Intersectional Implicit Bias: Evidence for Asymmetrically Compounding Bias and The Predominance of Target Gender

Paul Connor[1], Matthew Weeks[2], Jack Glaser[3], Serena Chen[4], Dacher Keltner[4]

[1]Department of Psychology, 1190 Amsterdam Ave, Columbia University, New York, New York 10027.

[2]Department of Psychology, 2000 North Parkway, Rhodes College, Memphis, Tennessee, 38112.

[3]Goldman School of Public Policy, 2607 Hearst Ave, University of California, Berkeley, Berkeley, California, 94720.

[4]Department of Psychology, 2121 Berkeley Way, University of California, Berkeley, Berkeley, California, 94720.

Correspondence concerning this article may be addressed to Paul Connor: paulrobertconnor@gmail.com, 1190 Amsterdam Ave, Columbia University, New York, New York 10027.

Acknowledgments

This research was supported by grants from Haas Business School's XLab and the NSF (SPRF 2104594). Special thanks goes to An Nghiem, Seunghun Lee, Mriganka Singh, and Mimi Yoo for their indispensable work as research assistants, Jazmin Brown-Iannuzzi and Stephen Antonoplis for their helpful comments and advice, and both the RASCL lab and the XLab at UC Berkeley for assistance with software access.

Abstract

Little is known about implicit evaluations of complex, multiply categorizable social targets. Across five studies (N = 5,204), we investigated implicit evaluations of targets varying in race, gender, social class, and age. Overall, the largest and most consistent evaluative bias was pro-women/anti-men bias, followed by smaller but nonetheless consistent pro-upper-class/anti-lower-class biases. By contrast, we observed less consistent effects of targets' race, no effects of targets' age, and no consistent interactions between target-level categories. An integrative data analysis highlighted a number of moderating factors, but a stable pro-women/anti-men and pro-upper-class/anti-lower-class bias across demographic groups. Overall, these results suggest that implicit biases compound across multiple categories asymmetrically, with a dominant category (here, gender) largely driving evaluations, and ancillary categories (here, social class and race) exerting relatively smaller additional effects. We discuss potential implications of this work for understanding how implicit biases operate in real-world social settings.

Keywords: implicit bias, intersectionality, social class, person perception, social cognition

Intersectional Implicit Bias: Evidence for Asymmetrically Compounding Bias and The Predominance of Target Gender

People display implicit evaluative biases—differences in patterns of automatic and often unconscious responses to varying kinds of stimuli—with respect to a wide variety of social categories, including race, gender, social class, and age (Greenwald & Lai, 2020; Nosek, 2005). These biases may have weighty social consequences, influencing decision making in contexts including employment, medicine, and voting (e.g., Greenwald, Banaji, & Nosek, 2015; Jost et al., 2009).

In most human interactions, individuals display multiple intersecting social identities, such as race, gender, social class, and age. Yet within the empirical literature on implicit bias, biases regarding such categories have typically been studied in isolation from each other, and most measures of implicit bias have been designed to isolate and measure biases regarding a single binary categorical preference at a time. For example, Nosek (2005) employed Implicit Association Tests (IATs; Greenwald, McGhee, & Schwartz, 1998) to demonstrate that US participants display implicit evaluative biases favouring White targets over Black targets, women over men, the rich over the poor, the young over older adults, and many others. However, IATs measure only a single categorical preference at a time, and do not speak to how multiple identities jointly contribute to implicit bias. Does a White, rich, young woman prompt implicit evaluations four times more positive than a Black, poor, old man? Are some social categories more influential than others? Do the categories interact with each other, such that, for example, implicit gender bias operates differently depending on the race, social class, age, weight, or sexual orientation of targets?

To date, psychologists have produced few answers to these questions, despite the rising prominence of a intersectional approaches within psychological science (e.g., Cole, 2009; Goff, & Kahn, 2013; Kang & Bodenhausen, 2015). There is, however, considerable evidence that implicit evaluations are sensitive to multiple aspects of target stimuli. Wittenbrink, Judd, and Park (2001) found implicit racial bias to be moderated by the visual contexts in which targets were presented. When Black and White targets were depicted on a street corner, participants displayed greater anti-Black bias compared to when targets were depicted inside a church. Similarly, Barden, Maddux, Petty, and Brewer (2004) found moderation of implicit bias by visual context and targets' clothing. When Black and White targets were depicted inside a jail,

participants displayed pro-White bias when targets were shown in prison clothes, but pro-Black bias when targets were shown in suits and ties. In keeping with this theme of moderation, participants showed greater implicit bias against Black targets with more racially prototypical features (Livingston & Brewer, 2002), and toward Black targets with neutral facial expressions compared to smiling Black targets (Steele, George, Cease, Fabri, & Schlosser, 2018). Each of these findings suggests that implicit evaluative biases respond to multiple variables within target stimuli. By implication, when targets are multiply categorizable—as in most everyday social interactions—it follows that implicit evaluations will likely be shaped by multiple dimensions of social categorization.

Models of Intersectional Intergroup Bias

Several schools of thought have considered how intergroup biases respond when multiply social categories are displayed by social targets (for recent reviews, see Nicolas, de la Fuente, & Fiske, 2017, and Petsko & Bodenhausen, 2019). Here, we consider in detail select treatments, focusing upon those most relevant to the present work and results.

Compounding Biases: Additive and Interactive Models

One thesis is that negative and positive biases *compound* when multiple social identities are displayed simultaneously. In early work, Brown and Turner (1979) relied on Tajfel and Turner's (1979) social identity theory to predict that separate intergroup biases would combine *additively* in the presence of multiple dimensions of social categorization. Their reasoning held that intergroup bias will increase in a linear fashion according to the number of dimensions on which a social target is perceived to be an out-group member, and decrease according to the number of dimensions on which they are perceived as an in-group member. A similar thesis is the *averaging* model of Singh, Yeoh, Lim, and Lim (1997), which proposes that intergroup bias is a function of the number of perceived out-group memberships divided by the total number of available social categorizations.

Other scholars have suggested that biases may compound across categories in interactive ways. Grounded in the writings of Black feminist activist Frances Beale (1970), Ransford (1980) proposed the *multiple jeopardy/advantage hypothesis,* which posits that individuals belonging to multiple stigmatized social categories are vulnerable to 'multiple jeopardy:' a negative bias that exceeds the sum of the negative biases associated with each category. By contrast, individuals belonging to multiple positively-valued social categories may benefit from 'multiple advantage:'

a positive bias that exceeds the sum of the positive biases associated with each category (see also Almquist, 1975; King, 1988; Landrine, Klonoff, Alcaraz, Scott, & Wilkins, 1995). In her widely known early treatment of "intersectionality", Crenshaw (1989) described a paradigmatic case of multiple jeopardy in the US legal system: despite General Motors hiring disproportionately fewer Black women, the company was exculpated of both race and gender discrimination due to employing sufficient numbers of (White) women and (male) Black people (*DeGraffenreid v. GENERAL MOTORS ASSEMBLY DIV.*, 1976).

Today, scholarship animated by the concept of intersectionality often presupposes compounding effects of multiple marginalized social identities (especially pertaining to Black women in the USA; Cooper, 2015). Within this literature, however, it has not always been clear whether intersectionality necessarily implies interactive (i.e., multiplicative) effects between social categories, or simply that individuals with multiple marginalized social identities suffer from multiple consequences as a result of their various identities. Indeed, scholars of intersectionality have at times been divided on the question of whether the concept can or should be reduced to these kinds of quantitative predictions (e.g., Cole, 2009; Bowleg, 2008).

Nonetheless, numerous researchers have sought to document the simultaneous effects of multiple intersecting social categorizations on the expression of intergroup bias. At times, evidence has been most consistent with multiple additive main effects on intergroup bias compounding across different social categorizations (e.g., Crisp, Hewstone, & Rubin, 2001, Study 1; Hewstone, Islam, & Judd, 1993; Islam & Hewstone, 1993, Study 2; Singh, Yeoh, Lim, & Lim, 1997; Vanbeselaere, 1991; van Oudenhoven, Judd, & Hewstone, 2000). At other times, evidence has been consistent with multiplicative disadvantages stemming from combined stigmatized social identities (e.g., Brown & Turner, 1979; Diehl, 1990; Marcus-Newhall, Miller, Holz, & Brewer, 1993; Vanbeselaere, 1991), or with multiplicative advantages stemming from combined positively-valued social identities (Brewer, Ho, Lee, & Miller, 1987; Eurich-Fulcher, & Schofield, 1995).

Thus, despite some ambiguity regarding the presence and pattern of interaction effects, theories of compounding bias make clear predictions with regard to the specific sub-groups of multiply categorizable targets that should evoke the most positive or negative implicit evaluations. In the case of implicit bias, for example, prior evidence suggests that Americans' implicit evaluative biases typically favour White over Black targets (Nosek, Banaji, &

Greenwald, 2002), women over men (Richeson & Ambady, 2001, Rudman & Goodwin, 2004), the upper class over the lower class (Horwitz & Dovidio, 2017; Rudman, Feinberg, & Fairchild, 2002), and the young over older adults (Nosek, 2005). Theories of compounding bias therefore predict that among targets varying in race, gender, social class, and age, the most negative implicit evaluative biases should be displayed toward lower-class, older Black men, whereas the most positive biases should be displayed toward upper-class, younger White women.

Category Dominance

Other researchers have challenged the claim that separate biases will necessarily compound in additive or interactive ways toward multiply categorizable targets.[1] One alternate view is the *category dominance model* (Macrae, Bodenhausen, & Milne, 1995), which is premised on the notion that due to the complexity of social stimuli, humans must by necessity act as 'cognitive misers' (Fiske & Taylor, 1991). When facing multiply categorizable targets, this view holds, people will often rely on a single social category to guide social perception. Which specific category becomes dominant depends on many factors, such as the situational or chronic salience of different categories, the goals of perceivers, and/or perceivers' prejudices. Once the dominant category is activated, it will inhibit the activation of competing categories. In support of this, Macrae and colleagues showed that when participants were primed with a specific social category (i.e., Asian or woman) and observed a multiply categorizable target (i.e., an Asian woman), concepts associated with the primed category became more cognitively accessible, while concepts associated with the non-primed category became less cognitively accessible (see also Dijksterhuis & Van Knippenberg, 1996).

The category dominance model therefore predicts that in evaluations of targets varying in race, gender, social class, and age, a single dominant categorization will drive bias. Importantly, the model does not necessarily predict what the dominant category will be—if no specific category is primed by researchers, the dominant category will depend upon the perceivers' attention, goals, and pre-existing biases.

---

[1] Other perspectives that challenge the notion of compounding bias include Urada, Stenstrom, and Miller's (2007) threshold-based *feature detection* model, and Kang and Chasteen's (2009) category salience-based *selective inhibition model*. For the sake of brevity, we do not discuss these theories in the present manuscript, though our data is arguably relevant to, and fails to show support for, either model.

Existing Evidence Regarding Intersectional Implicit Bias

Select studies have investigated implicit bias toward multiply categorizable targets. Thiem, Neel, Simpson and Todd (2019) used a weapon identification task (Payne, 2001) and sequential priming tasks to measure automatic associations between weapons and headshots of targets varying in race (Black and White), gender, and age. Consistent with compounding bias accounts, each social category influenced responses, with participants displaying a greater tendency to associate Black, male, and adult targets with weapons compared to White, female, and child targets. Additionally, there was some evidence of a multiplicative multiple-jeopardy effect, with Black male targets appearing to evoke stronger associations with threat than could be explained by main effects of race and gender alone. Similarly, Perszyk, Lei, Bodenhausen, Richeson, and Waxman (2019) used the Affective Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) to measure children's implicit evaluations of headshots of child targets varying in race (White and Black) and gender. In this study a race × gender interaction emerged, with Black boys eliciting more negative evaluations than could be explained by main effects of race and gender alone.

Other work has considered the intersecting effects of race and class. Moore-Berg, Karpinski, and Plant (2017) presented images of the upper bodies of targets varying in race (Black and White) and social class (signalled via targets' wearing either t-shirts or suits) within a 'shoot/don't-shoot' task (Correll, Park, Judd, & Wittenbrink, 2002). Similarly, Mattan, Kubota, Li, Venezia, and Cloutier (2019) used an Evaluative Priming Task (EPT; Fazio, Sanbonmatsu, Powell, & Kardes, 1986) to measure implicit evaluations of headshots of targets varying in race (Black and White) and background color (red and blue), with participants trained to associate background colors with higher or lower social status. The results of these studies varied, with five unique patterns of results emerging from five separate experiments. However, one consistent result was that in each experiment, upper-class White targets were relatively favored by responses (though not always more so than lower-class White targets or upper-class Black targets). These studies can therefore also be considered broadly consistent with compounding bias models, with upper-class Whites appearing to be the sub-group most favored by displayed biases.

By contrast, other studies have yielded results more consistent with the category dominance model. Mitchell, Nosek, and Banaji (2003) presented Black athletes and White

politicians as stimuli within an IAT, but had participants categorize targets either via profession (Athlete vs. Politician) or race (Black vs. White). When targets were categorized by profession, biases favoured Black athletes, but when targets were categorized by race, biases favoured the White politicians. The same authors also presented Black female and White male targets within a Go/No-Go Association Test (Nosek & Banaji, 2001), and manipulated the relative salience of targets' race and gender. Results indicated that when race was salient, participants evaluated White males more positively than Black females, but when gender was salient, participants evaluated Black females more positively than White males. Similarly, Yamaguchi and Beattie (2019) found that when Black and White female and male targets were categorized according to race within IATs, participants displayed substantial anti-Black/pro-White implicit racial bias, but little implicit gender bias. But when targets were categorized according to gender, participants displayed pro-female/anti-male implicit gender bias, but little implicit racial bias.

Further evidence suggests that the direct manipulation of category salience is not always necessary for a single category to dominate responses to multiply categorizable targets. Jones and Fazio (2010) used a weapon identification task to measure participants' tendency to perceive objects as guns versus tools while exposed to images of primes varying in race (Black and White), gender, and occupational status (high or low, e.g., professor, sanitation worker). In this study, participants instructed to attend to primes' race displayed an implicit racial bias were relatively more likely to perceive guns/tools while exposed to Black/White primes, but showed little gender- or occupation-based bias. However, when participants were not instructed to attend to any specific social category, the only bias displayed was gender-based, with participants relatively more likely to perceive guns/tools when exposed to male/female targets.

Finally, other researchers have argued that category dominance in implicit evaluation tasks also depends on the task employed. Gawronski, Cunningham, LeBel, and Deutsch (2010) measured implicit evaluations of targets varying in race (Black and White) and age via EPTs and AMPs, while instructing participants to attend either to targets' race or age. Results suggested that non-attended categories affected evaluations on the AMP but not the EPT, leading the authors to argue that tasks structured to induce response interference—such as the EPT—may be especially conducive to category dominance, whereas other tasks such as the AMP are not.

The Present Research

In most social interactions, individuals can be categorized in multiple ways. Thus, understanding how implicit evaluative bias operates toward multiply categorizable targets is likely to be critical to understanding how it operates in everyday life. However, current evidence concerning implicit bias and multiply categorizable targets is inconclusive. Whereas some work supports theories of compounding bias, and suggests that implicit biases tend to compound across multiple social categories, other work aligns better with the category dominance perspective, and suggests that implicit evaluations are often driven by a single dominant categorical dimension.

Guided by these contrasting perspectives, we conducted four studies investigating implicit evaluations of multiply categorizable targets. In Study 1, we measured evaluations of full-body target photographs of males varying in race (Black or White) and social class status. In Study 2, we extended on this approach, and incorporated target images varying in race, gender, and social class, as well as a data-driven approach to determine the primary dimensions of perceived target-level variation and their respective influence on implicit evaluations. In Study 3, we again measured implicit evaluations of targets varying in race, gender, social class and age, but presented targets via full-body or upper-body photographs, and tightened experimental control over potential confounds by shuffling targets' faces and bodies. In Study 4, we tested the generalizability of our results by obtaining data from two nationally representative samples of US adults, and by comparing results across different measurement methods. Finally, in Study 5 we conducted an integrative data analysis of the data from Studies 2-4 to test the extent to which patterns of results differed among different sub-groups of respondents, and to better elucidate the precise explanation for our patterns of results.

The present research offers theoretical, empirical, and methodological advances for the study of intersectional implicit bias. At the theoretical level, this work presents a novel perspective regarding the simultaneous influence of multiple social categories on intergoup biases—asymmetrically compounding bias—which in part reconciles competing theories of compounding bias and category dominance. At the empirical level, the present work is to our knowledge the first to measure implicit evaluations of targets systematically varying in the variables of race, gender, social class, and age, each of which tend to be simultaneously perceptible among the majority of real-world social targets. And at the methodological level, the present work is to our knowledge the first to focus specifically upon measuring and modelling

implicit evaluations of multiply categorizable targets at the individual target level, which we argue carries multiple advantages over previous approaches. All data and code used in the current project are accessible via the Open Science Framework (https://osf.io/sbpna/).

Study 1

In Studies 1a and 1b we measured implicit evaluations of full-body images of male targets varying in race (Black or White) and social class. Theories of compounding bias predict that pro-White/anti-Black biases and pro-upper-class/anti-lower-class biases should both occur, resulting in lower-class Black targets being evaluated most negatively, and upper-class White targets being evaluated most positively. They also suggest possible interaction effects, with either lower-class Black targets producing especially negative responses (multiplicative multiple jeopardy), or upper-class White targets producing especially positive responses (multiplicative multiple advantage). Conversely, the category dominance model suggests that either race or social class will emerge as the dominant category driving implicit bias.

