Intersectional Implicit Bias: Evidence for a Category Dominance Hierarchy and The Predominance of Target Gender

By

Paul R Connor

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Dacher Keltner, Chair
Professor Serena Chen
Professor Jack Glaser

Summer 2021

Abstract

Intersectional Implicit Bias: Evidence for a Category Dominance Hierarchy and The Predominance of Target Gender

by

Paul Connor

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Dacher Keltner, Chair

Individuals demonstrate implicit evaluative biases with respect to multiple dimensions of social categorization. However, little is known about how such implicit biases manifest toward targets displaying simultaneously intersecting social categories. Across four studies (N = 4,314) we used Single-Target IATs (Studies 1-4) and Evaluative Priming Tasks (Study 4) to test competing hypotheses concerning implicit evaluations of multiply categorizable targets varying in race, gender, social class, and age. Overall, we observed a dominant pro-female/anti-male bias, which accounted for more target-level variation in implicit evaluations than race-, class-, or age-related biases. We also documented smaller and less consistent pro-upper-class/anti-lower-class biases, and pro-Asian/pro-White/anti-Black racial biases. We observed little evidence of consistent interactions between social categories, or of effects differing between student samples (Studies 1-3) and a representative US sample (Study 4), or as a function of presenting targets as full-body or upper-body photographs (Studies 3 & 4). Taken together, these results suggest that implicit evaluations of multiply categorizable targets may operate according to a category dominance hierarchy, with a single category (here, gender) predominantly driving evaluations, but ancillary categories producing compounding levels of bias toward individuals displaying multiple stigmatized or positively-valued social identities.

Table of Contents

Intersectional Implicit Bias: Evidence for a Category Dominance Hierarchy and The Predominance of Target Gender

People display implicit evaluative biases—automatic associations between categories and positive or negative valance —with respect to a wide variety of social categories, including race, gender, social class, and age (Greenwald & Lai, 2020; Nosek, 2005). Evidence suggests that these biases have weighty social consequences, influencing decision making in contexts including employment, medicine, and voting (e.g., Greenwald, Banaji, & Nosek, 2015; Jost et al., 2009).

However, little is known about how implicit bias operates with respect to multiply categorizable social targets. In most human interactions, individuals display multiple intersecting social identities, most notably 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 measures of implicit bias have been designed to isolate and measure biases with regard to a single binary categorical preference (e.g., Black vs. White, female vs. male). For example, Nosek (2005) employed Implicit Association Tests (IATs; Greenwald, McGhee, & Schwartz, 1998) to demonstrate that US participants display implicit evaluative biases favouring Whites over Blacks, females over males, the rich over the poor, the young over the elderly, and many others. However, because IATs measure only a single categorical preference at a time, these methods 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? Or are some social categories more influential than others? Alternatively, 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 widespread advocacy of an intersectional approach within psychological science (e.g., Cole, 2009; Goff, & Kahn, 2013; Kang & Bodenhausen, 2015). There is, however, evidence that implicit evaluative biases can be simultaneously affected by 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 indicates that implicit evaluative biases are sensitive to multiple aspects of target stimuli. By implication, when targets are multiply categorizable—as in most everyday social interactions—it is likely that implicit evaluations will be shaped by multiple dimensions of social categorization. The central aim of the present investigation is to better understand this process.

Models of Intersectional Intergroup Bias

Several scholars have theorized about how multiple simultaneous social categorizations affect intergroup bias (for recent reviews, see Nicolas, de la Fuente, & Fiske, 2017, and Petsko & Bodenhausen, 2019). Here, we consider in detail select treatments, focusing upon those theories that make clear and testable predictions with regard to intersectional implicit bias.

Compounding Biases: Additive and Interactive Models

Perhaps the most prominent school of thought on intersectionality and intergroup bias is the thesis 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 can be found in 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 complex, interactive ways. Grounded in the writings of Black feminist activist Frances Beale (1970), Ransford (1980) proposed the *multiple jeopardy-advantage hypothesis:* when individuals are perceived as belonging to multiple stigmatized social categories, they are vulnerable to 'multiple jeopardy', a multiplicative negative evaluation that exceeds the sum of the negative biases associated with each category. By contrast, when individuals are perceived as belonging to multiple positively-valued social categories, the result can be 'multiple advantage' that exceeds the combined positive biases of the relevant categories (see also Almquist, 1975; King, 1988; Landrine, Klonoff, Alcaraz, Scott, & Wilkins, 1995). Crenshaw (1989) popularized these ideas (and introduced the term "intersectionality"), describing 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) Blacks (*DeGraffenreid v. GENERAL MOTORS ASSEMBLY DIV.*, 1976).

