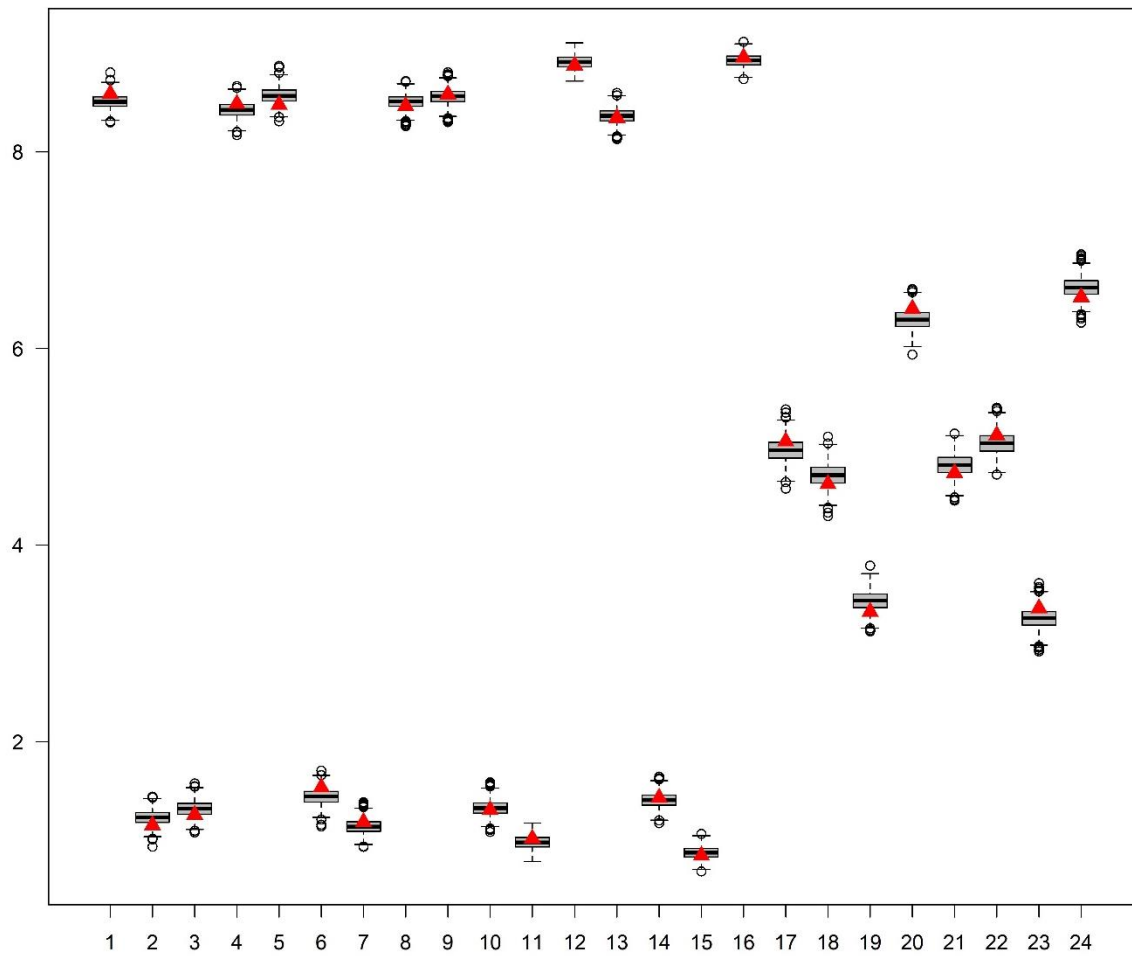
Plot of Posterior Predictive and Observed Covariances (Gender Classification)



Notes. Red triangle markers reflect observed covariances. Boxplots reflect summarized covariances sampled from the fitted model’s posterior distribution. Even-numbered response categories (“Woman” judgments) are the complement of the odd-numbered response categories (“Man” judgments) and are, therefore, excluded from the plot.