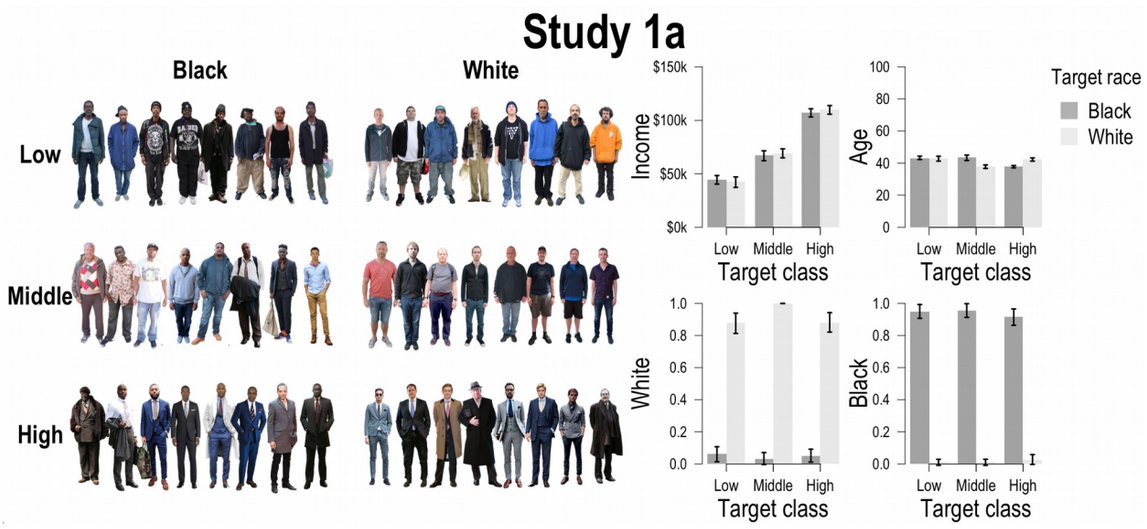
Stimuli Creation and Pilot Studies

We gathered 130 full-body color photographs of Black and White adults (60 Black, 70 White). Targets appeared on plain white backgrounds facing forward with neutral expressions. Photographs were presented to 1788 U.S. adults recruited via MTurk, who rated the photographs on perceived yearly income (ICC = 0.43), perceived age (ICC = 0.70), and whether they perceived targets to be Black (ICC = 0.88) or White (ICC = 0.95). Raters offered judgments of an average of 29.73 (*SD* = 13.61) randomly selected photographs, and each photo was rated on each trait by an average of 52.58 raters (*SD* = 23.08).

Based on photographs' mean ratings, we assembled groups of eight photos varying in race (Black and White) and income, but matched in age (see Figure 1). In each study, targets' mean perceived income varied significantly across class categories (all $p < .001$) but not race categories (all $p > 0.69$), whereas targets' mean perceived race varied significantly across race categories (all $p < .001$) but not class categories (all $p > .08$).[2] Additionally, there were no

[2] The *p* value of 0.08 referred to resulted from a t-test comparing Study 1b's 16 lower-class and 16 upper-class targets on their mean categorizations as White (see the bottom-left bar plot in Figure 1). Although not ideal, this result is un-problematic for interpreting Study 1b's results. As shown in Figure 2, Study 1b's Black targets (who were categorized as White 3% of the time) produced more positive evaluations than Study 1b's White targets (who were categorized as

significant interactions between race and class categories in predicting perceived income or race (all $p > 0.19$), and no significant main effects or interactions of race and class categories in predicting perceived age (all $p > 0.32$).



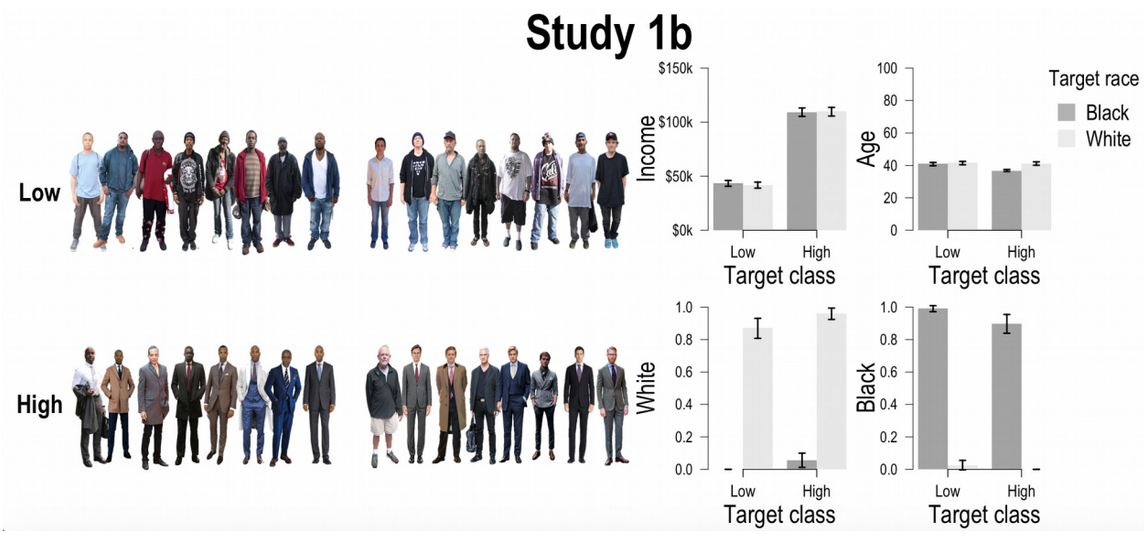


Figure 1. Target groups used in Studies 1a and 1b, and figures displaying raters' judgments of perceived income, age, and race ratings of each group. Bars indicate 95% confidence intervals.

White 91% of the time). It is therefore highly unlikely that participants responded more positively to the upper-class targets (who were categorized as White 50% of the time) than the lower-class targets (who were categorized as White 43% of the time) due to a race confound.

Participants and Procedure

Participants for Study 1a (N = 307, 196 women, 100 men, 11 missing gender data, M_{age} = 20.3, SD_{age} = 1.9, 129 Asian,[3] 125 White, 28 Latino, 9 Black, 5 other race, 11 missing race data) and Study 1b (N = 533, 340 women, 170 men, 1 non-binary, 22 missing gender data, M_{age} = 20.5, SD_{age} = 2.63, 268 Asian, 173 White, 54 Latino, 10 Other race, 6 Black, 22 missing race data) were undergraduates who participated for course credit. Study 1a used a within-subjects design, with participants' responding to all six of the target groups in a randomized order, whereas Study 1b used a between-subjects design, with participants randomly assigned to respond to one of the four target groups.

Single-Target IATs

In both studies we measured implicit evaluations of target groups via evaluative Single Target IATs (ST-IATs; Bluemke & Friese, 2008; Wigboldus, Holland, & van Knippenberg, 2004).[4] Each ST-IAT began with a practice block, in which the labels "Good" and "Bad" appeared at the top left and right of participants' computer screens. Across 20 trials participants then classified words appearing on their screens as either good (e.g., Beautiful) or bad (e.g., Agony) as quickly as possible via timed computer key presses. Following this, the word "Person" also appeared at either the top left of screens (in 'compatible' blocks), or the top right of screens (in 'incompatible' blocks), and participants categorized words as "Good" or "Bad" and targets as a "Person." Participants were randomly assigned either to complete two compatible blocks (of 20 then 40 trials) followed by two incompatible blocks (of 20 then 40 trials), or *vice versa* (see Table 1).

In Study 1b we also used a wealth ST-IAT to measure implicit associations between target groups and the concepts of wealth and poverty. In this measure the labels "Good" and "Bad" were replaced with "Wealth" and "Poverty," and the positively and negatively valanced

[3] Our demographic survey did not delineate between sub-categories of Asian-identifying students, so likely includes participants of East, South-, and Southeast-Asian descent.

[4] ST-IATs are highly similar to the Single-Category IAT (SC-IAT) introduced by Karpinski and Steinman (2006). We follow Bluemke and Friese (2008) in distinguishing between the tasks on the basis that the SC-IAT uses an in-task response maximum latency window while the ST-IAT does not. In the present manuscript, we did not use a limited response latency window, so classify our task as a ST-IAT, not a SC-IAT.

words were replaced with words evoking wealth (e.g., Rich, Wealth, Affluent) and poverty (e.g., Poor, Poverty, Destitute).

Table 1
Single Target IAT procedure

Block	Task description	Left key (E)	Right key (I)	Trials
1	Practice block	Positive[a]/Wealth words[c]	Negative[b]/Poverty[d] words	20
2	Compatible block 1	Positive/Wealth words + target images	Negative/Poverty words	20
3	Compatible block 2	Positive/Wealth words + target images	Negative/Poverty words	40
4	Incompatible block 1	Positive/Wealth words	Negative/Poverty words + target images	20
5	Incompatible block 2	Positive/Wealth words	Negative/Poverty words + target images	40

[a]Positive words = Beautiful, Glorious, Joyful, Lovely, Marvellous, Pleasure, Superb, Wonderful
[b]Negative words = Agony, Awful, Horrible, Humiliate, Nasty, Painful, Terrible, Tragic
[c]Wealth words = Rich, Wealthy, Affluent, Prosperous, Well Off, Loaded, Fortune, Lucrative
[d]Poverty words = Poor, Poverty, Destitute, Needy, Impoverished, Broke, Bankrupt, Penniless
Note: the order of the target/valence pairing was randomised, meaning that for half of participants, incompatible blocks 4 & 5 preceded compatible blocks 2 & 3.

To quantify participants' implicit responses, we used the D Score summary measure (Greenwald, Nosek, & Banaji, 2003). On this measure, higher/lower scores indicate greater automatic associations between target groups and positive/negative concepts in evaluative ST-IATs, and between target groups and wealth/poverty in wealth ST-IATs. D Scores from ST-IATs display comparable psychometric properties to the more commonly used two-category IAT (Greenwald & Lai, 2020). We estimated the average split-half reliability of the valence and wealth ST-IATs to be 0.66[5] and 0.68, respectively (the valence ST-IAT figure combines data from Studies 1a and 1b). All implicit measures in the present manuscript were administered online via Inquisit Web software.

Demographics

In both studies demographic information (age, gender, race, and political orientation) was collected at the end of the experiment.

Results

For Study 1a we fitted a 2 (target race: Black, White) × 3 (target class: low, middle, high) repeated measures ANOVA predicting participants' D scores on the evaluative ST-IAT. For Study 1b we fitted separate 2 (target race: Black, White) × 2 (target class: low, high) independent

[5] These figures (and all split-half reliability figures reported in this paper) are based on average split-half correlations from 100 random splits of the ST-IAT data corrected according to the Spearman-Brown prophecy formula (Revelle & Condon, 2019).

samples Analyses of Variance (ANOVA) predicting D scores on both the evaluative and wealth ST-IATs. All analyses were conducted in R version 3.6.1 (R Core Team, 2019).

Evaluative ST-IATs

In both studies there was a significant main effect of targets' social class, Study 1a: $F(2,598) = 18.93$, $p < .001$, $\eta_p^2 = 0.02$, Study 1b: $F(1,516) = 5.27$, $p = 0.02$, $\eta_p^2 = 0.01$, with participants responding more positively to upper-class targets than lower-class targets. In Study 1a, participants responded more positively to upper-class targets than middle-class targets and to middle-class targets than lower-class targets, although this latter difference did not reach statistical significance (see Figure 2). By contrast, there were no significant main effects of race in either study: Study 1a, $F(1,299) = 2.07$, $p = 0.15$, $\eta_p^2 = 0.001$, Study 1b, $F(1,516) = 2.47$, $p = 0.12$, $\eta^2 = 0.005$, nor any significant race × class interactions: Study 1a, $F(2,598) = 0.28$, $p = 0.75$, $\eta_p^2 = 0.0003$, Study 1b, $F(1,516) = 0.58$, $p = 0.45$, $\eta_p^2 = 0.001$.

Wealth ST-IAT

In the wealth ST-IAT in Study 1b, there was again a main effect of target class, $F(1,518) = 23.72$, $p < 0.001$, $\eta_p^2 = 0.04$, with upper-class targets producing stronger relative associations with wealth than lower-class targets (see Figure 2). There was no significant effect of target race, $F(1,518) = 0.0008$, $p = 0.98$, $\eta_p^2 < 0.001$, and no significant race × class interaction, $F(1,518) = 3.13$, $p = 0.08$, $\eta_p^2 = 0.01$.

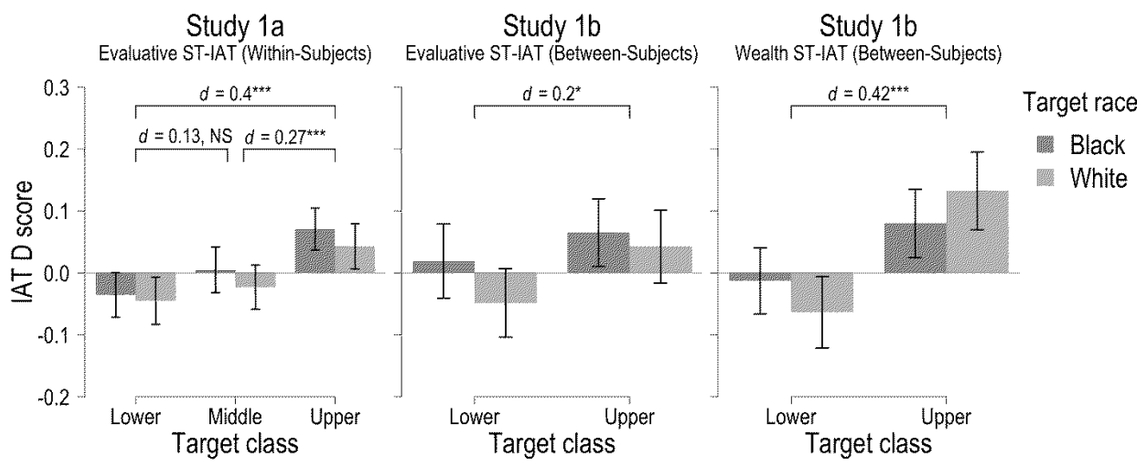


Figure 2. Mean IAT D scores by target group for Studies 1a and 1b. Bars indicate 95% confidence intervals. Cohens' *d* and statistical significance of t tests between social class groups collapsing across races are also reported (NS = not significant, * = $p < 0.05$, *** = $p < 0.001$).

Simulation-based power sensitivity analyses suggested that Analyses of Variances (ANOVAs) in both studies were well-powered to detect small main and interaction effects. Study 1a achieved 80% power to detect smaller effects ($\eta_p^2 = 0.005$) than Study 1b ($\eta_p^2 = 0.015$). See Supplementary Materials for details.

Discussion

In Studies 1a and 1b, participants evaluated targets with higher perceived incomes more positively than targets of lower perceived incomes. By contrast, evaluations were not significantly affected by target groups' race, nor did we observe any significant race × class interaction effects. These results are most consistent with the category dominance model, and diverge from previous findings regarding the effects of race and class on implicit bias (Mattan et al., 2019; Moore-Berg et al., 2017; though Mattan and colleagues observed a similar result in their third study). Those past studies, it is worth noting, did not hold perceived social class constant across races. By contrast, our Black and White target groups were pre-matched on explicit ratings of perceived incomes, and our wealth ST-IAT in Study 1b verified that automatic associations between target groups and wealth did not differ significantly across races. Our use of full-body target photographs may also have been a factor, elevating the influence of targets' bodies—a primary source of social class cues (e.g., Becker, Kraus, & Rheinschmidt-Same, 2017; Gillath, Bahns, Ge & Crandall, 2012; Schmid-Mast & Hall, 2004)—relative to the influence of targets' faces—a source of race cues—on evaluations. Both previous studies used stimuli which devoted a more equal share of visual space to cues of race and class.

Study 2

In Study 2 we tested participants' responses to targets varying more widely in terms of race (we incorporated Asian as well as Black and White targets[6]), as well as on social class, gender, and age. We also addressed whether the lack of pro-White/anti-Black implicit racial bias observed in Study 1 might have occurred simply due to our specific sampling population possessing little pro-White/anti-Black implicit bias (Studies 1a's and 1b's samples were largely

[6] The choice to include Asian rather than another race of targets was partly pragmatic, due to their availability within our photograph database, but was also informed by an interest in the potential for our majority Asian student samples to show greater racial bias if their racial ingroup were included as targets.

made up of female college students). To investigate this, we also measured participants on a traditional two-category Race IAT (Greenwald et al., 1998).

Toward a Target-Level Analysis: The Target D Score

Studying intersectionality encounters pragmatic limitations. For example, measuring evaluations of target groups displaying three different races (e.g., Asian, Black, and White), two genders (female vs. male), two levels of social class (high vs. low), and two levels of age (old vs. young) using the methods of Study 1 would require 24 separate experimental conditions. This method is inefficient, however, as it ignores systematic variation in implicit evaluations within target groups.

In Study 2 we developed a more efficient approach by quantifying implicit evaluations at the level of individual targets via *Target D Scores*. This measure relies on a similar logic to a standard ST-IAT D Score, but rather than measuring an individual participant's response to a target group in compatible vs. incompatible trials, Target D Scores measure an entire sample's response to an individual target in compatible vs. incompatible trials. This allows researchers to study systematic variation in implicit evaluations both between and within groups of targets, and thereby to more efficiently model the simultaneous effects of multiple simultaneously varying target-level variables.