Today, scholarship animated by the concepts of intersectionality and multiple jeopardy has sought to understand the unique challenges faced by individuals possessing multiple marginalized social identities (especially those of Black women in the USA; Cooper, 2015). However, within this literature, it has not been clear whether intersectionality necessarily implies interaction (i.e., multiplicative) effects between social categories, or simply that individuals with multiple marginalized social identities suffer from multiple consequences of their identities. Indeed, scholars of intersectionality have at times been divided as to whether the concept can or should be reduced to these kinds of quantitative predictions (e.g., Cole, 2009; Bowleg, 2008).

Nonetheless, numerous researchers have attempted to quantify the simultaneous effects of multiple intersecting social categorizations on the expression of intergroup bias, and have often found evidence broadly consistent with both additive and interactive models of compounding biases. 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 most consistent with interaction effects producing 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 interaction effects producing multiplicative advantages stemming from combined positively-valued social identities (Brewer, Ho, Lee, & Miller, 1987; Eurich-Fulcher, & Schofield, 1995).

Thus, despite ambiguity regarding the presence and pattern of interaction effects, theories of compounding bias offer relatively clear predictions with regard to the specific sub-groups of multiply categorizable targets. Specifically, given prior evidence that Americans' implicit evaluative biases typically favour Whites over Blacks (Nosek, Banaji, & Greenwald, 2002), females over males (Richeson & Ambady, 2001, Rudman & Goodwin, 2004), the upper class over the lower class (Horwitz & Dovidio, 2017; Rudman, Feinberg, & Fairchild, 2002), and the young over the elderly (Nosek, 2005), theories of compounding bias 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 males, whereas the most positive biases should be displayed toward upper-class, younger White females.

Category Dominance

Other researchers have challenged the claim that biases will compound across multiple social categorizations.[1] One divergent perspective is the *category dominance model* (Macrae,

[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)

Bodenhausen, & Milne, 1995). This theory is premised on the notion that humans are by necessity 'cognitive misers' (Fiske & Taylor, 1991), who must strive to parse and interpret an overwhelming amount of social information as efficiently as possible. The category dominance model asserts that when social perceivers are faced with complex, multiply categorizable social targets, a single dominant categorization dimension will guide social perception and behavior. Which specific category becomes dominant depends on many factors, including the situational or chronic salience of different categories, the goals of perceivers, and/or perceivers' prejudices. But importantly, it is theorized that this dominant category, once activated, will subsequently 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, and 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 dimension will drive bias. Importantly, the category dominance model does not necessarily predict what the dominant category will be—if no specific category is primed by researchers, the choice of dominant category will in theory rely upon the perceivers' attention, goals, and pre-existing biases. Rather, this model predicts that due to the necessity of efficiently parsing an enormous amount of information from our social contexts, a single category will tend to emerge as dominant within specific contexts and suppress the effects of alternate categories.

Evidence Regarding Intersectional Implicit Bias

Only a handful of studies has 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 a compounding bias account, each target-level variable influenced perceptions, with participants displaying a greater tendency to associate Black, male, and adult targets with weapons than White, female, and child targets, respectively. Further, there was also 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, and found a race × gender interaction, with Black boys eliciting more negative evaluations than could be explained by main effects of race and gender alone.

Two further recent studies manipulated the race (Black and White) and perceived social class of targets within implicit bias tasks. In the first, 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). In the second, 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), and trained participants to associate the coloured backgrounds with higher or lower social class status. The results of this pair of studies varied, with five different patterns of results emerging from five separate experiments. However, one consistent result was that upper-class, White targets were favored within each observed pattern of bias (though not always more so than lower-class White targets or upper-class Black targets). Taken together, these results can be seen as consistent with theories of compounding bias, in that upper-class Whites were the most favored group across the full set of results.

By contrast, other studies have yielded results more consistent with the category dominance model, and its assertion that single categories will dominate responses when participants'

---

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.

attention is directed toward them. 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 participants categorized targets by profession, biases favoured Black athletes, but when participants categorized targets 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. For example, in one condition, Black female targets' race was made salient by presenting them alongside White females and males, in another, their gender was made salient by presenting them alongside White and Black males. 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. When targets were categorized according to gender, participants displayed pro-female/anti-male implicit gender bias, but little implicit racial bias.

Other 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. The authors concluded that "given sufficiently complex stimuli, the racial dimension may not always dominate categorization" (p. 1078).