Figure C3

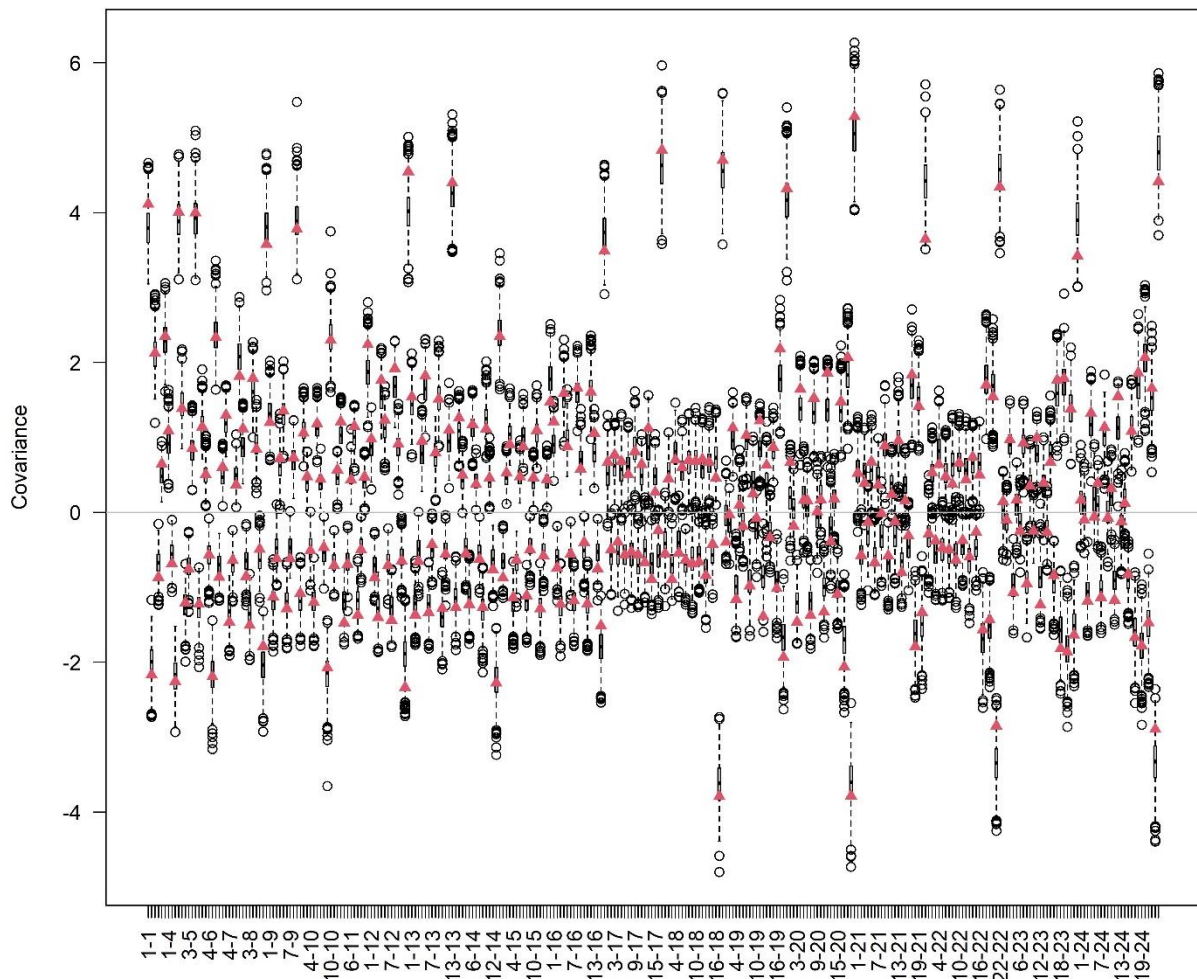
Plot of Posterior Predictive and Observed Mean Frequencies (Emotion Classification)



Notes. Red triangle markers reflect observed mean frequencies. Boxplots reflect summarized frequencies from the fitted model's posterior distribution. Even-numbered response categories ("Woman" judgments) are the complement of the odd-numbered response categories ("Man" judgments) and are, therefore, excluded from the plot.

Figure C4

Plot of Posterior Predictive and Observed Covariances (Emotion Classification)



Notes. Red triangle markers reflect observed covariances. Boxplots reflect summarized covariances sampled from the fitted model’s posterior distribution. Even-numbered response categories (“Woman” judgments) are the compliment of the odd-numbered response categories (“Man” judgments) and are, therefore, excluded from the plot.

Additional Analyses

Gender Classification

Parameter comparisons revealed a credible and large effect of Ambiguity on sex processing, $\Delta C_1 = .62$, $BCI_{95\%} [.59, .64]$, indicating that the use of sex cues to classify faces by gender decreased with ambiguity. Although ambiguity in expression cues also decreased the use of sex cues to classify faces by gender, $\Delta C_1 = -.04$, $BCI_{95\%} [-.06, -.02]$, this effect was >30 times smaller than the effect of ambiguity in sex cues on sex processing, $\Delta C_{1.sex} / \Delta C_{1.expression} = 32$.

There was also a credible effect of Expression Ambiguity on expression processing, $\Delta C_2 = .23$, $BCI_{95\%} [.16, .29]$, indicating that the use of facial expressions to classify faces by gender decreased when they were ambiguous. The effect of expression ambiguity on expression processing was >5 times larger than its effect on sex processing, $\Delta C_2 / \Delta C_1 = 5.75$.

Emotion Classification

Parameter comparisons revealed a credible and large effect of Ambiguity on expression processing, $\Delta C_2 = .64$, $BCI_{95\%} [.60, .67]$, indicating that the use of facial expressions to classify faces by emotion decreased with ambiguity. Although ambiguity in sex cues also decreased the use of facial expressions to classify faces by emotion, $\Delta C_2 = -.03$, $BCI_{95\%} [-.05, >-.01]$, this effect was >25 times smaller than the effect of expression ambiguity on expression processing, $\Delta C_{2.sex} / \Delta C_{2.expression} = 25.22$.

Task Relevance

To explore whether each cue is used more if it is relevant versus irrelevant to the intended judgment, we compared parameter estimates between gender and emotion classification tasks. Sex cues were used more when they were task-relevant versus task-irrelevant, $\Delta C_1 \geq .13$, $BCI_{95\%}$

[$\geq .11$, $\geq .15$]. Facial expressions were used more often when they were task-relevant versus task-irrelevant, $\Delta C_2 \geq .14$, $BCI_{95\%}$ [$\geq .11$, $\geq .17$] (see Table C1).

Table C1

Analyses of C1 and C2 Parameters Across Conditions of Cue Ambiguity

Effect	Mean Difference	SD	$BCI_{95\%}$
<i>Use of Sex Cues (C₁) During Gender Classification</i>			
Unambiguous Sex and Expression Cues	.74	.01	[.72, .77]
Ambiguous Sex Cues	.13	.01	[.11, .15]
Ambiguous Expression Cues	.78	.01	[.75, .81]
<i>Use of Expression Cues (C₂) During Emotion Classification</i>			
Unambiguous Sex and Expression Cues	.55	.04	[.48, .62]
Ambiguous Sex Cues	.59	.02	[.56, .63]
Ambiguous Expression Cues	.14	.02	[.11, .17]