A Data-Driven Approach to Person-Perception

We were *a priori* interested in implicit evaluations of targets varying in race, gender, social class, and age, as these categories are perceptible in most social interactions, and have been the focus of much of the previous work into implicit evaluative bias. However, we did not wish to presume in advance how participants would spontaneously perceive and categorize such complex targets. In recent work, Koch, Imhoff, Dotsch, Unkelbach, and Alves (2016) studied the content of social perceptions in a data-driven way. Rather than rating targets on pre-chosen traits, participants provided ratings of the similarity/dissimilarity of pairs of targets,, which were then subjected to Multidimensional Scaling (MDS, for a review, see Borg & Groenen, 2005) to identify the primary dimensions underlying participants' judgments. We used this method to ascertain whether indeed race, class, gender and age spontaneously shape implicit bias. Study 2 was pre-registered at https://aspredicted.org/87gw6.pdf.[7]

[7] We deviated from this pre-registration by predicting Target D Scores calculated according to the algorithm described below rather than logged response times between 300ms

Target Photographs

We selected 54 images (18 Asian[8], 18 Black, and 18 White targets) from a large database of 726 full-body target images (54 Asian female, 63 Asian Male, 115 Black female, 154 Black male, 140 White female, 200 White male). In addition to the images, the database contains 490,359 explicit ratings of the targets made by 3,311 US adults (1,875 women, 1031 men, 24 non-binary, 381 missing gender data, M_{age} = 23.8, SD_{age} = 8.6, 1,116 Asian, 1,089 White, 414 Latino, 117 Black, 575 other race or unreported) on 24 different personality and demographic traits selected as central to person perception. Traits measured were: warm (ICC = 0.23), competent (ICC = 0.31), honest/moral (ICC = 0.13), dominant (ICC = 0.16), submissive (ICC = 0.11), hard-working (ICC = 0.18), extraverted/enthusiastic (ICC = 0.15), reserved/quiet (ICC = 0.12), sympathetic/warm (ICC = 0.15), critical/quarrelsome (ICC = 0.07), dependable/self-disciplined (ICC = 0.21), disorganized/careless (ICC = 0.20), calm/emotionally stable (ICC = 0.14), anxious/easily upset (ICC = 0.08), open to new experiences/complex (ICC = 0.15), conventional/uncreative (ICC = 0.09), attractive (ICC = 0.33), income (ICC = 0.39), education (ICC = 0.27), occupational prestige (ICC = 0.39), subjective socioeconomic status (ICC = 0.43), age (ICC = 0.72), political orientation (ICC = 0.26), and race (measured via a multiple choice categorical response; ICCs for dummies indicating Asian, Black, and White categorizations = 0.87, 0.90, and 0.80, respectively).

For each race (Asian, Black, and White), we selected 9 female and 9 male targets varying in social class and age. There was some minor non-orthogonality between target-level variables (maximum r = 0.15, see Table 2). However, we were able to control for such imbalances by estimating effects of targets' race while controlling for their precise levels of perceived social class, and *vice versa*, as is done in conjoint experimental designs with multivariate analyses (Hainmueller, Hopkins, & Yamamoto, 2014).

---

and 10,000ms. This deviation reflects our evolving understanding of how best to model and analyze ST-IAT data at the individual target level, and had only a minor impact on conclusions (see Supplementary Materials).

[8] All Asian targets used in the present manuscript appear subjectively to be of prototypically East Asian appearance, though it is a limitation of the present manuscript that neither our data nor the Chicago Face Database norming data relied upon for the Studies 3 & 4 targets distinguishes between different sub-categories within the overarching category of 'Asian.'

Table 2
Descriptive statistics of targets chosen for Study 2

Correlations	1.	2.	3.	4.	5.	6.
1. Asian categorization						
2. Black categorization	-0.49					
3. White categorization	-0.52	-0.47				
4. Female[a]	-0.01	-0.01	0.03			
5. Age	-0.02	-0.01	0.03	-0.04		
6. SES[b]	0.15	-0.15	-0.01	-0.002	-0.02	
Descriptives						
M(*SD*) Overall	0.33(0.47)	0.31(0.45)	0.32(0.42)	0.5(0.5)	43.6(12.93)	0(1)
M(*SD*) Asian Females	0.97(0.03)	0.01(0.03)	0.02(0.05)	1(0)	40.59(11.34)	0.18(0.67)
M(*SD*) Asian Males	0.99(0.03)	0(0)	0.01(0.02)	0(0)	46.05(13.52)	0.24(0.87)
M(*SD*) Black Females	0.01(0.02)	0.91(0.15)	0.08(0.08)	1(0)	44.87(13.42)	-0.21(1.08)
M(*SD*) Black Males	0.01(0.03)	0.95(0.05)	0.01(0.02)	0(0)	41.6(14.35)	-0.15(1.20)
M(*SD*) White Females	0(0)	0.01(0.02)	0.89(0.15)	1(0)	43.84(13.38)	0.02(1.01)
M(*SD*) White Males	0(0)	0.01(0.02)	0.9(0.1)	0(0)	44.64(14.37)	-0.08(1.27)

[a] Female is a manually coded dummy (1 = Female, 0 = Male)
[b] SES is a *z*-scored average of *z*-scored ratings on income, education, occupational prestige, and subjective SES

Participants and Procedure

Participants were 371 undergraduate students who participated for course credit (281 women, 66 men, 1 non-binary, 23 missing gender data, M_{age} = 20.44, SD_{age} = 2.5, 194 Asian, 93 White, 32 Latino, 6 Black, 16 other race, 30 missing race data).

ST-IATs

Participants completed three separate evaluative ST-IATs, following the procedures described above. The three ST-IATs used as target stimuli the 18 Asian, 18 Black, and 18 White targets, respectively, and were presented in a randomized order.

Race IAT

Participants also completed a two-category Race IAT using black-and-white partial face images of Black and White targets as stimuli.[9] This involved a similar procedure to the ST-IAT, except that in test trials the labels "White American" and "Black American" appeared on opposite sides of participants' screens, alongside the labels "Good" and "Bad." Participants categorized positive words or White faces via a single computer key and negative words or Black faces via an alternative key (compatible trials), or categorized positive words or Black faces via a single computer key, and negative words or White faces via an alternative key (incompatible trials). We computed D scores according to Greenwald and colleagues' (2003) algorithm, with higher D scores (split-half reliability = 0.75) indicating anti-Black implicit bias. The order of the ST-IATs and the Race IAT was randomly counter-balanced.

[9]We used the "Racism IAT" available from Millisecond.com (https://www.millisecond.com/download/library/iat/raceiat/)

Difference Ratings

Following the implicit measurement tasks, participants were presented with 60 randomly selected pairs of the 54 targets and asked to indicate "how different or similar are these people" on 0-100 sliders ranging from "Very Similar" to "Very Different." This resulted in an average of 14.8 ratings (*SD* = 3.57) each of 1,431 possible target pairs (ICC = 0.29).

Demographics

Finally, participants reported demographic information, including subjective SES measured via the MacArthur ladder measure (Adler, Epel, Castellazzo, & Ickovics, 2000).

Results

Multi-Dimensional Scaling

We computed the mean perceived difference between each of the 1,431 unique target pairs and subjected the resulting distance matrix to MDS using the majorization approach assuming an interval scale (SMACOF; De Leeuw & Mair, 2009). A five-dimension solution proved to be the most parsimonious solution providing good fit (scaling stress of 0.116 and r^2 of 0.79; stress of 0.15 or less is generally considered acceptable, Dugard, Todman, & Staines, 2010; see Supplemental Materials for more information).

We calculated correlations between targets' scores on each MDS dimension and the explicit trait ratings of each target (Table 3). The first dimension correlated strongly with targets' subjective SES (r = 0.91),[10] the second with both categorization as Asian (r = -0.81) and categorization as Black (r = 0.79),[11] the third with categorization as White (r = 0.78), the fourth with categorization as Female (r = 0.81), and the fifth with targets' age (r = 0.91). These results suggested that targets were spontaneously perceived as varying based on core demographic variables: social class, race, gender, and age.

Table 3
Target-level correlations between targets' MDS-derived dimension scores and mean explicit trait ratings. Correlations weaker than 0.2 are not displayed.

	MDS Dimensions

[10] In the original MDS solution Dimension 1 correlated negatively with measures of social class. We have reversed its scores throughout the manuscript for ease of interpretation. This has no effect on any of the reported results beyond reversing their direction.

[11] The fact that two race dimensions emerged—one (Dimension 2) separating Asian and Black targets, and the other (Dimension 3) separating White from Asian and Black targets—is sensible given that two linear dimensions are necessary to separate the three racial groups represented.

	1	2	3	4	5
Subjective SES	0.91				
Occupational Prestige	0.89				
Education	0.85				
Income	0.81		0.22	-0.24	0.24
Attractiveness	0.8				-0.31
Competence	0.79				
Disorganized/Careless	-0.74				-0.33
Dominant	0.73	0.25			
Dependable/Self disciplined	0.67			-0.21	0.21
Calm/Emotionally stable	0.61			-0.22	
Submissive	-0.6	-0.33			-0.23
Hard working	0.55		-0.26	-0.31	0.33
Extraverted/Enthusiastic	0.52	0.33		0.28	-0.3
Reserved/Quiet	-0.51	-0.27		-0.34	0.24
Asian[a]	0.2	-0.81	-0.34	-0.3	
Black[a]		0.79	-0.47		
Liberal		0.62	-0.27		-0.37
Conventional/Uncreative	-0.33	-0.38			0.29
White[a]			0.78	0.45	
Honest/Moral			-0.34		
Critical/Quarrelsome	0.22		0.22		
Female[b]	0.26		-0.43	0.81	
Anxious/Easily upset	-0.41			0.51	
Sympathetic		0.22	-0.24	0.3	
Warmth		0.23	-0.2	0.25	
Age	-0.26				0.91
Open To New Experience/Complex	0.44	0.31			-0.5

[a]Asian, Black, and White represent means of dummies indicating categorical categorization as appearing to be of each respective race
[b]Female represents a manually coded dummy (1 = female target, 0 = male target)

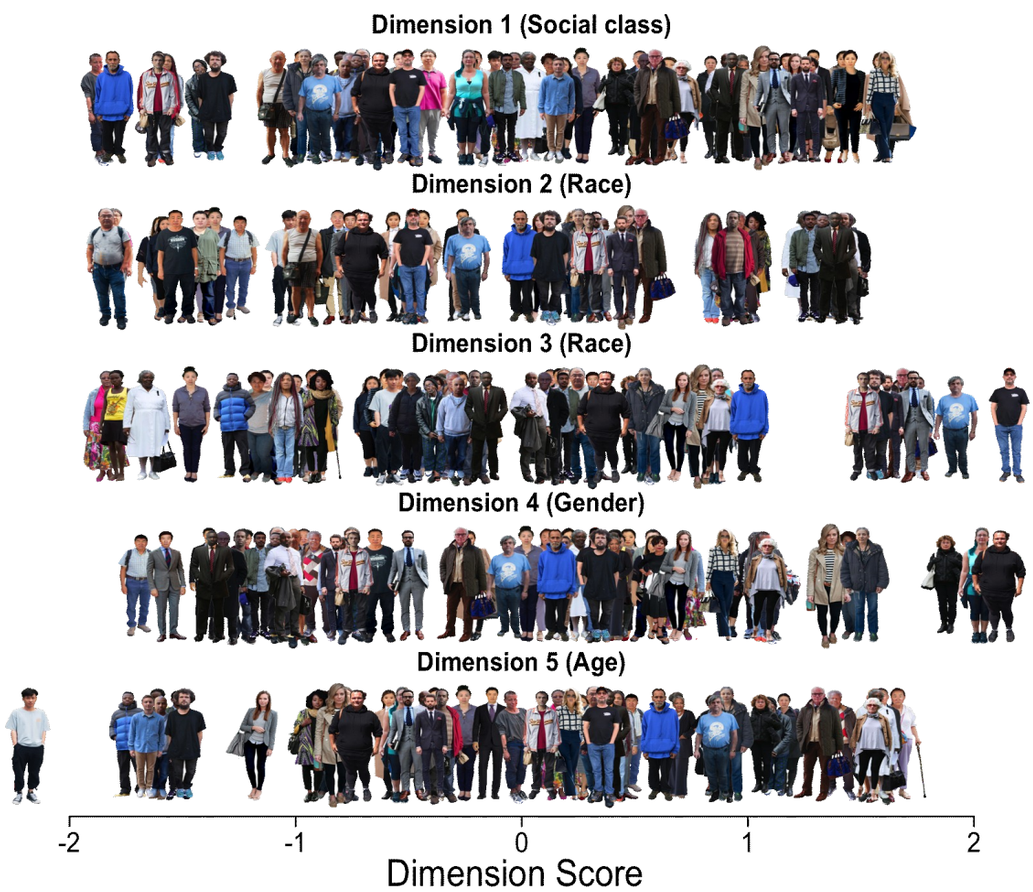


Figure 3. Study 2 targets arranged according to their scores on each of the 5 spontaneously emerging dimensions underlying relative similarity/dissimilarity judgments.

Calculating and Validating Target D Scores

To identify the optimal scoring algorithm for Target D Scores, we compared different algorithms with regard to both their internal reliability, as indexed by split-half reliability estimates, and their convergent validity, as indexed by the strength of their relationships with target-level characteristics shown in previous research and the present manuscript to be associated with implicit evaluations (see Supplementary Materials for more details). The scoring algorithm for Target D Scores producing the optimal results[12] involved (a) identifying all raw response times toward a specific target in ST-IATs trials, including error trials, (b) eliminating response times below 100 milliseconds and above 4000 milliseconds (12% and 0.02% of trials, respectively), (c) penalizing error trials, in which the wrong computer key was pressed in response to a target (6.5% of all trials) by replacing their latency with participants' individual mean response latency in compatible/incompatible trials plus 600ms, (d) taking the natural log of each of the remaining response times, (e) computing a difference score for each target representing the mean logged response time in incompatible trials minus the mean logged response time in compatible trials. To aid interpretability, these difference scores were then divided by the overall standard deviation of all logged response times between 100 and 4000 milliseconds. Higher/lower Target D Scores indicate that participants responded relatively faster/ slower to a target in compatible compared to incompatible trials.

To test the utility of modelling implicit evaluations at the target level, we calculated Target D Scores for each of the 69 unique targets used in Study 1 (Study 1a split-half reliability = 0.57, Study 1b split-half reliability = 0.66). Not only was there was a significant positive raw correlation between target's mean income ratings and Target D Scores, $r(67) = 0.35$, $p = .003$, income ratings remained a significant predictor of Target D Scores in a multiple regression controlling for targets' group membership, $\beta = 0.91 (SE = 0.36)$, $t(58) = 2.58$, $p = 0.013$, $\eta_p^2 = 0.10$.[13] Thus, even within target groups, targets judged to have higher incomes produced higher

[12] This algorithm also produced the highest internal reliability, so would have been chosen if internal reliability were the only criterion.

[13] β here represents a standardized slope, with Target D Scores and targets' mean income ratings both z-scored. Target group membership was entered into the model as a categorical predictor.

Target D Scores. This systematic variation had previously been obscured within Study 1's target group-level analyses.[14]

Predicting Target D Scores from Multi-Dimensional Scaling Dimensions

To assess the relationship between each MDS dimension and implicit bias, we fit multiple regression models predicting the Target D Scores (split-half reliability = 0.71) of each of the 54 Study 2 targets from each of the multi-dimensional scaling dimensions. Results (Table 4) revealed significant associations between Target D Scores and Dimension 1 (Social class), $\hat{\beta}\left(SE_{\hat{\beta}}\right)$ = 0.06(0.02), t(48) = 4.07, $p < .001$, $\eta_p^2 = 0.26$, with bias favouring higher class over lower class targets. We also observed a significant effect of Dimension 3 (Race), $\hat{\beta}\left(SE_{\hat{\beta}}\right)$ = -0.04(0.02), t(48) = -2.71, $p = .01$, $\eta_p^2 = 0.13$, with bias favouring Asian and Black targets over White targets, and Dimension 4 (Gender), $\hat{\beta}\left(SE_{\hat{\beta}}\right)$ = 0.06(0.02), t(48) = 3.89, $p < .001$, $\eta_p^2 = 0.24$, with bias favouring female targets over male targets.

In a second model, we included each two-way interaction between dimensions. This significantly improved model fit, $F(9,39) = 3.43$, $p = 0.003$. Main effects of Dimensions 1 (Social class), 3 (Race), and 4 (Gender) each remained significant (see Table 4), but the effects of Dimensions 1 and 4 were qualified by a significant two-way interaction, $\hat{\beta}\left(SE_{\hat{\beta}}\right)$ = 0.06(0.02), t(39) = 4.29, $p < .001$, $\eta_p^2 = 0.32$, with the positive interaction slope suggesting a stronger effect of the social class dimension among female targets (higher scores on Dimension 4 = female targets). Including three-way interactions between dimensions did not improve model fit, $F(7,32) = 0.48$, $p = 0.84$.

A simulation-based power sensitivity analyses suggested that our linear regressions achieved 80% power to detect main effects of approximately $\eta_p^2 = 0.10$ and two-way interaction effects of approximately $\eta_p^2 = 0.08$ (see Supplementary Materials for details).

Table 4
Study 2 results of multiple regressions predicting Target D Scores

[14] Target-level variation in implicit evaluations can also be studied via more complex models predicting raw or logged response times (e.g., Thiem et al., 2019; Mattan et al., 2019). We discuss Target D Scores' advantages over these methods in our general discussion.