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 the real world. However, current evidence concerning implicit bias and multiply categorizable targets remains limited, and ambiguous. Whereas some work supports theories of compounding bias, and suggests that implicit biases tend to compound across multiple social categories, other evidence aligns better with the category dominance model, and suggests that implicit evaluations are often driven by a single dominant categorical dimension.

Guided by these contrasting perspectives, we conducted four studies investigating the influences of multiple simultaneously displayed social categories upon implicit evaluative biases. In Study 1, we measured implicit evaluations of full-body target photographs of males varying in race (Black or White) and social class status. In Study 2, we developed a set of full-body target images varying simultaneously in race, gender, and social class, and pursued 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 shuffled targets' faces and bodies to achieve greater control over potential confounds, and varied the salience of categories by presenting targets via full-body or upper-body photographs. Finally, in Study 4, we tested the generalizability of our results by obtaining data from a nationally representative sample of US adults, and by comparing results across different methods of measuring implicit bias.

The present research offers theoretical, empirical, and methodological advances for the study of intersectional implicit bias. At the theoretical level, our work points toward a reconciliation between 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 varying in race, gender, social class, and age. And at the methodological level, the present work is, to our knowledge, the first investigation to focus specifically upon measuring and modelling

implicit evaluations of multiply categorizable targets at the individual target level. This method carries advantages over traditional approaches, as it allows researchers to study variation in implicit evaluations within target groups, which is not possible via target-group based analyses, while retaining the ability to assess of measurement reliability, which is not possible via response time-level analyses. All data and code used in the current project are accessible via the Open Science Framework (https://osf.io/sbpna/?view_only=645d6fea96f74ad5a59339da0920908e).

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. Here, 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 Blacks being evaluated the most negatively and upper-class Whites being evaluated the most positively. Additionally, they also suggest possible interaction effects, with either lower-class Blacks producing especially negative responses (multiplicative multiple jeopardy), or upper-class Whites producing especially positive responses (multiplicative multiple advantage). By contrast, 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) facing the camera with neutral expressions. Targets appeared on plain white backgrounds. Photographs were then presented to 1788 American 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 of income, race, and age, we assembled groups of eight photos each varying in race (Black and White) and income (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 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 $p > 0.32$).

[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). While 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 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.