	Multi-Dimensional Scaling dimensions							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	η_p^2	r^2	$\hat{\beta}(SE_{\hat{\beta}})$	p	η_p^2	r^2
(Intercept)	0.019(0.015)	0.216			0.019(0.012)	0.139	NA	
Dimension 1 (Social class[a])	0.061(0.015)	<.001	0.257		0.062(0.013)	<.001	0.364	
Dimension 2 (Race[b])	0.002(0.015)	0.871	0.001		-0.001(0.014)	0.929	0.005	
Dimension 3 (Race[c])	-0.041(0.015)	0.009	0.132		-0.037(0.014)	0.009	0.163	
Dimension 4 (Gender[d])	0.059(0.015)	<.001	0.24		0.059(0.013)	<.001	0.334	
Dimension 5 (Age)	-0.008(0.015)	0.602	0.006		-0.013(0.013)	0.342	0.005	
Dimension 1 × Dimension 2					-0.023(0.017)	0.171	0.047	
Dimension 1 × Dimension 3					0.01(0.015)	0.526	0.01	
Dimension 1 × Dimension 4					0.063(0.015)	<.001	0.321	
Dimension 1 × Dimension 5					-0.024(0.018)	0.173	0.047	
Dimension 2 × Dimension 4					-0.015(0.015)	0.335	0.024	
Dimension 2 × Dimension 5					0.014(0.013)	0.279	0.03	
Dimension 3 × Dimension 4					0.002(0.02)	0.928	<.001	
Dimension 3 × Dimension 5					-0.025(0.013)	0.056	0.09	
Dimension 4 × Dimension 5					0.002(0.015)	0.896	<.001	
				0.453				0.695
	Explicit target ratings							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	η_p^2	r^2	$\hat{\beta}(SE_{\hat{\beta}})$	p	η_p^2	r^2
(Intercept)	-0.026(0.031)	0.41			-0.011(0.034)	0.751		
Social class[e]	0.038(0.016)	0.02	0.108		-0.028(0.025)	0.274	0.142	
Asian[f]	-0.041(0.038)	0.283	0.024		-0.063(0.049)	0.206	0.039	
White[f]	-0.054(0.038)	0.156	0.041		-0.069(0.048)	0.157	0.087	
Female[f]	0.153(0.031)	<.001	0.339		0.127(0.048)	0.011	0.419	
Age[g]	-0.023(0.016)	0.147	0.043		0.019(0.028)	0.498	0.047	
Social class × Asian					0.007(0.04)	0.865	0.001	
Social class × White					0.058(0.032)	0.072	0.081	
Social class × Female					0.096(0.029)	0.002	0.225	
Social class × Age					-0.016(0.016)	0.32	0.025	
Asian × Female					0.039(0.069)	0.579	0.008	
Asian × Age					-0.022(0.036)	0.534	0.01	
White × Female					0.023(0.067)	0.73	0.003	
White × Age					-0.068(0.035)	0.061	0.087	
Female × Age					-0.019(0.028)	0.503	0.012	
				0.423				0.632

Note: Black is the reference category for race contrasts in the Explicit target ratings models
[a]Higher scores on Dimension 1 = higher perceived social class
[b]Higher scores on Dimension 2 = Black, lower scores = Asian
[c]Higher scores on Dimension 3 = White
[d]Higher scores on Dimension 4 = Female
[e]Social class = a *z*-scored composite of targets' perceived income, subjective SES, occupational prestige, and education
[f]Asian, White, and Female are dummy variables indicating Asian, White, and Female targets
[g]Age is targets' perceived age, *z*-scored

Predicting Target D Scores from Explicit Target Ratings

Next, we predicted Target D Scores directly from explicit ratings of target's social class (the average of *z*-scored mean ratings of subjective SES, occupational prestige, education, and income, Cohen's $\alpha = 0.98$), binary indicators of Asian race, White race, and female gender[15], and *z*-scored mean ratings of targets' age. We observed significant effects of targets' perceived social class, $\hat{\beta}(SE_{\hat{\beta}}) = 0.04(0.02)$, t(48) = 2.45, $p = .02$, $\eta_p^2 = 0.11$, with bias favouring higher class over lower class targets, and targets' gender, $\hat{\beta}(SE_{\hat{\beta}}) = 0.15(0.03)$, t(48) = 4.96, $p < .001$, $\eta_p^2 = 0.34$, with bias favouring female over male targets. In contrast to MDS dimensions, there were no

[15] Targets were coded as Asian, Black, and White if they were categorized as such by raters > 90% of the time. Gender was manually coded by the lead author.

significant effects of target race, suggesting that the previously observed effect of Dimension 3 may have occurred due to its overlap with (see Table 4).

Next, we included each two-way interaction between predictors (except between the two race indicators), which again significantly improved model fit, $F(9,39) = 2.46$, $p = 0.02$. Target gender remained a significant predictor, but was qualified by a significant two-way interaction with target social class, $\hat{\beta}(SE_{\hat{\beta}}) = 0.10(0.03)$, $t(39) = 3.37$, $p = .002$, $\eta_p^2 = 0.23$. The pattern of this interaction suggested a strong effect of social class with regard to female targets, with upper-class female targets eliciting positive evaluations, but little effect of social class for male targets (see Figure 4).

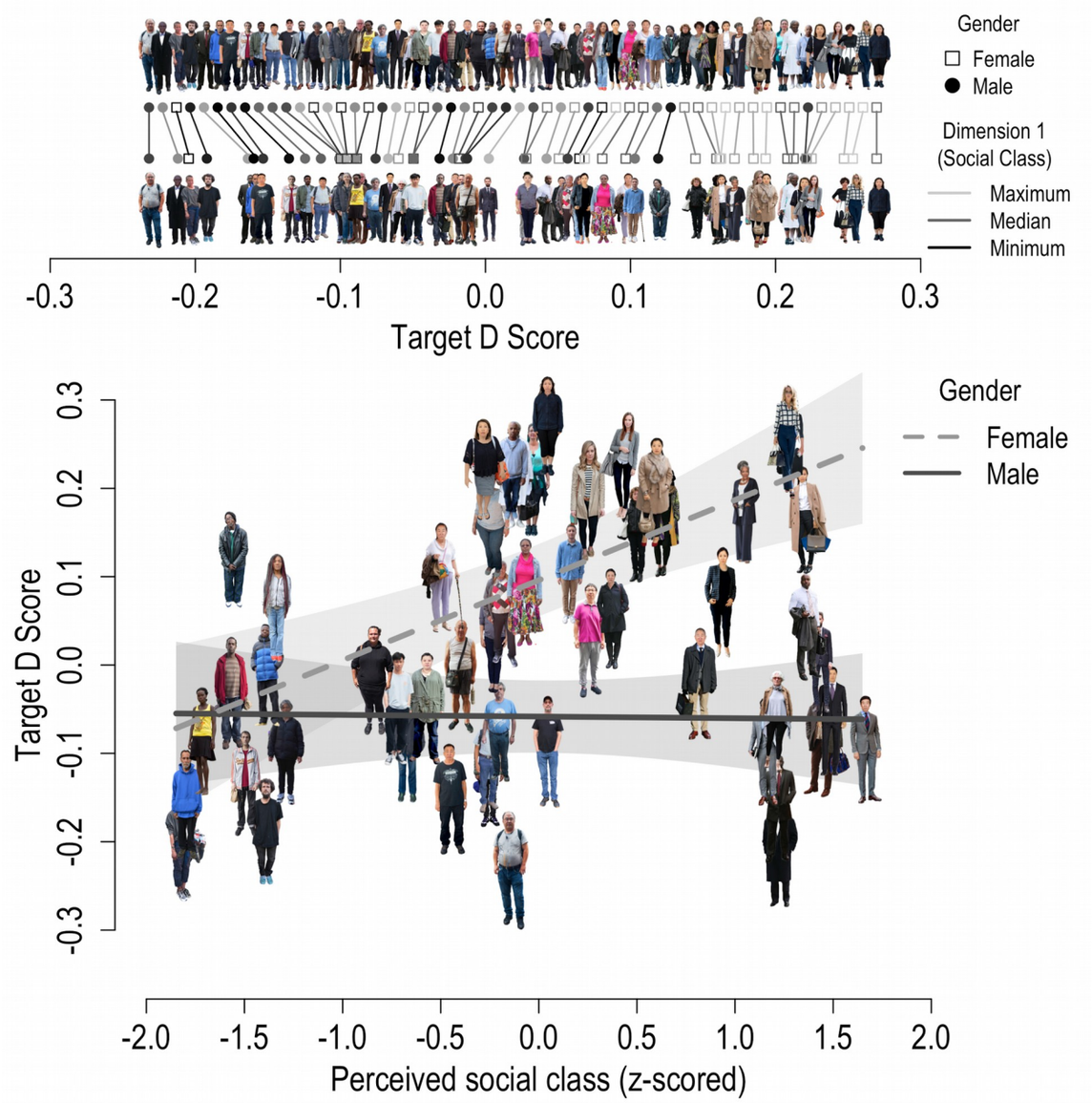


Figure 4. The top panel displays targets ordered by their Target D Scores (the row above) and arranged according to their exact Target D Scores (the row below). The bottom panel displays the interaction between targets' gender and perceived social class (a z-scored composite of

targets' perceived income, subjective SES, occupational prestige, and education) in predicting Target D Scores.

Race IAT Results

A single sample t-test on participants' D Scores suggested that the sample exhibited significant anti-Black/pro-White bias from the traditional two-category Race IAT (M = 0.30, SD = 0.4), $t(367) = 14.11$, $p < .001$, Cohen's $d = 0.75$, 95% CI = [0.26,0.34].

Discussion

In Study 2 we observed implicit evaluations to be largely driven by an interaction between targets' gender and social class, with upper-class female targets eliciting especially positive evaluations. This interaction emerged regardless of whether we predicted evaluations from MDS dimension scores or from explicit ratings of targets. By contrast, target race yielded more equivocal effects, with an apparent anti-White bias emerging from MDS Dimension scores, but failing to emerge when Target D Scores were predicted from targets' explicit race categorizations. Target age exhibited no significant effects.

These results do not align neatly with theories of compounding bias or the category dominance model. Theories of compounding bias are consistent with especially positive evaluations of upper-class female targets, but offer little explanation as to why we observed little evidence of anti-Black bias in our ST-IATs (if anything, we observed weak evidence of anti-White bias). Meanwhile, the category dominance model can make sense of equivocal or absent race and age effects, as well as the relatively large effect of target gender. However, it does not provide an easy explanation of interaction effects, which require at least some participants to be sensitive to multiple categories at once.[16]

Additionally, despite showing little evidence of pro-White/anti-Black bias within ST-IATs, our sample displayed a robust pro-White/anti-Black bias on the traditional Race IAT. This suggests that the ST-IAT results cannot be explained as being simply a function of sampling bias.

Study 3

[16] If each participant's responses were dominated by a single category, gender-biased participants should produce a main effect of gender, and class-biased participants should produce a main effect of class. Such participants could collectively display main effects of both class and gender, but should not, in theory, display an interaction between the two categories

In Study 3, we incorporated a number of methodological improvements. First, we exerted tighter experimental control over our target stimuli, swapping the same target faces onto multiple target bodies, thus holding constant body shape and clothing across target race categories, and holding constant facial features exactly constant across social class categories. Second, all racial groups were presented together within ST-IAT tasks. In Studies 1 and 2, targets of different races were presented within separate ST-IAT tasks, raising the possibility that participants may have used recoding strategies that suppressed implicit racial biases (e.g., Meissner & Rothermund, 2013). Third, we investigated whether the use of full-body targets in Studies 1 and 2 had elevated the influence of targets' bodies—a primary source of social class cues (e.g., Becker et al., 2017; Gillath et al., 2012; Schmid-Mast & Hall, 2004)—relative to the influence of targets' faces—likely the primary source of race cues—due to targets' bodies dominating the visual space of stimuli. To probe this, in Study 3 we presented targets both as upper-body images from the waist up (Study 3a) and as full-body images (Study 3b).

Stimuli Development

Faces

We selected 24 unique faces from the Chicago Face Database (CFD; Ma, Correll, & Wittenbrink) varying in race (8 Asian, 8 Black, 8 White), gender (12 male, 12 female), and age (12 old, 12 young), with two faces chosen to represent each race/age/gender subgroup. Based on CFD norming data, there were no significant differences among the chosen faces in perceived attractiveness or racial prototypicality between race, age, or gender groups (all $F < 1.27$, all $p > 0.27$)., nor differences in female or male categorization between race or age groups (all $F < 0.002$, all $p > 0.98$), nor significant differences in Asian, Black, or White categorization between gender or age groups (all $F < 0.02$, all $p > 0.89$), and no significant differences in perceived age between race or gender groups (all $F < 0.03$, all $p > 0.97$).

Bodies

The 24 bodies we selected varied in terms of gender (12 male, 12 female), age (12 old, 12 young), and perceived socioeconomic status (12 high-SES, 12 low-SES), with three bodies chosen to represent each gender/age/SES subgroup. Based on explicit ratings[17] in which each body was rated by an average of 84.1 raters (SD = 111.0), there were no significant differences in

[17] Ratings of each body were made with different faces attached to each body, rendering these data only a rough guide to the specific influence of the bodies.

perceived attractiveness between race, age, or gender groups (all $F < 2.80$, all $p > 0.10$), no significant differences in perceived age between gender or SES groups (all $F < 2.14$, all $p > 0.15$), and no significant differences in perceived SES or income between gender or age groups (all $F < 0.64$, all $p > 0.43$). Unavoidably, due to the strong correlation between ratings of attractiveness and subjective SES in the data ($r = 0.53$), there was a significant difference in perceived attractiveness between SES groups, with the high-SES bodies ($M = 53.9$, $SD = 10.4$) rated significantly more attractive than the low-SES bodies ($M = 30.6$, $SD = 7.6$), $F(1,22) = 39.3$, $p < 0.001$.

Attaching Faces to Bodies

We used Adobe Photoshop software to attach each of the 6 faces to each of the 6 bodies within each age/gender subgroup. This resulted in 144 total stimuli, which were then assembled into six target groups. Each group contained 8 Asian, 8 Black, and 8 White targets, 12 female and 12 male targets, 12 young and 12 old targets, and 12 high-SES and 12 low-SES targets (see Figure 5). See Supplementary Materials for more details.

Figure 5. The 24 faces and 24 bodies combined to create 144 unique targets arranged into six groups in which each face and body appears once. Both upper-body presentation (Study 3a) and full-body presentation (Study 3b) are displayed.

Participants and Procedure

Participants for Study 3a (N = 871, 591 women, 223 men, 11 non-binary, 46 missing gender data, M_{age} = 23.0, SD_{age} = 8.0, 411 Asian, 253 White, 77 Latino, 26 Black, 30 other race, 39 missing race data) and Study 3b (N = 656, 489 women, 149 men, 7 non-binary, 11 missing gender data, M_{age} = 20.83, SD_{age} = 2.8, 364 Asian, 145 White, 84 Latino, 10 Black, 36 Other race, 17 missing race data) were undergraduate students who participated for course credit. We excluded ST-IAT data from five participants in Study 3b who experienced technical issues during the ST-IAT task resulting in mean response times that were unreasonably large (> 3000ms). Study 3a was pre-registered at https://aspredicted.org/bv4jy.pdf[18] Study 3b was pre-registered at https://aspredicted.org/qz5yu.pdf.[19]

Single Target IATs (ST-IATs)

After providing informed consent, participants were randomly assigned to one of the six target groups, and completed two consecutive ST-IATs containing their target group as stimuli following the procedures described above.[20] In Study 3a participants viewed targets in upper-body presentation; in Study 3b participants viewed targets in full-body presentation.

---

[18] After the original planned sample size was reached in Study 3a (N = 379), the split-half reliability of the Target D Scores was so low (0.37) that we decided to collect additional data, and re-pre-registered the study at https://aspredicted.org/cr938.pdf. At this point we also made some minor changes to the study design, omitting similarity/difference ratings of pairs of targets and the Symbolic Racism Scale, and adding explicit ratings scales of targets' attractiveness, competence, political orientation, and photo blurriness. These changes had minor effects on the conclusions of the study (see Supplementary Materials for more information).

[19] We again deviated slightly from each of these pre-registrations as a result of our evolving understanding of how best to model and present our results. See Supplementary Materials for more details.

[20] We included two ST-IATs because in Study 3 there were 24 targets per ST-IAT, compared with 8 and 18 targets per ST-IAT in Studies 1 and 2. We therefore wanted to increase the number of trials for each target.

Difference Ratings

In Study 3a we initially measured similarity/difference ratings of pairs of targets to confirm that targets' race, gender, social class, and age would again emerge as the primary spontaneous dimensions underlying such judgments. Following Study 3a's initial data collection (see footnote 15), we considered this to be sufficiently established, and omitted the difference ratings from the additional data collected for Study 3a and from Study 3b (see Supplementary Materials for details).

Explicit Ratings of Targets

Participants also rated their 24 targets via 0-100 sliders on perceived gender (ICCs = 0.89, 0.87 in Studies 3a and 3b, respectively), race (three separate sliders measuring perceptions of targets as Asian, ICCs = 0.87, 0.86, Black, ICCs = 0.91, 0.89, and White, ICCs = 0.85, 0.84) social class (ICCs = 0.55, 0.59), and age (ICCs = 0.61, 0.58). We also measured perceptions of targets' warmth (ICCs = 0.22, 0.21), extroversion (ICCs = 0.11, 0.14), attractiveness (ICCs = 0.20, 0.22), competence (ICCs = 0.30, 0.31), political orientation (ICCs = 0.26, 0.27), and photo blurriness (ICCs = 0.70, 0.10) as factors we considered might be predictive of implicit evaluations.

Demographics

Participants reported the same demographic information as in Study 2.

Results

Manipulation Checks

As tests of our manipulations, we inspected correlations between participants' explicit ratings of the targets and our *a priori* categorizations of targets as male, Asian, Black, White, high-SES, and older/younger. Correlations indicated that each variable was manipulated as intended (see bolded correlations in Table 5). There was also relatively little non-orthogonality between these variables, with the exception of a correlation between SES and age ratings (Study 3a: $r = 0.15$, Study 3b: $r = 0.12$). To control for this non-orthogonality, we again used target-level analyses modelling targets' social class and age as continuous variables.

Table 5

Correlations between our *a priori* categorizations and participants' subjective ratings of targets

	Female ratings	Asian ratings	Black ratings	White ratings	SES ratings	Age ratings
Study 3a						
Asian ratings	0.01					
Black ratings	0.004	-0.489				

White ratings	-0.017	-0.464	-0.545			
SES ratings	-0.028	0.074	-0.028	-0.034		
Age ratings	-0.035	0.078	0.012	-0.096	0.151	
Female categorization	**0.998**	0.01	-0.004	-0.009	-0.025	-0.032
Asian categorization	0	**0.998**	-0.495	-0.456	0.071	0.073
Black categorization	0.007	-0.493	**0.999**	-0.54	-0.031	0.004
White categorization	-0.008	-0.505	-0.504	**0.996**	-0.039	-0.077
SES categorization	-0.003	0.001	0.002	0.005	**0.911**	0.039
Age categorization	0	0.004	0.005	-0.018	0.127	**0.947**
Study 3b						
Asian ratings	0.021					
Black ratings	-0.002	-0.493				
White ratings	-0.028	-0.472	-0.533			
SES ratings	0.018	0.073	0.03	-0.087		
Age ratings	0.055	0.131	-0.039	-0.102	0.12	
Female categorization	**0.997**	0.017	-0.012	-0.014	0.021	0.063
Asian categorization	0.004	**0.997**	-0.494	-0.467	0.069	0.124
Black categorization	0.01	-0.497	**0.999**	-0.527	0.027	-0.046
White categorization	-0.014	-0.5	-0.504	**0.994**	-0.096	-0.079
SES categorization	-0.003	0.005	0.002	0.008	**0.955**	0.088
Age categorization	-0.005	0.003	0.006	-0.022	0.033	**0.927**

Note: intercorrelations between dummy variables are omitted because these are all necessarily $r = 0$, except the race dummies which correlate at $r = 0.5$

Predicting Target D Scores

Because the same faces and bodies were shared by multiple targets, we fitted cross-classified hierarchical linear models (HLMs) predicting Target D Scores (Study 3a split-half reliability = 0.54, Study 3b split-half reliability = 0.59), and included in each model random intercepts for the 24 unique target faces and 24 unique target bodies (see Table 6). For all HLMs we used the R packages lme4 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, Christensen, 2017).

Study 3a. We first predicted Target D Scores from *z*-scored mean ratings of targets' subjective SES, *z*-scored mean ratings of targets' age, and dummy variables indicating Asian race, White race, and female gender. We observed significant effects of target race, with both Asian targets, $\hat{\beta}\left(SE_{\hat{\beta}}\right) = 0.10(0.02)$, t(18.85) = 4.30, $p < .001$, $r^2_{sp} = 0.13$[21], and White targets, $\hat{\beta}\left(SE_{\hat{\beta}}\right) = 0.09(0.02)$, t(18.69) = 4.07, $p < .001$, $r^2_{sp} = 0.12$, evaluated more positively than Black

[21] r^2_{sp} refers to semi-partial r^2 values (Edwards, Muller, Wolfinger, Qaqish, & Schabenberger, 2008) computed using the standardized generalized variance approach with the *r2glmm* R package (Jaeger, 2017).

targets (for the simultaneous addition of both race dummies Δr^{2} [22] = 0.07). There was no significant difference between evaluations of Asian and White targets, t(18.97) = -0.24, p = 0.81. Female targets were also evaluated more positively than male targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.20(0.02), t(13.36) = 8.71, $p < .001$, r^2_{sp} = 0.49. Neither targets' social class nor age exhibited significant unique effects on implicit evaluations. (see Table 6).

In a second model, we added two-way interactions between each target-level factor. Doing so did not significantly improve model fit, $\chi^2(9) = 7.53$, $p = 0.58$, so we relegate these results to Supplementary Materials. Finally, in a third model, we tested if the effects observed in our initial model were robust to controlling for targets' z-scored mean ratings on perceived warmth, extroversion, attractiveness, competence, political liberalism, and photograph blurriness. In this model target gender remained a significant predictor, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.20(0.02), t(26.64) = 6.23, $p < .001$, r^2_{sp} = 0.31, but all other target level variables were non-significant (See Table 6).

Table 6
Results from hierarchical linear models in Study 3a and Study 3b

	Study 3a (upper-body targets)							
	Model 1				Model 3			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	$r^{2\,a}_{sp}$	SD	$\hat{\beta}(SE_{\hat{\beta}})$	p	$r^{2\,a}_{sp}$	SD
Fixed effects								
(Intercept)	-0.129(0.021)	<.001			-0.107(0.032)	0.002		
Social class	0.007(0.011)	0.569	0.004		-0.026(0.045)	0.563	0.003	
Asian	0.096(0.022)	<.001	0.127		0.075(0.038)	0.059	0.032	
White	0.091(0.022)	<.001	0.115		0.041(0.062)	0.514	0.004	
Female	0.2(0.023)	<.001	0.488		0.203(0.033)	<.001	0.306	
Age	0.006(0.011)	0.598	0.003		0(0.017)	0.995	<.001	
Warmth					-0.004(0.022)	0.851	<.001	
Extroversion					0.003(0.018)	0.869	<.001	
Attractiveness					0.023(0.024)	0.334	0.009	
Competence					0.016(0.049)	0.739	<.001	
Liberal					-0.043(0.029)	0.143	0.019	
Blurry					0.016(0.013)	0.249	0.019	
			0.534				0.536	
Random effects								
Face				0.007				0.015
Body				0.034				0.042
Residual				0.107				0.106
	Study 3b (full-body targets)							
	Model 1				Model 3			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	$r^{2\,a}_{sp}$	SD	$\hat{\beta}(SE_{\hat{\beta}})$	p	$r^{2\,a}_{sp}$	SD
Fixed effects								
(Intercept)	-0.152(0.028)	<.001			-0.123(0.029)	<.001		
Social class	0.044(0.016)	0.01	0.12		0.022(0.042)	0.594	0.002	
Asian	0.101(0.027)	0.001	0.108		0.06(0.038)	0.117	0.021	

[22] Δr^2 refers to refers to differences in full model r^2 values computed using the standardized generalized variance approach with the *r2glmm* R package (Jaeger, 2017) between full models and models with predictors removed.

White	0.092(0.026)	0.003	0.091		0.037(0.061)	0.547	0.003
Female	0.232(0.033)	<.001	0.491		0.237(0.033)	<.001	0.349
Age	-0.009(0.016)	0.585	0.005		-0.006(0.017)	0.74	0.001
Warmth					-0.016(0.022)	0.487	0.004
Extroversion					-0.016(0.015)	0.297	0.011
Attractiveness					0.035(0.026)	0.186	0.016
Competence					-0.012(0.043)	0.772	<.001
Liberal					-0.022(0.028)	0.425	0.006
Blurry					-0.044(0.012)	<.001	0.131
			0.507				0.617
Random effects							
Face				0.026			0.012
Body				0.061			0.031
Residual				0.113			0.117

Note: Black is the reference category for race contrasts

[a] r^2_{sp} refers to semi-partial r^2 statistics, except the bottom-most values, which indicates r^2 for the full model.

Study 3b. We fitted the same series of cross-classified HLMs predicting Target D Scores for the Study 3b full-body targets. Again, we observed a significant effect of target race, with both Asian targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.10(0.03), t(18.44) = 3.80, p = 0.001, r^2_{sp} = 0.05, and White targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.09(0.03), t(18.41) = 3.46, p = 0.003, r^2_{sp} = 0.04, evaluated more positively than Black targets (for the simultaneous addition of both race dummies Δr^2 = 0.06), but no significant differences between Asian and White targets, t(19.17) = -0.35, p = 0.73. We also observed significant effects of target gender, with female targets evaluated more positively than males, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.23(0.03), t(19.79) = 7.06, p < .001, r^2_{sp} = 0.38, and of target social class, with upper-class targets evaluated more positively than lower-class targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.04(0.02), t(21.59) = 2.83, p = .01, r^2_{sp} = 0.05. Targets' age did not significantly affect implicit evaluations (See Table 6).

As in Study 3a, adding two-way interactions did not significantly improve model fit, χ^2 (9) = 11.99, p = 0.21 (see Supplementary Materials), and target gender was the only manipulated factor that remained a significant predictor over and above the control variables, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.24(0.03), t(23.46) = 7.31, p < .001, r^2_{sp} = 0.34. In this model we also observed a significant effect of photo blurriness, with more blurry photos eliciting more negative evaluations, $\hat{\beta}(SE_{\hat{\beta}})$ = -0.04(0.01), t(25.78) = -3.79, p < .001, r^2_{sp} = 0.13.

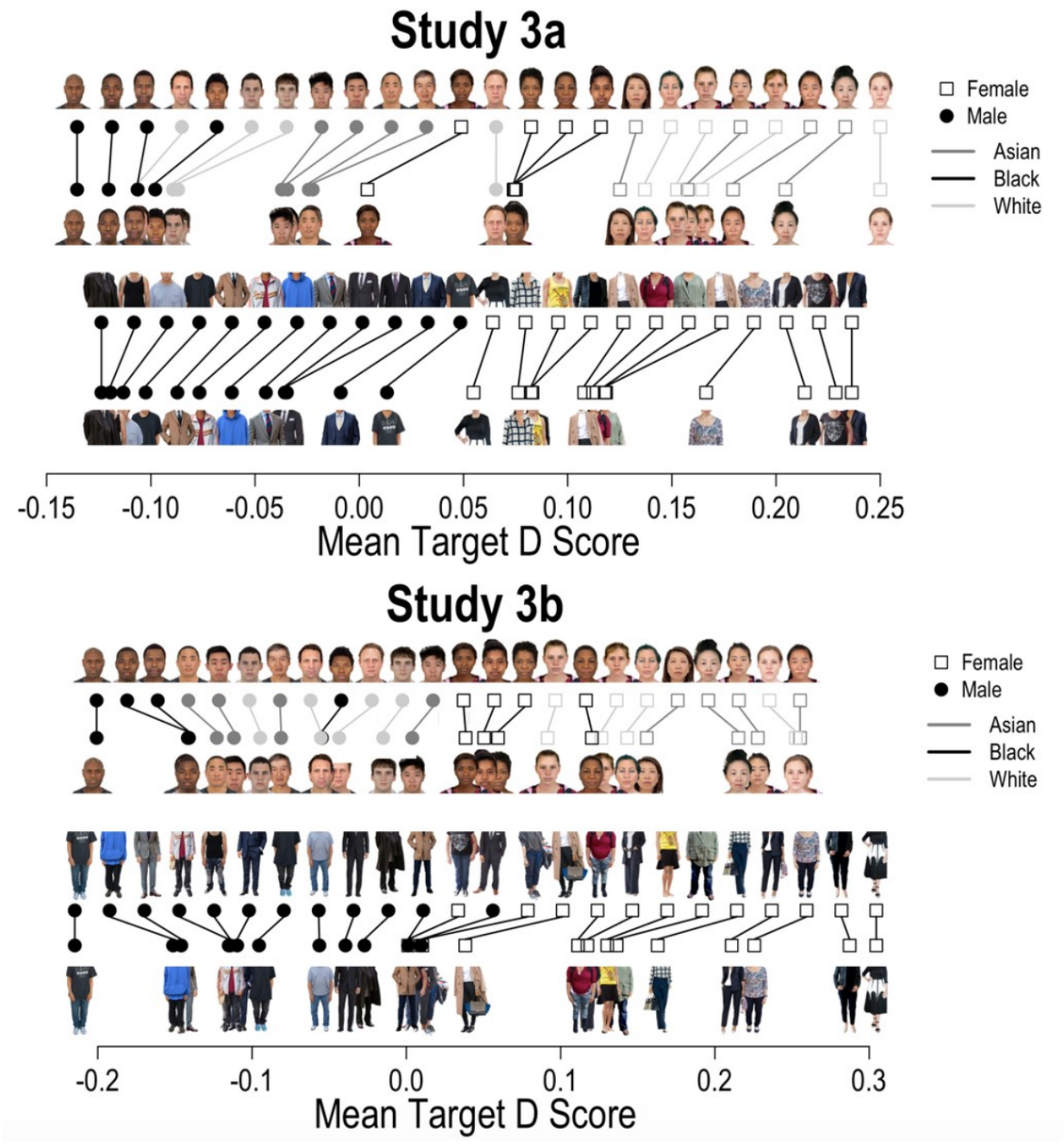


Figure 6. The effects of target race and gender in Study 3a and 3b visualized by showing each unique face and body arranged according to their mean Target D Scores (lower rows) and rank-ordered by their mean Target D Scores (upper rows).

Simulation-based power sensitivity analyses suggested that due to the package lmerTest's (Kuznetsova et al., 2017) use of the Satterthwaite degrees of freedom method, statistical power varied between effects. Both Study 3a and 3b achieved 80% power to detect main effects between approximately r^2_{sp} = 0.10 and r^2_{sp}= 0.15 and interaction effects between approximately r^2_{sp} = 0.05 and r^2_{sp}= 0.15 (for more details see Supplementary Materials).

Discussion

In Study 3, we again measured implicit evaluations of targets varying in race, gender, social class, and age. Across both methods we observed a dominant effect of target gender, which exerted effects many times larger than any other target-level factor. This result is most

consistent with the category dominance model, which posits that responses to multiply categorizable targets will be driven by single dominant categories. This theory is agnostic to which category will dominate when participants are not primed or manipulated in specific ways, and our finding that gender emerged as the dominant category in the present context is notable.

However, this dominance of gender was not absolute. We also observed effects of targets' race, with Asian and White targets evaluated more positively than Black targets in both studies, and social class, with upper-class targets evaluated more positively than lower-class targets in Study 3b. These results therefore also provide some level of support for the notion of compounding bias, as they suggest that implicit biases do combine additively, at least to some extent, across multiple social categories.

Three further results of Study 3 were noteworthy. First, the presence of anti-Black bias in both studies was consistent with the idea that such biases may have been suppressed in Studies 1 and 2, perhaps as a result of recoding strategies (Meissner & Rothermund, 2013). Second, the observation of a significant effect of social class only for the full-body targets in Study 3b aligned with the idea that full-body target images may increase the relative salience of social class. Third, Study 3, with its more tightly controlled design, did not replicate the interaction between target gender and social class observed in Study 2, calling into question the generalizability of that result.

Study 4

Study 3 revealed gender to be a predominant category driving implicit evaluations of multiply categorizable social targets varying in race, gender, social class, and age. Two issues, though, animated our last studies. First, Study 3 used non-representative samples of university students (71% and 75% female and 49% and 55% Asian in Studies 3a and 3b, respectively). Second, Study 3 relied solely on ST-IATs to measure implicit evaluations. As discussed above, Gawronski and colleagues (2010) have argued that different measurement procedures might produce different patterns of implicit biases toward multiply categorizable targets. In Study 4 we sought to address these issues by (a) recruiting a nationally representative sample of U.S. adults, and (b) measuring implicit evaluations via three different methods: ST-IATs, EPT (Fazio et al., 1986) and AMP (Payne et al., 2005).

Participants and Procedure

We recruited two separate samples of U.S. adults nationally representative on gender, age, and race via Prolific (Study 4a: N = 1620, 803 women, 790 men, 20 non-binary, 7 missing gender data, M_{age} = 38.6, SD_{age} = 14.2, 1167 White, 155 Black, 140 Asian, 103 Latino, 38 other race, 17 missing race data; Study 4b: N = 846, 423 women, 415 men, 4 non-binary, 4 missing gender data, M_{age} = 44.5, SD_{age} = 20.8, 620 White, 117 Black, 58 Asian, 34 Latino, 8 other race, 9 missing race data).

All participants were randomly assigned to evaluate one of the six target groups used in Study 3, which they viewed in in either full-body or upper-body presentation (Study 4a) or upper-body presentation only (Study 4b). In Study 4a participants completed two ST-IATs and one EPT, with the tasks randomly ordered. In Study 4b participants completed one AMP. Study 4a was pre-registered at https://aspredicted.org/r2ea2.pdf. Study 4b was pre-registered at https://aspredicted.org/8m3ux.pdf. As pre-registered, in Study 4a we excluded ST-IAT data from 9 participants and EPT data from 6 participants for having mean response times greater than 3000ms.[23] In Study 4b we excluded 38 participants for uniform responses on the AMP, and 9 participants for having mean response times greater than 3000ms.

ST-IATs

ST-IATs in Study 4a followed the same procedure as those administered in Study 3.

Evaluative Priming Task

EPTs in Study 4a began with 10 practice trials in which the symbols "***" were presented in the center of participants' screens for 200ms, followed by an interstimulus gap of 100ms, and then one of 24 positive words or 24 negative target words (e.g., "honor", "lucky", "evil", "cancer", Draine & Greenwald, 1998). Participants were tasked with categorizing the target words as either "Good" or "Bad" as quickly as possible via E or I computer key presses, with the assignment of valences to keys randomised between participants. Following this, participants performed 96 test trials (4 per target) in which the multiply categorizable target images were presented as primes in place of the "***" symbols. Each multiply categorizable target image was presented prior to two positive and two negative target words, and there was a

[23] We deviated slightly from our pre-registration due to our evolving understanding of the optimal algorithm for computing ST-IAT Target D Scores by using response time cut-offs of 100ms and 4000ms instead of 100ms and 6000ms, and by penalizing error trials. As reported in Supplementary Materials, these deviations had little effect on our results.

2500ms gap between the presentation of each prime/target pairing. Participants took breaks after the 32nd and 64th trials, and proceeded when ready.[24]

Affect Misattribution Procedure

In the AMP in Study 4b, the words 'Unpleasant'/'Pleasant' appeared at the top left/right of participants' screens. In each trial a multiply categorizable target was displayed as a prime for 75ms, followed by an inter-stimulus gap of 125ms, followed by one of 200 Chinese characters displayed for 100ms, followed by a pattern mask. Participants were tasked with categorizing the Chinese characters as either less pleasant than average or more pleasant than average via their E and I keys, respectively. Participants completed 10 practice trials, followed by 150 test trials, with a break after the 75th trial.

Explicit Ratings of Targets

Participants in Study 4a rated each of the 24 targets in their assigned target group via 0-100 sliders on perceived gender (ICC = 0.91), race (three separate sliders measuring perceptions of targets as Asian, ICC = 0.88, Black, ICC = 0.92, and White, ICC = 0.84) social class (ICC = 0.53), age (ICC = 0.59), attractiveness (ICC = 0.18), and photo blurriness (ICC = 0.48).

Demographics

Finally, participants reported the same demographic information as in Studies 2 and 3.

Results

Calculating Target D Scores

For the ST-IAT data, we calculated Target D Scores for each of the 288 unique target images (144 targets presented in both full- and upper-body formats) according to the algorithm described above (split-half reliability = 0.40). For the EPT and AMP data, we again undertook a data-driven process to determine which scoring algorithm would produce the highest combined internal reliability and convergent validity. This process suggested that both EPT and AMP data require different scoring algorithms to optimize measurement. This was especially the case for the EPT: applying the ST-IAT algorithm to the EPT data resulted in virtually zero internal reliability (see Supplementary Materials for details).