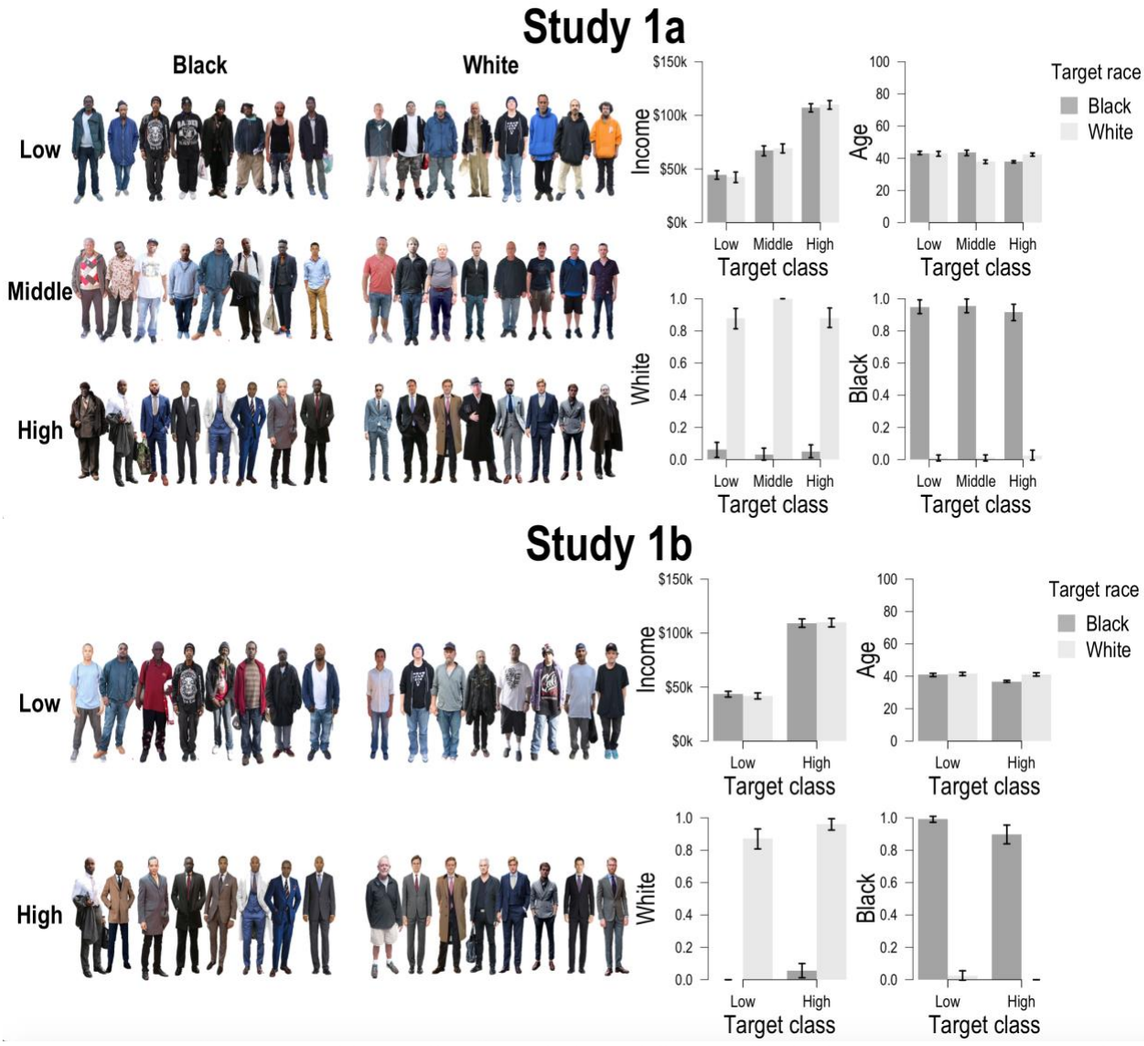

Figure 1. Target groups used in Studies 1a and 1b, and figures displaying raters' judgments of perceived income, age, and race ratings of each group. Bars indicate 95% confidence intervals.

Participants and Procedure

Participants for Study 1a (N = 298, 196 female, 2 missing gender data, M_{age} = 20.3, SD_{age} = 1.9, 129 Asian, 125 White, 29 Latino, 9 Black, 5 other race, 2 missing race data) and Study 1b (N = 533, 340 female, M_{age} = 20.5, SD_{age} = 2.63, 268 Asian, 173 White, 54 Latino, 6 Black, 10 Other race, 22 missing race data) were undergraduates who participated for course credit. Study 1a used a within-subjects design, measuring participants' implicit evaluations of all six target groups in a randomized order, whereas Study 1b used a between-subjects design, with participants randomly assigned to one of four target groups, and implicit methods used to measure both evaluations of the assigned target group and associations between target groups and wealth/poverty.

Single-Target IATs

We measured implicit evaluative bias via evaluative Single Target IATs (ST-IATs; Bluemke & Friese, 2008; Wigboldus, Holland, & van Knippenberg, 2004),[3] which measure the relative positivity of individuals' automatic responses toward a single target group. Each ST-IAT began with a practice block, in which the labels "Good" and "Bad" appeared at the top left and top right, respectively, 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

[3] 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.

either the top left of screens (in 'compatible' blocks), or the top right of screens (in 'incompatible' blocks), and participants then categorized words as "Good" or "Bad" or target photographs as a "Person." Participants were randomly assigned either to complete either two compatible blocks (of 20 then 40 trials) followed by two incompatible blocks (of 20 then 40 trials), or *vice versa* (see Table 1). Using the same procedure, in Study 1b we also used a wealth ST-IAT measuring 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 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 associations with each target group, we used the D Score summary measure (Greenwald, Nosek, & Banaji, 2003). On this measure, scores above/below zero indicate automatic associations between target groups and positive/negative concepts (in evaluative ST-IATs) or 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). In the present research we estimated the average split-half reliability of the valence and wealth ST-IATs to be 0.66[4] and 0.68, respectively (the valence ST-IAT figure combines data from Studies 1a and 1b). All implicit tasks in the present manuscript were administered online via Inquisit Web software.

Demographics

In both studies demographic information (age, gender, and race) 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 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: $F(2,594) = 19.16$, $p < .001$, $\eta^2 = 0.02$, Study 1b: $F(1,516) = 5.27$, $p = 0.02$, $\eta^2 = 0.01$, 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, $F(1,297) = 2.21$, $p = 0.14$, $\eta^2 = 0.001$, Study 1b, $F(1,516) = 2.47$, $p = 0.12$, $\eta^2 = 0.005$, nor any significant race × class interactions: Study 1a, $F(2,594) = 0.33$, $p = 0.72$, $\eta^2 < 0.001$, Study 1b, $F(1,516) = 0.58$, $p = 0.45$, $\eta^2 = 0.001$.

[4] 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).