---

[24] . We chose 96 trials to obtain a roughly equivalent amounts of potentially useable trials per participant for the ST-IAT and EPT measures (in total, two ST-IATs provide approximately 80 potentially useable trials per participant).

For the EPT, the method providing the best measurement involved (a) identifying all raw response times toward a specific target in EPT trials, (b) eliminating response times below 175 milliseconds and above 1000 milliseconds (0.006% and 0.097% of trials, respectively), (c) taking the natural log of the remaining response times, (e) computing a difference score for each target representing the mean logged response time to the target in incompatible trials minus the mean logged response time to the target in compatible trials. For interpretability, we again divided these differences by the overall standard deviation of all logged EPT response times between 175 and 1000 milliseconds. This procedure yielded an estimated split-half reliability for the EPT Target D Scores of 0.28.

For the AMP, the method providing the best measurement involved (a) identifying all responses following each specific target prime, (b) eliminating responses faster than 75 milliseconds or slower than 4500 milliseconds (0.006% and 0.013% of trials, respectively), (c) computing the proportion of the Chinese characters judged more positive than average following each target prime (M = 0.62, SD = 0.03, range = 0.53-0.70). This procedure yielded an estimated split-half reliability for the AMP Target D Scores of 0.76.[25]

Predicting Target D Scores

For each Target D Score (ST-IAT, EPT, and AMP), we fitted a separate series of cross-classified HLMs. To test for differences between full-body and upper-body presentation in Study 4a, separate full-body and upper-body Target D Scores were computed for each target, and both were included in each model. For Study 4b, a single Target D Score was computed for each target.

An initial model predicted Target D Scores from fixed effects of z-scored mean ratings of targets' subjective SES, dummy variables indicating Asian race, White race, and female gender, and z-scored mean ratings of targets' age. As in Study 3, we included in each model random intercepts for targets' faces and bodies. A second model added a dummy variable indicating whether targets were observed in full-body or upper-body format (0 = upper-body, 1 = full-body), and a third model added two-way interactions between each target-level factor and the full-body indicator to test whether the effect of targets' social class, race, gender and age were

[25] This result is similar to that of Gawronski and colleagues (2010), who also found the AMP to provide a much more reliable measurement tool for measuring evaluations of multiply categorizable targets than the EPT.

moderated by presentation format. If these interaction terms failed to significantly improve fit compared to the second model, they were removed. A fourth model added two-way interactions between each target-level factor. Again, if these interaction terms failed to significantly improve fit compared to the previous model, they were removed. A fifth and final model added *z*-scored mean ratings of targets' attractiveness and photo blurriness.

ST-IAT Target D Scores. For ST-IAT Target D Scores, in the initial model we observed significant effects of target social class, with higher-class targets evaluated more positively than lower-class targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.03(0.01), t(23.12) = 5.1, $p < .001$, r^2_{sp}= 0.13. We also observed significant effects of target gender, with female targets evaluated more positively than male targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.14(0.01), t(20.15) = 11.49, $p < .001$, r^2_{sp}= 0.43, and target race, with both Asian targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.06(0.01), t(266.43) = 3.91, $p < .001$, r^2_{sp} = 0.07, and White targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.05(0.01), t(263.94) = 3.78, $p < .001$, r^2_{sp} = 0.07, evaluated more positively than Black targets (for the simultaneous addition of both race dummies Δr^2 = 0.04). There was no significant difference between evaluations of Asian and White targets, t(273.32) = -0.13, p = 0.89. Targets' age had no significant effect on implicit evaluations (see Table 7). In the second model, we observed a significant effect of the full-body target indicator, with full-body targets evaluated more negatively than upper-body targets, $\hat{\beta}(SE_{\hat{\beta}})$ = -0.05(0.01), t(261.03) = -4.52, $p < .001$, r^2_{sp} = 0.09. Model fit was not significantly improved by adding two-way interactions between the full-body target indicator and each of the target-level factors, $\chi^2(5) = 4.25$, $p = 0.51$, or by adding two-way interactions between each of the target-level factors, $\chi^2(9) = 4.98$, $p = 0.84$. Fixed effects estimates remained virtually unchanged after controlling for attractiveness and photo blurriness (results of Models 1 and 5 are reported in Table 7; for full results of all models see Supplementary Materials).

EPT Target D Scores. For EPT Target D Scores in Study 4a, in the initial model we observed significant effects of target social class, with higher-class targets evaluated more positively than lower-class targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.02(0.01), t(23.1) = 3.5, p = .002, r^2_{sp}= 0.07. We also observed significant effects of target gender, with female targets evaluated more positively than male targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.05(0.01), t(20.01) = 4.05, $p < .001$, r^2_{sp} = 0.09, and target race, with Asian targets evaluated more positively than both Black targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.03(0.02), t(266.46)

= 2.10, p = .04, r^2_{sp} = 0.02, and White targets,[26] $\hat{\beta}(SE_{\hat{\beta}})$ = -0.04(0.02), t(273.54) = -2.34, p = .02, r^2_{sp} = 0.03 (for the simultaneous addition of both race dummies Δr^2 = 0.03). There was no significant difference between evaluations of White and Black targets, t(263.87) = -0.24, p = 0.81. Targets' age also had no significant effect on implicit evaluations (see Table 7). In the second model, there was no significant effect of the full-body target indicator, t(260.88) =- 0.19, p = 0.85. Model fit was not significantly improved by adding two-way interactions between the full-body target indicator and each of the target-level factors, $\chi^2(5)$ = 5.52, p = 0.36, or by adding two-way interactions between each of the target-level factors, $\chi^2(9)$ = 5.31, p = 0.81. After controlling for attractiveness and photo blurriness, the gender and pro-Asian/anti-Black biases remained significant, but the effect of social class and the difference between Asian and White targets became non-significant (see Table 7).

AMP Target D Scores. For the AMP Target D Scores in Study 4b, in the initial model we observed significant effects of target social class, with higher-class targets evaluated more positively than lower-class targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.01(0.002), t(138) = 3.99, p < .001, r^2_{sp} = 0.12. We also observed significant effects of target gender, with female targets evaluated more positively than male targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.02(0.005), t(138) = 5.01, p < .001, r^2_{sp}= 0.18, and target race, with both Asian targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.02(0.01), t(138) = 3.68, p < .001, r^2_{sp}= 0.10, and Black targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.02(0.01), t(138) = 2.98, p = 0.003, r^2_{sp} = 0.07, evaluated more positively than White targets (for the simultaneous addition of both race dummies Δr^2 = 0.08). There was no significant difference between evaluations of Asian and Black targets, t(138) = 0.74, p = 0.46. Targets' age also had no significant effect on implicit evaluations (see Table 7). Model fit was not significantly improved by adding two-way interactions between each of the target-level factors, $\chi^2(9)$ = 5.75, p = 0.76. Only target gender remained significant after controlling for attractiveness and photo blurriness (see Table 7).

Simulation-based power sensitivity analyses suggested that Study 4a achieved 80% power to detect main effects of between approximately r^2_{sp} = 0.05 and r^2_{sp} = 0.075, and interaction effects between approximately r^2_{sp} = 0.025 and r^2_{sp} = 0.075, while Study 4b achieved 80% power

[26] The Asian-White result refers to a model fit with Asian set as the reference level for the race variable.

to detect main effects between approximately r^2_{sp} = 0.075 and r^2_{sp}= 0.10, and interaction effects between r^2_{sp} = 0.05 and r^2_{sp}= 0.10. (see Supplementary Materials for details).

Table 7
Results from hierarchical linear models in Study 4

	ST-IAT Target D Scores (Study 4a)							
	Model 1				Model 5			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*
Fixed effects								
(Intercept)	-0.119(0.012)	<.001			-0.09(0.013)	<.001		
Social class	0.032(0.006)	<.001	0.128		0.032(0.011)	0.006	0.04	
Asian	0.056(0.014)	<.001	0.07		0.056(0.015)	<.001	0.068	
White	0.054(0.014)	<.001	0.066		0.054(0.016)	0.001	0.052	
Female	0.144(0.013)	<.001	0.428		0.144(0.015)	<.001	0.328	
Age	-0.01(0.006)	0.137	0.013		-0.009(0.007)	0.197	0.009	
Full-body target					-0.057(0.013)	<.001	0.088	
Attractiveness					-0.002(0.013)	0.887	<.001	
Blurry					-0.008(0.007)	0.254	0.007	
			0.493				0.534	
Random effects								
Face				<.001				<.001
Body				0.011				0.01
Residual				0.1				0.096
	EPT Target D Scores (Study 4a)							
	Model 1				Model 5			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*
Fixed effects								
(Intercept)	0.108(0.013)	<.001			0.109(0.014)	<.001		
Social class	0.024(0.007)	0.002	0.067		0.004(0.012)	0.756	<.001	
Asian	0.032(0.015)	0.037	0.021		0.041(0.016)	0.011	0.032	
White	-0.004(0.015)	0.813	<.001		0.014(0.018)	0.441	0.003	
Female	0.054(0.013)	<.001	0.088		0.035(0.016)	0.039	0.024	
Age	-0.003(0.007)	0.639	0.001		0.004(0.008)	0.566	0.002	
Full-body target					0.0002(0.014)	0.989	<.001	
Attractiveness					0.025(0.014)	0.07	0.017	
Blurry					-0.006(0.007)	0.375	0.004	
			0.168				0.186	
Random effects								
Face				<.001				<.001
Body				0.013				0.009
Residual				0.105				0.104
	AMP Target D Scores (Study 4b)							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	r^2_{sp} [a]	*SD*
Fixed effects								
(Intercept)	0.616(0.005)	<.001			0.616(0.005)	<.001		
Social class	0.01(0.002)	<.001	0.119		0.005(0.004)	0.173	0.016	
Asian	0.004(0.006)	0.462	0.005		0.007(0.006)	0.256	0.011	
White	-0.018(0.006)	0.003	0.07		-0.012(0.007)	0.069	0.028	
Female	0.024(0.005)	<.001	0.175		0.019(0.006)	0.002	0.075	
Age	-0.002(0.003)	0.361	0.007		0(0.003)	0.983	<.001	
Attractiveness					0.008(0.005)	0.075	0.027	
Blurry					0.003(0.003)	0.202	0.014	
			0.328				0.346	
Random effects								
Face				<.001				<.001
Body				<.001				<.001
Residual				0.029				0.029

Note: Black is the reference category for race contrasts
[a] r^2_{sp} refers to semi-partial r^2 statistics, except the bottom-most values, which indicates r^2 for the full model.

Discussion

In Study 4 we measured implicit evaluations of targets varying in race, gender, social class, and age using ST-IATs, EPTs, and AMPs. Target gender again emerged as the most

important predictor of implicit evaluations, with female targets evaluated more positively than males, and target gender explaining more variation in ST-IAT, EPT, and AMP Target D Scores than any other factor. We also observed smaller but consistent effects of target social class, with upper-class targets evaluated more positively than lower-class targets via all three methods. By contrast, the effects of race were inconsistent, with participants favoring White and Asian over Black targets in ST-IATs, Asian over White and Black targets in EPTs, and Asian and Black over White targets in AMPs. We observed no significant effects of target age, no significant interactions between target-level factors, and no significant moderation of effects by presenting targets in upper-body compared with full-body target presentation.

These results suggest that the dominant effect of gender in Study 3 was not due to non-representative sampling. In a sample of U.S. adults nationally representative with regard to race, gender, and age, target gender exhibited a similar-sized effect on ST-IAT Target D Scores (r^2_{sp} = 0.43) as it had in Study 3 (r^2_{sp} = 0.49). However, these results also suggest that the dominance of target gender in Study 3 may have been amplified by its exclusive reliance on ST-IATs. Although target gender was the largest effect across all three methods used, its relative dominance was less pronounced in EPTs and AMPs.

Study 5

We conducted Study 5 to address two final questions. First, we were curious how patterns of responses to multiply categorizable targets varied for different sub-groups of our respondents. For example, although we found pro-female/anti-male evaluative biases to be the most important driver of our results, past work has found that such biases tend to larger in women than in men (Richeson & Ambady, 2001). In Study 5a we conducted an integrative data analysis (Curran & Hussong, 2009) to investigate a number of such potential moderators of our results.

Second, observed effects of gender, social class, and race are compatible with multiple explanations. It was possible that participants had simultaneously attended to and displayed bias with respect to all three categories: gender, class, and race, but it was also possible that specific groups of participants had attended and shown bias with respect solely to target gender, social class, or race, respectively. In Study 5b we tested between these competing accounts.

Study 5a: Exploring Moderators

In Study 5a we tested the extent to which measurement tasks (ST-IAT, EPT, AMP), sample sources (students, Prolific), participants' gender, participants' race, participants' age,

participants' SES, and participants' political affiliations moderated the effects of target gender, race, and social class. To do so, we combined all the raw implicit evaluation data from Studies 2, 3, and 4.[27] We then computed implicit evaluation scores for each unique participant/target/task combination in the data. For example, if participant X was exposed to target Y in a ST-IAT, the corresponding participant/target/task evaluation score was computed by isolating all of participant X's responses to target Y within ST-IATs, and then applying the ST-IAT Target D Score algorithm to this data to compute an evaluation score specific to the participant/target pairing. Because ST-IAT and EPT Target D Scores require the calculation of difference scores, evaluation scores for these tasks were calculated only for participant/target pairs with at least one usable response in both compatible and incompatible trials.

To allow comparability across tasks, we z-scored the resulting evaluation scores within tasks (ST-IAT, EPT, AMP). We also converted targets' perceived social class into a binary predictor via median split, and converted participants' age, subjective SES, and political affiliation into three-category predictors. We excluded any participants missing data on moderators, as well as insufficiently represented racial or gender categories (see Table 8). This left a final sample of 103,715 unique participant/target/task evaluation scores, representing 3,659 participants' implicit evaluations of 198 targets (because we found no significant differences between full- and upper-body presentations in Study 4, we treated responses to targets across both presentation modes as responses to the same target).

Table 8
Descriptive statistics of moderators included in integrative data analysis

Moderator	Categories included	Categories excluded
Task	ST-IAT: 2,221	
	EPT: 679	
	AMP: 759	
Sample	Students: 1,418	
	Prolific: 2,241	
Gender	Women = 2,186	Non-binary = 43
	Men = 1,473	
Race	Asian = 1,003	Other race = 131
	Black = 295	
	Latino = 283	
	White = 2,078	
Age	>50 years = 641	
	31-50 years = 1,035	
	18-30 years = 1,983	
Subjective SES[a]	High (8-10) = 522	
	Medium (5-7) = 2,220	
	Low (1-4) = 917	

[27] Study 1 data was not included in the integrative data analysis because Study 1 participants were not measured on subjective SES, or exposed to female or Asian targets.

Political affiliation	Liberal (8-10) = 1,918 Moderate (5-7) = 1,281 Conservative (1-4) = 460

[a] Subjective SES was measured via the MaCarthur ladder scale
[b] Political affiliation was measured via a 1-10 Likert scale with 1 = Extremely conservative and 10 = Extremely liberal

Results

We first measured the main effects of target gender, target race, and target social class by fitting a cross-classified hierarchical linear model predicting evaluation scores from fixed effects of each target-level factor, plus random intercepts for participants and targets. These results are denoted in Figure 7 as 'Overall effects,' with associated $\chi^2$and Δr^2 values representing the model fit improvement from each factor being added to this initial model.

Following this, we tested how the effects of target gender, race, and social class differed depending on each moderator. To do so, we fitted a full model including fixed effects of each moderator and each possible two-way interaction between moderators and target-level factors. This meant that each interaction was tested while controlling for all other interactions. This was desirable given high levels of covariation among the moderators (e.g., student samples were largely Asian and largely women, EPT and AMP samples were on average older than ST-IAT samples). Each two-way interaction is visualized in Figure 7, with the associated $\chi^2$and Δr^2 values representing model comparisons between this full model and models with all interactions except the focal interaction. Given the number of tests run, we adjusted *p* values using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995; for full model results see Supplementary Materials).

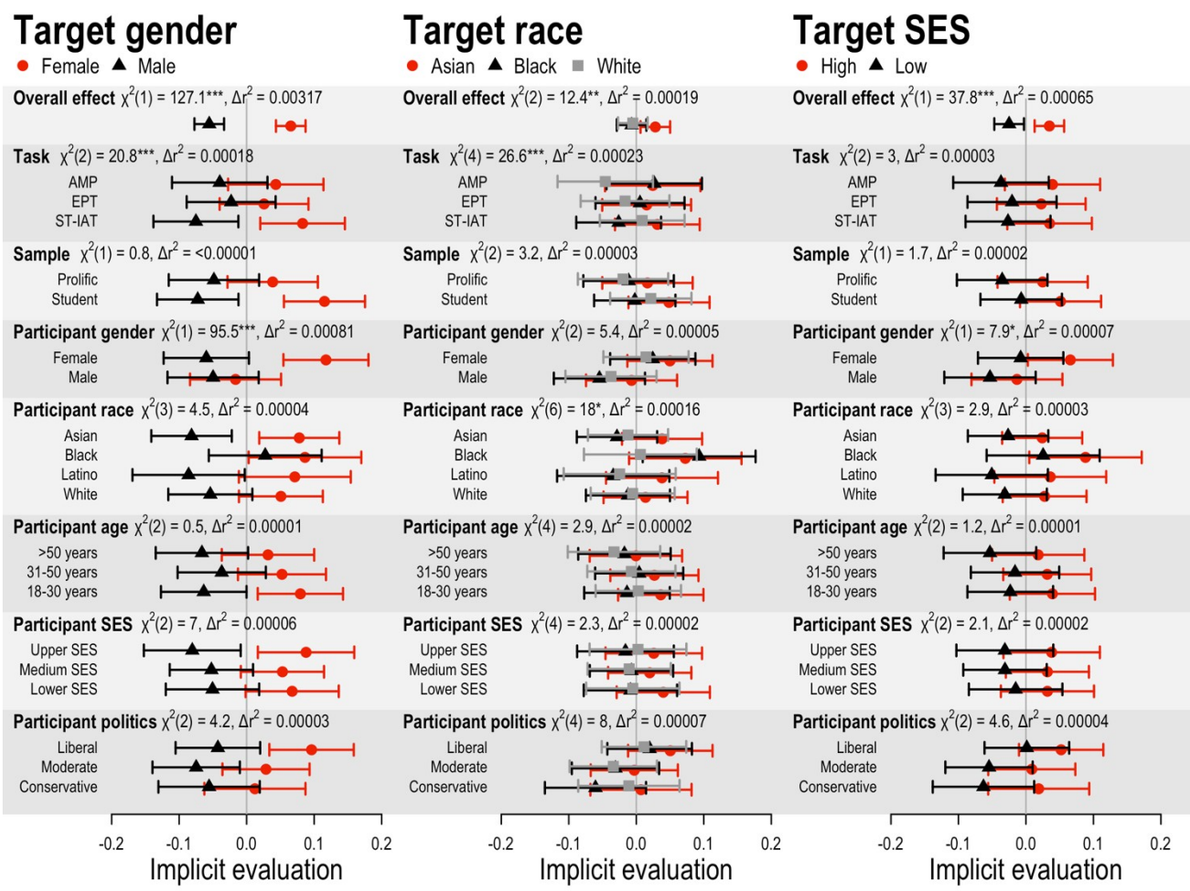


Figure 7. Effects of target gender, race, and social class, and how these were moderated by task, sample, and participant gender, race, age, SES, and political affiliation. 'Overall effect' $\chi^2$and Δr^2 values represent model fit improvements from adding each target-level predictor to a model containing both other target-level predictors. All other $\chi^2$and Δr^2 values represent model fit improvements from adding each two-way interaction to a model containing all other two-way interactions between moderators and target-level predictors; *** = p < .001, ** = p < .01, * = p < .05.

Results showed that the effect of target gender was significantly moderated by task, $\chi^2(2) = 20.8$, $p < .001$, $\Delta r^2 = 0.0002$, and participant gender, $\chi^2(1) = 95.5$, $p < .001$, $\Delta r^2 = 0.0008$. Bias favoring female targets was displayed in all tasks and among all sub-groups of participants, but the effect was stronger in ST-IATs (gender difference = 0.16, *SE* = 0.01) than in EPTs (gender difference = 0.05, *SE* = 0.01) and in AMPs (gender difference = 0.08, *SE* = 0.01), and was stronger among women (gender difference = 0.18, *SE* = 0.01) than men (gender difference = 0.03, *SE* = 0.01).

The effect of target race was moderated by task, $\chi^2(4) = 26.6$, $p < .001$, $\Delta r^2 = 0.0002$, and participant race, $\chi^2(6) = 18.0$, $p = .01$, $\Delta r^2 = 0.0002$. Participants displayed pro-Asian/anti-Black bias in ST-IATs (Asian – Black difference = 0.05, *SE* = 0.02; all other differences *NS*), little

racial bias in EPTs (all differences *NS*), and anti-White bias in the AMP (Asian – White difference = -0.07, *SE* = 0.02; Black – White difference = 0.06, *SE* = 0.02, Asian – Black difference *NS*). Asian participants displayed a pro-Asian bias (Asian – Black difference = 0.07, *SE* = 0.02; Asian – White difference = 0.05, *SE* = 0.02, White – Black difference *NS*), Black participants displayed an anti-White bias (Black – White difference = 0.08, *SE* = 0.03; Asian – White difference = 0.07, *SE* = 0.03, Black – Asian difference *NS*), Latino participants displayed a pro-Asian/anti-Black bias (Asian – Black difference = 0.07, *SE* = 0.03, all other differences *NS*), and White participants displayed little racial bias (all differences *NS*).

The effect of target social class was significantly moderated by participant gender, $\chi^2(1) = 7.9$, $p = .005$, $\Delta r^2 = 0.00007$, with women showing a greater bias (upper – lower difference = 0.07, *SE* = 0.02) than men (upper – lower difference = 0.04, *SE* = 0.01). No other interactions reached significance.

Discussion

In Study 5a we explored how task type, sample source, and participants' gender, race, age, social class, and political affiliation moderated the effects of targets' gender, race, and social class. Although some notable interactions emerged, there was striking consistency across results. For example, implicit gender bias was greater among women than men, and greater in ST-IAT tasks than EPTs and AMPs. However, every sub-group of respondents displayed a pro-female/anti-male bias. Similarly, implicit social class bias was stronger among women than men, yet every sub-group of respondents displayed a pro-upper-class/anti-lower-class bias. Taken together, these results suggest that while the relative magnitude of implicit gender and social class biases may vary across demographic groups, the fundamental directions of these biases are relatively stable.

By contrast, the effect of race was less consistent, with participants displaying pro-Asian/anti-Black bias in the ST-IAT, little detectable racial bias in the EPT,[28] and anti-White bias in the AMP. Additionally, Asian participants displayed a clear ingroup bias favoring Asian over Black and White targets, Black participants favored Asian and Black targets over Whites, Latino

[28] Via the Target D Score analysis in Study 4a, a pro-Asian/anti-Black/anti-White bias was detected via the EPT data. This difference likely reflects the data exclusions and different scoring method used in the integrative data analysis.

participants favored Asian over Black targets, and White participants displayed no significant racial bias overall.

With the exception of the differences in implicit gender bias between women and men (Richeson & Ambady, 2001), the majority of these interactions involve novel observations. We are not aware of any prior work that would have predicted the effect of target gender to be strongest in ST-IATs, the effect of target social class to be stronger among women than men, White participants to show the least racial bias of any racial group, or a robust anti-White bias to emerge on AMPs. Each of these findings may warrant further attention and research, yet given their exploratory and unanticipated nature, each should also be regarded as preliminary and suggestive only.

Study 5b: Testing for Category-Dominant Sub-Groups

As discussed above, participants may have simultaneously attended to the separate categories of target gender, race, and social class, or alternatively, separate sub-groups of participants may have attended to each category. The noisiness of implicit bias data makes it difficult to tease these alternate explanations apart, but one way to do so is to assess the relationship between separate biases at the level of individual participants. In our case, we focused on the relationship between participants' implicit gender bias and participants' implicit social class bias, as these were the two most consistent biases displayed in our data, and could both be easily quantified. [29]

Here, the reasoning is that if our observation of both gender and social class biases came about via distinct groups of participants attending to either targets' gender or to targets' social class, this would be expected to produce a negative correlation between the two biases among participants. This is due to the expected distributions of each kind of bias among each sub-group of participants. The gender-focused group would produce a distribution of gender bias scores centered above zero, and a distribution of social class bias scores centered near zero.[30] Meanwhile, the class-focused group would produce a distribution of social class bias scores

[29] For this analysis we ignored racial bias due to the inconsistency of racial biases in our data, and the complexity involved in creating racial bias scores from evaluations of three categories of targets.

[30] We say centered near zero here because the zero point (equivalent response times in compatible vs. incompatible trials) does not necessarily indicate a lack of bias.

centered above zero, and a distribution of gender bias scores centered near zero. This would mean that even with substantial amounts of added measurement error, individuals exhibiting relatively higher gender bias scores would be more likely to belong to the gender-focused group, and so would be more likely to exhibit relatively lower social class bias scores. Conversely, individuals exhibiting relatively higher social class bias scores would be more likely to belong to the class-focused group, and so would be more likely to exhibit relatively lower gender bias scores. This should produce a negative correlation between the two kinds of bias, which we demonstrate below via simulation.

Results

To quantify participants' gender and social class bias, we used the evaluation scores from in Study 5a, and for each of 3,657 participants[31] computed gender bias scores (participants' mean evaluation scores for female targets minus their mean evaluation scores for male targets; M = 0.13, SD = 0.46), and social class bias scores (participants' mean evaluation scores for upper class targets minus their mean evaluation scores for lower class targets; M = 0.06, SD = 0.46). These scores displayed a significant positive correlation, r = 0.07, t(3655) = 3.99, $p < .001$ (see Figure 8). To assess how unlikely this correlation would be if the data were produced by distinct groups focused on separate categories, we simulated samples of 3,656 gender and social class bias scores with means and standard deviations matching our observed data, but manipulated the data such that half the sample was 'gender-focused' (producing a distribution of gender bias centered above zero and a distribution of social class bias centered at zero), while the other half of the sample was 'class-focused' (producing in a distribution of social class bias centered above zero and a distribution of gender bias centered at zero). For each simulated distribution, we computed the correlation between the two biases. From 10,000 simulated datasets, 98% of these correlations fell below zero, and no correlations were higher than r = 0.03 (see the right panel of

[31] Two participants were missing data on responses to males or females.

Figure 8).

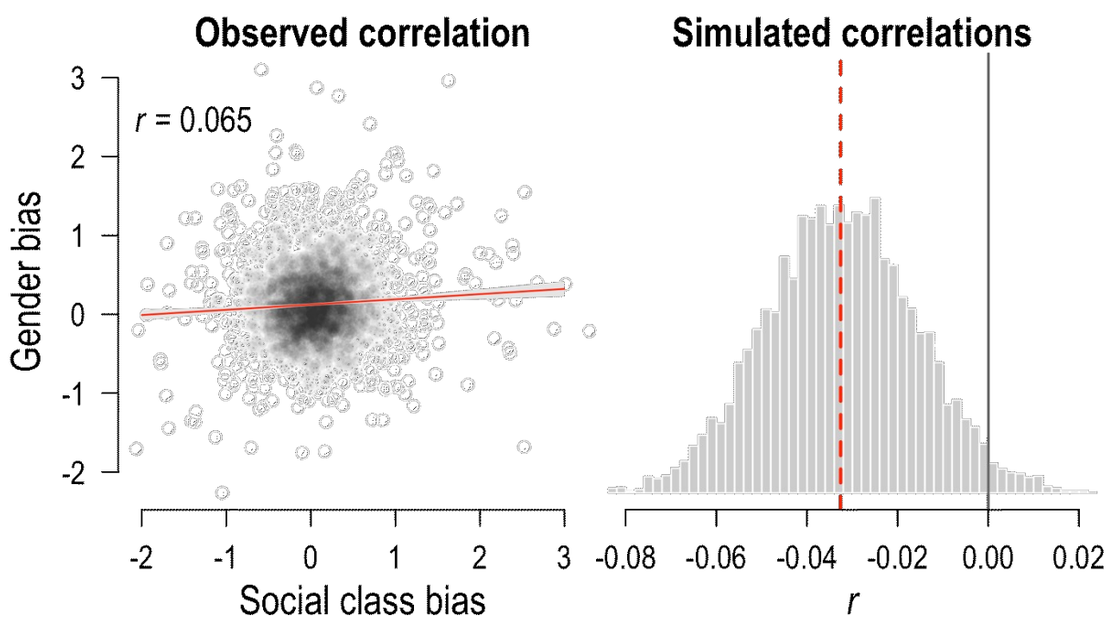


Figure 8. The observed correlation between participants' implicit gender bias and implicit social class bias scores (left), and the distribution of correlations between simulated bias scores matching our data in *N*, means, and standard deviations, but derived from distinct gender-focused and class-focused sub-groups (right).

Discussion

In Study 5b we observed a small but significant positive correlation between participants' gender and social class bias scores. Via simulation, we then demonstrated that such a correlation would be extremely unlikely if observed gender and social class biases result from distinct groups of participants attending solely to gender or to social class. This suggests it is highly unlikely our results were driven by distinct single-category-focused groups of participants.

General Discussion

Implicit bias is central to the study of social cognition. Given that people are multiply categorizable, understanding the influences of such intersectionality upon implicit bias is likely to be vital for understanding its effects in everyday social contexts. In the present research, we examined implicit evaluations of multiply categorizable social targets, testing two competing theories about intersectional intergroup bias. We also developed and tested the reliability of a novel method of measuring and modelling implicit bias at the level of individual targets.

In Study 1 we observed implicit evaluations of Black and White males to be driven solely by targets' social class, with bias favoring upper-class over lower-class targets. In Study 2, we measured implicit evaluations of targets varying in race, gender, social class, and age, and found results to be primarily driven by a specific positive bias favoring upper-class female targets. In Study 3, we used similarly intersectional targets, and explored the impact of portraying targets in full-body versus upper body photographs on implicit evaluations. Here, we observed effects of targets' race, with Asian and White targets evaluated more positively than Black targets, and of targets' social class, with upper-class targets evaluated more positively than lower-class targets (though only when targets were displayed in full-body presentation). Most striking, however, was the dominant effect of target gender, with positive/negative evaluations of female/male targets accounting for the majority of variance in implicit bias.

In Study 4 we tested the generalizability of these results by recruiting representative samples of US adults, and measuring implicit evaluations not just via ST-IATs, but also via EPTs and AMPs. Across all measures, we observed target gender to be the largest driver of implicit evaluations, though its dominance was less pronounced in EPTs and AMPs than in ST-IATs. We also again observed effects of targets' social class and race, though the effect of race was inconsistent across tasks, with participants displaying anti-Black bias in the ST-IAT, pro-Asian bias in the EPT, and anti-White bias in the AMP. Finally, in Study 5 we conducted an integrative data analysis to test a number of potential moderating factors. Results showed that while all groups of participants displayed pro-female implicit gender bias and pro-upper-class implicit social class bias, both biases were stronger among women than men. Results also showed the effect of race varied across racial groups, with Asians displaying a preference for Asian over White and Black targets, Black participants displaying a preference for Asian and Black targets over White targets, Latinos displaying a preference for Asian over Black targets, and Whites displaying no significant racial bias.

The present work makes theoretical, empirical, and methodological contributions to the study of implicit evaluative bias toward multiply categorizable targets. On a theoretical level, we believe our results are best accounted for by a synthesis of compounding bias and category dominance approaches. Consistent with the category dominance model (Macrae et al., 1995), we observed a single social category to exert a dominant influence on implicit evaluations of intersectional targets in each of our studies. In Study 1, social class was dominant. In Studies 3 and 4, target gender was dominant. And even in Study 2, despite its more complex results, target gender still uniquely accounted for substantially more variation in Target D Scores than any other target-level predictor. These results are consistent with the notion that when faced with complex social stimuli, social perceivers act as 'cognitive misers', and make implicit evaluations that are strongly influenced by specific social categories, and are relatively unaffected by others.

However, our results are also consistent with the notion that implicit biases compound—at least to some extent—across multiple categories at once. In Studies 3 and 4, which used the most tightly controlled set of targets, we observed simultaneous effects of targets' gender, race, and social class. And in Study 5b, we found little evidence that these results represented separate groups of participants attending solely to each factor. So, although we found little consistent evidence for the kind of multiplicative interaction effects suggested by the multiple jeopardy/advantage hypothesis (Ransford, 1980), we did find evidence of biases compounding additively—albeit highly asymmetrically—across multiple social categories, with the most negative implicit evaluations consistently being made of targets displaying multiple intersecting stigmatized social identities (in this case, lower SES males), and the most positive implicit evaluations being made of individuals displaying multiple intersecting positively-valued social identities (in this case, upper SES females).

The overall picture emerging from the present work is therefore one of theoretical compromise. Implicit evaluative biases toward complex multiply categorizable targets do appear to compound across categories, but do so asymmetrically, with a dominant category (here, target gender) playing a leading role, less dominant categories (here, target race and social class) exerting relatively small additional effects, and peripheral categories (here, target age) having little detectable influence.

This compromise position offers a novel rationale for grappling with intersectionality in psychological science. Often, arguments in favour of intersectional approaches stress the

importance of examining the experiences of individuals possessing multiple marginalized social identities, or the idea that specific category intersections give rise to emergent phenomena that cannot be understood by studying each category in isolation (e.g., Cole, 2009; Ghavami & Peplau, 2012; Goff, & Kahn, 2013; Kang & Bodenhausen, 2015). The present work does not invalidate these perspectives, but complements them, by suggesting that there may also be specific phenomena—such as implicit evaluations—in which responses to intersectional targets are best described by an asymmetrical compounding account.

Importantly, just like emergent intersectional effects, these asymmetries may also only be discoverable via intersectional research programs. For example, in past research on implicit evaluative bias, targets' race, gender, social class, and age have tended to produce biases of roughly comparable size (Nosek, 2005). However, this work has rarely used an intersectional lens, and the present results suggest that such methods may provide little guidance regarding the relative influence of each category when participants respond to complex, multiply categorizable targets. Indeed, even our traditional two-category Race IAT used in Study 2 provided a poor guide to responses to more complex targets, with participants displaying robust anti-Black bias on the two-category Race IAT, but no evidence of anti-Black bias when responding to multiply categorizable targets in ST-IATs. Given that intersectionality is a fact of everyday social encounters, this suggests that advancing understanding of how implicit bias operates in real-world contexts is likely to be severely limited by the absence of studying responses to realistically complex, intersectional targets.

On an empirical level, it is noteworthy that gender emerged as the dominant driver of implicit evaluations of multiply categorizable targets. This is consistent with one previous study, in which target gender was the sole significant predictor of categorization errors in a weapon identification task (Jones & Fazio, 2010). However, this prior work involved both a relatively small and non-representative sample (79 college students), and as a relatively small and idiosyncratic set of stimuli (8 total stimuli varying in race, gender, and occupation, with occupations not matched across races or genders, and no reported pre-testing of stimuli). The present results therefore provide a substantially more robust demonstration of this phenomenon.

It has long been established that individuals display pro-female evaluative biases via binary implicit measures (Nosek, 2005). However, compared with evaluative biases regarding race, or implicit associations between genders and specific social roles or abilities (e.g., Carlana,

2019; Levinson & Young, 2010), this phenomenon has attracted relatively little attention. However, its dominance in the present results suggests the greater attention to gender-based biases might have an important role to play in building our understanding of the causes and consequences of implicit evaluative bias.

One possible explanation for this result is that the dominance of gender was mediated by its overall visual salience. While race was conveyed within our stimuli by targets' faces and exposed skin, and social class was conveyed by targets' clothing, gender was conveyed by both faces and clothing. This may have made gender the most visually salient category, producing its dominant effect. Notably, however, even if this were the underlying mechanism, this would not preclude our results from generalizing to real-world interactions, as in most everyday contexts individuals' faces and bodies/clothing both tend to be visible, and to communicate gender.

Finally, from a methodological perspective, we believe that Target D Scores provide a promising path forward for studying intersectional implicit biases. Previously, researchers in this area have used one of two approaches. One approach has been to measure and model implicit attitudes at the level of target groups, either by calculating stand-alone measures of evaluations of target groups representing intersectional category combinations (e.g., Jones & Fazio, 2010; Mitchell et al., 2003, Studies 4 & 5; Moore-Berg et al., 2017; Perszyk et al., 2019), or by calculating multiple binary preferences from responses to targets varying on multiple categories (e.g., Gawronski et al., 2010; Mitchell et al., 2003, Studies 1-3; Yamaguchi & Beattie, 2019). However, this approach obscures systematic variation in implicit evaluations within target groups. By allowing investigators access to such within-target-group variation, Target D Scores allows for the investigation of the simultaneous influence of a greater number of target-level factors than is possible via target-group-based approaches. Additionally, target group-level approaches such as these require target groups to be orthogonal with respect to both manipulated variables and potential confounds, which is often not possible. As discussed above, Target D Scores allow for greater control over non-orthogonalities and confounds by allowing researchers to estimate effects of target-level predictors while controlling for targets' precise levels of other variables of interest, in a manner akin to conjoint studies (Hainmueller et al., 2014).

A second prior approach has been to measure and model responses to multiply categorizable targets at the level of individual (usually logged) response times (e.g., Mattan et al., 2019; Thiem et al., 2019). Like Target D Scores, this method allows researchers to study

systematic variation in implicit evaluations within target groups, and to control for target-level confounds. However, in contrast to such approaches, Target D Scores provide an intuitive, simple measure of samples' overall implicit evaluations of individual targets, and allow for the fitting of more straightforwardly interpretable models compared to raw response time models, which typically require interaction terms between target-level characteristics and indicators of compatible/incompatible trials. Moreover, unlike response time-level analyses, Target D Scores allow researchers to assess the reliability of measured evaluations of targets. This allows distinguishing between ranges of response times that contribute reliable information regarding implicit evaluations, and ranges of response times that contribute only unhelpful random noise.[32]

Some limitations regarding the present research should be noted. The first regards our restricted ability to detect higher order three-way or four-way interactions between target-level factors. We were reasonably well-powered to detect two-way interactions, which the multiple jeopardy/advantage hypothesis predicts to be present even if there are higher-order interactions.[33] However, there are other possible three- or four-way interaction patterns which do not entail the presence of two-way interactions, and we did not test for these given our limited number of stimuli. It is also plausible that there exist interactions which imply two-way interactions but whose effect sizes fall below levels we were sufficiently powered to detect. Consequently, while the present results do speak against the idea that certain patterns of interactions—including multiple jeopardy/advantage effects—are among the most important drivers of implicit evaluations of multiply categorizable targets, they do not speak to the existence of such effects at small effect sizes, or other more complex interactions.

---

[32] This was well illustrated in Study 4, where we observed Target D Scores to capture virtually zero reliable variation when we applied our ST-IAT algorithm directly to the EPT data. If we had relied on response time-level modelling in the present project, we would not have known that the EPT data required a different scoring algorithm altogether to obtain some level of internally reliable measurement.

[33] For example, if there were a three-way multiple jeopardy effect resulting in especially negative evaluations of lower SES Black male targets, tests of two-way interactions should in theory detect especially negative evaluations of Black male targets, lower SES Black targets, lower SES male targets, or any combination of these three sub-groups.

A second limitation is ambiguity regarding how to interpret discrepancies in results between measurement tasks. As discussed above, we observed a substantially more dominant effect of target gender in ST-IATs than the EPT and AMP. Previous researchers have argued that tasks reliant on the mechanism of response interference—such as the ST-IAT—are especially likely to produce category dominance (Gawronski et al., 2010). However, these researchers theorized that EPTs—which also rely on response interference—would produce greater category dominance effects than AMPs. By contrast, we observed a more dominant effect of gender in the AMP than the EPT, suggesting it is unlikely our category dominance results were a function of response interference tasks alone. One potentially important difference separating the ST-IAT method from the EPT and AMP methods is its reliance on key presses made directly in response to the multiply categorizable targets themselves, rather than subsequently displayed words (the EPT) or Chinese characters (the AMP). Plausibly, there may be a temporal factor involved in the evaluation of multiply categorizable targets, with category dominance strongest immediately after stimulus presentation, and thereafter reduced, or a focal effect, whereby tasks requiring responses directly to targets focus attention on targets' dominant categories in a way that other tasks do not. We leave this question for future research.

It is also unclear why the effect of target race varied across measurement tasks. Here, the most anomalous result was the anti-White bias displayed in the AMP, which runs counter to the anti-Black evaluative bias typically displayed by U.S. adults (e.g., Nosek, 2005), and previously demonstrated via AMPs using multiply categorizable targets (Gawronski et al., 2010). This result also ran counter to the anti-Black bias displayed by our samples in Studies 3 and 4a via ST-IATs. However, it is not unprecedented to obtain results counter to expectations when using the AMP to detect implicit prejudice (Teige-Mocigemba, Becker, Sherman, Reichardt, & Klauer, 2017). Given the number of studies run in the present manuscript, as well as the number of effects measured in each study, an anomalous result of this nature is perhaps not surprising.

Nonetheless, it is worth noting that target race in general tended to produce relatively inconsistent effects compared to target gender and social class, regardless of the measurement method. In Studies 1a, 1b, and 2, we observed no robust effect of race, and only in Studies 3a, 3b, and 4a did we observe robust anti-Black race effects in-keeping with prior literature. As discussed above, one explanation for these results may be that because targets of different race were presented in separate ST-IATs in Studies 1 and 2, participants used recoding strategies

(Meissner & Rothermund, 2013) to suppress anti-Black bias in these studies. Another is that due to perceived causal effects of race on social class (Pew Research Centre, 2019), and the process of *augmentation* (Kelley, 1973), matching target groups on explicit ratings of social class in Studies 1 and 2 may have led to the Black targets being perceived as higher on other traits conferring social class status, such as competence or industriousness. But neither of these explanations accounts for the anti-White bias observed in the AMP task in Study 4b. Given the consistency with which anti-Black bias is typically displayed in two-category IATs, the inconsistency of race effects in the present work is itself noteworthy, as it provides further evidence that we are yet to fully understand implicit bias in the context of complex, multiply categorizable targets.

Other major challenges for future research include incorporating even greater naturalistic complexity within target stimuli. In the present research, we focused on target-level variation in race, gender, social class, and age—four target dimensions that are perceptible in many if not most social interactions. Of course, real-world social targets vary on far more than just these four variables; modelling such complexity will require the study of other social variables, including variation in body shape (Bessenoff & Sherman, 2000; Teachman, Gapinski, Brownell, Rawlins, & Jeyaram, 2003), sexual orientation (Banse, Seise, & Zerbes, 2001; Steffens & Buchner, 2003), social and physical contexts (Barden et al., 2004; Wittenbrink et al., 2001), facial expressions (Steele et al., 2018), and more.

Finally, the present work focused only on identifying basic implicit evaluative biases defined by the facilitation/impedance of response times in timed categorization tasks. It will therefore be vital to assess how well implicit evaluations of multiply categorizable targets align with explicit bias measures, and how well each kind of measure predicts discriminatory behaviors. One key criticism of traditional implicit bias tests has been their relatively low correlations with discriminatory behavior (e.g., Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; but see Jost et al., 2009; Greenwald, Banaji, & Nosek, 2015). It may be the case that participants' spontaneously displayed implicit biases toward multiply categorizable targets will better predict behavior in real social contexts than traditional binary measures. This possibility is worthy of further investigation.

Ultimately, understanding how individual social perceivers, themselves members of multiple intersecting social categories, automatically respond to other complex, multiply

categorizable human beings is a daunting challenge. Nonetheless, we believe these challenges of intersectionality are vital to the future study of implicit bias.

References

Almquist, E. M. (1975). Untangling the effects of race and sex: The disadvantaged status of Black women. *Social Science Quarterly*, 129-142.

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health psychology*, *19*(6), 586.

Banaji, M.R., & Hardin, C.D. (1996). Automatic stereotyping. *Psychological Science, 7*, 136–141.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für experimentelle Psychologie*, *48*(2), 145-160.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: the impact of social roles on controlled and automatically activated attitudes. *Journal of personality and social psychology*, *87*(1), 5.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Beale, F. (1970). Double jeopardy: To be Black and female. In T. Cade (Ed.), *The Black woman* (pp. 109–122). New York, NY: New American Library.

Becker, J., Kraus, M. W., & Rheinschmidt-Same, M. L. (2017). Cultural expressions of social class and their implications for beliefs and behavior. *Journal of Social Issues*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*(4), 329-353.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. *European journal of social psychology*, *38*(6), 977-997.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bowleg, L. (2008). When Black + lesbian + woman ≠ Black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex roles*, *59*(5), 312-325.

Brewer, M. B., Ho, H. K., Lee, J. Y., & Miller, N. (1987). Social identity and social distance among Hong Kong schoolchildren. *Personality and Social Psychology Bulletin*, *13*(2), 156-165.

Brown, R. J., & Turner, J. C. (1979). The criss-cross categorization in intergroup discrimination. *British Journal of Social and Clinical Psychology*, *18*, 371–383.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163-1224.

Cooper, B. (2015). Intersectionality. In L. Disch & M. Hawkesworth (Eds.), *The Oxford handbook of feminist theory*. New York, NY: Oxford University Press

Cole, E. R. (2009). Intersectionality and research in psychology. *American psychologist*, *64*(3), 170.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology*, *83*(6), 1314.

Crisp, R. J., Hewstone, M., & Rubin, M. (2001). Does multiple categorization reduce intergroup bias?. *Personality and social psychology bulletin*, *27*(1), 76-89.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological methods*, *14*(2), 81.

DeGraffenreid v. GENERAL MOTORS ASSEMBLY DIV., ETC., 413 F. Supp. 142 (E.D. Mo. 1976). https://law.justia.com/cases/federal/district-courts/FSupp/413/142/1660699/

De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of statistical software*, *31*(1), 1-30.

Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings. *European review of social psychology*, *1*(1), 263-292.

Dijksterhuis, A., & Van Knippenberg, A. D. (1996). The knife that cuts both ways: Facilitated and inhibited access to traits as a result of stereotype activation. *Journal of experimental social psychology*, *32*(3), 271-288.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*(3), 286.

Dugard, P., Todman, J., & Staines, H. (2010). *Approaching multivariate analysis. A practical introduction.* Second Edition. Routledge: New York.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, *27*(29), 6137-6157.

Eurich-Fulcher, R. & Schofield, J.W. (1995). Correlated versus uncorrelated social categorisations: the effect on intergroup bias, *Personality and Social Psychology Bulletin*, *21*, 149–159.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline?. *Journal of personality and social psychology, 69*(6), 1013.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of personality and social psychology*, *50*(2), 229.

Fiske, S. T, & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

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*(6), 1008-1025.

Gillath, O., Bahns, A. J., Ge, F., & Crandall, C. S. (2012). Shoes as a source of first impressions. *Journal of Research in Personality, 46*(4), 423-430. https://doi.org/10.1016/j.jrp.2012.04.003

Goff, P. A., & Kahn, K. B. (2013). How psychological science impedes intersectional thinking. *Du Bois Review: Social Science Research on Race*, *10*(2), 365-384.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology, 108,* 553-561.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. Political analysis, 22(1), 1-30.

Hewstone, M., Islam, M. R., & Judd, C. M. (1993). Models of crossed categorization and intergroup relations. *Journal of Personality and Social Psychology*, *64*(5), 779.

Horwitz, S. R., & Dovidio, J. F. (2017). The rich—love them or hate them? Divergent implicit and explicit attitudes toward the wealthy. *Group Processes & Intergroup Relations*, *20*(1), 3–31. https://doi.org/10.1177/1368430215596075

Islam, M. R., & Hewstone, M. (1993). Intergroup attributions and affective consequences in majority and minority groups. *Journal of Personality and Social Psychology*, *64*(6), 936.

Jaeger, B. (2017). *r2glmm: Computes R Squared for Mixed (Multilevel) Models*. R package version 0.1.2. https://CRAN.R-project.org/package=r2glmm

Jones, C. R., & Fazio, R. H. (2010). Person categorization and automatic racial stereotyping effects on weapon identification. *Personality and Social Psychology Bulletin*, *36*(8), 1073-1085.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in organizational behavior*, *29*, 39-69.

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

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of personality and social psychology*, *91*(1), 16.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

King, D. K. (1988). Multiple jeopardy, multiple consciousness: The context of a Black feminist ideology. *Signs: Journal of Women in Culture and Society*, *14*(1), 42-72.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, *110*(5), 675.

Kuznetsova, A., Brockhoff, P, & Christensen, R. (2017). *lmerTest package: Tests in linear mixed effects models*. Journal of Statistical Software, 82(13), 1-26. doi: 10.18637/jss.v082.i13

Landrine, H., Klonoff, E.A., Alcaraz, R., Scott, J., & Wilkins, P. (1995). Multiple variables in discrimination. In B. Lott & D. Maluso (Eds.), *The social psychology of intergroup discrimination* (pp. 183-224). New York: Guilford Press.

Levinson, J. D., & Young, D. (2010). Implicit gender bias in the legal profession: An empirical study. *Duke J. Gender L. & Pol'y*, *18*, 1.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology, 82*(1), 5-18. http://dx.doi.org/10.1037/0022-3514.82.1.5

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

Macrae, C., Bodenhausen, G. V., & Milne, A. B. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology*, *69*, 397–407. doi:10.1037/0022–3514.69.3.397

Mattan, B. D., Kubota, J. T., Li, T., Venezia, S. A., & Cloutier, J. (2019). Implicit Evaluative Biases Toward Targets Varying in Race and Socioeconomic Status. *Personality and Social Psychology Bulletin*, 0146167219835230.

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*(1), 45.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*(3), 455.

Moore-Berg, S., Karpinski, A., & Plant, E. A. (2017). Quick to the draw: How suspect race and socioeconomic status influences shooting decisions. *Journal of Applied Social Psychology*, *47*(9), 482-491.

Nicolas, G., de la Fuente, M., & Fiske, S. T. (2017). Mind the overlap in multiple categorization: A review of crossed categorization, intersectionality, and multiracial perception. *Group Processes & Intergroup Relations*, *20*(5), 621-631.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*(4), 565.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19,* 625–666.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of personality and social psychology*, *105*(2), 171.

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., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An ink- blot 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

Perszyk, D. R., Lei, R. F., Bodenhausen, G. V., Richeson, J. A., & Waxman, S. R. (2019). Bias at the intersection of race and gender: Evidence from preschool-aged children. *Developmental science*, *22*(3), e12788.

Petsko, C. D., & Bodenhausen, G. V. (2019). Multifarious person perception: How social perceivers manage the complexity of intersectional targets. *Social and Personality Psychology Compass*, e12518.

Pew Research Centre. (2019). Race in America 2019. Retrieved from https://www.pewsocialtrends.org/wp-content/uploads/sites/3/2019/04/Race-report_updated-4.29.19.pdf

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ransford, H. E. (1980). The prediction of social behavior and attitudes. In V. Jeffries & H. E. Ransford (Eds.), *Social stratification: A multiple hierarchy approach* (pp. 265–295). Boston: Allyn & Bacon.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological assessment*, *31*(12), 1395.

Richeson, J. A., & Ambady, N. (2001). Who's in charge? Effects of situational roles on automatic gender bias. *Sex Roles*, *44*(9-10), 493-512.

Rudman, L. A., Feinberg, J., & Fairchild, K. (2002). Minority members' implicit attitudes: Automatic ingroup bias as a function of group status. *Social Cognition*, *20*(4), 294-320.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic ingroup bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, *87*, 494–509.

Singh, R., Yeoh, B. S., Lim, D. I., & Lim, K. K. (1997). Cross-categorization effects in intergroup discrimination: Adding versus averaging. *British Journal of Social Psychology*, *36*(2), 121-138.

Schmid-Mast, M., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status. J*ournal of Nonverbal Behavior*, *28*, 145–165. https://doi.org/10.1023/b:jonb.0000039647.94190.21

Steele, J. R., George, M., Cease, M. K., Fabri, T. L., & Schlosser, J. (2018). Not always Black and White: The effect of race and emotional expression on implicit attitudes. *Social Cognition*, *36*(5), 534-558.

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology*, *50*(1), 33.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

Teachman, B. A., Gapinski, K. D., Brownell, K. D., Rawlins, M., & Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: the impact of providing causal information and evoking empathy. *Health psychology*, *22*(1), 68.

Teige-Mocigemba, S., Becker, M., Sherman, J. W., Reichardt, R., & Klauer, K. C. (2017). The affect misattribution procedure. *Experimental Psychology*, *64*(3), 215-230.

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*(10), 1427-1439.

van Oudenhoven, J. P., Judd, C. M., & Hewstone, M. (2000). Additive and interactive models of crossed categorization in correlated social categories. *Group Processes & Intergroup Relations*, *3*(3), 285-295.

Vanbeselaere, N. (1991). The different effects of simple and crossed categorizations: A result of the category differentiation process or of differential category salience?. *European review of social psychology*, *2*(1), 247-278.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of personality and social psychology*, *81*(5), 815.

Wigboldus, D. H., Holland, R. W., & van Knippenberg, A. (2004). Single target implicit associations. *Unpublished manuscript*.

Xu, K., Nosek, B., & Greenwald, A. (2014). Data from the race implicit association test on the Project Implicit demo website. *Journal of Open Psychology Data, 2*(1). doi:10.5334/jopd.ac

Yamaguchi, M., & Beattie, G. (2019). The role of explicit categorization in the Implicit Association Test. *Journal of Experimental Psychology: General*.