Wealth ST-IAT

In the wealth ST-IAT in Study 1b, there was again a main effect of target class, $F(1,518) = 23.72$, $p < 0.001$, $\eta^2 = 0.04$, 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, $F(1,518) = 0.0008$, $p = 0.98$, $\eta^2 < 0.001$, and no significant race × class interaction, $F(1,518) = 3.13$, $p = 0.08$, $\eta^2 = 0.01$.

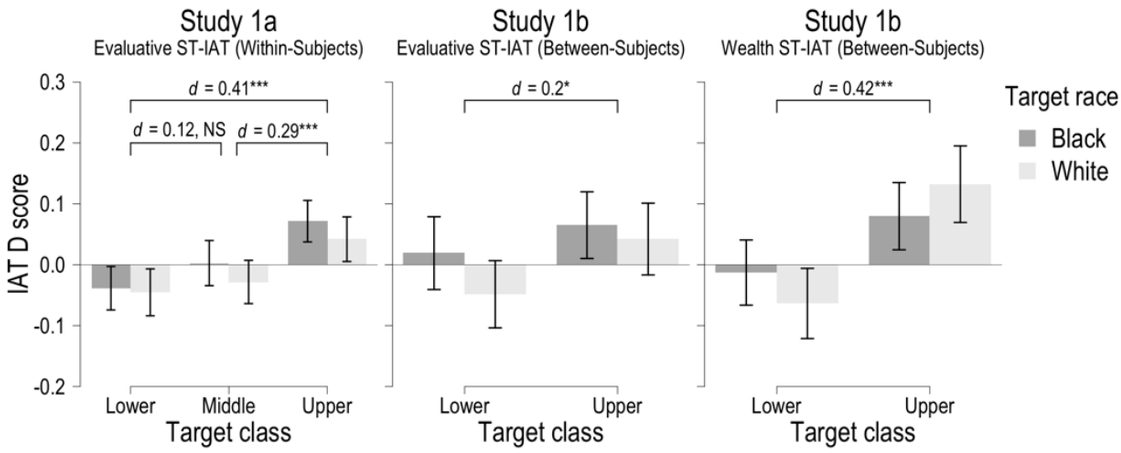

Figure 2. Mean IAT D scores by target group for Studies 1a and 1b. Bars indicate 95% confidence intervals. Cohens' *d* and statistical significance of t tests between social class groups collapsing across races are also reported (NS = not significant, * = $p < 0.05$, *** = $p < 0.001$).

Simulation-based power sensitivity analyses suggested that Analyses of Variances (ANOVAs) in both studies were well-powered to detect small main and interaction effects, though due to its within-subject design Study 1a achieved 80% power to detect smaller effects ($\eta^2 = 0.005$) than Study 1b ($\eta^2 = 0.015$). See Appendix L for details.

Discussion

In Studies 1a and 1b we found implicit evaluations of full-body male targets varying in race and social class to be driven by targets' social class. Across both studies, participants evaluated targets with higher perceived incomes more positively than targets of lower perceived incomes (though not significantly so in the case of the middle- and lower-class targets of Study 1). By contrast, participants' implicit evaluations were not significantly affected by target groups' race, nor did we observe any significant race × class interaction effects.

These results fail to align with theories of compounding bias, which predict that targets displaying multiple marginalized social identities (lower-class Blacks) and multiple positively-valued social identities (upper-class Whites) should elicit the most negative and positive evaluations, respectively. Instead, these results are more consistent with the category dominance model, with a single category (social class) emerging as dominant within the context of this particular study.

Notably, these results also diverged from previous empirical results 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). One explanation for this discrepancy is that these previous studies may not have held perceived social class was held constant across target groups representing different races; neither study reported evidence in this regard. 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. Another possibility, however, is that our methods may have amplified the salience of targets' social class compared to these past studies. For example, Moore-Berg and colleagues' use of a shoot/don't shoot paradigm may have increased the salience of race relative to class compared to our methods due to stereotypes associating Blacks with crime and violence (e.g., Devine & Elliot, 1995; Glaser & Knowles, 2008; Quillian & Pager, 2001). Additionally, our use of full-body target photographs may have

elevated 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—over evaluations, due to targets' bodies taking up relatively more of the visual space within our stimuli. Both previous studies used stimuli which devoted a more equal share of visual space to cues of race and class.

Study 2

In Study 1 we found implicit evaluations of Black and White male targets to be dominated by targets' social class. In the service of generalizability, and mindful of the category dominance model's assertion that different social categories will tend to become dominant in different contexts, in Study 2 we sought to extend this initial result, and tested participants' responses to targets varying more widely in terms of race (we incorporated Asian as well as Black and White targets), as well as on social class, gender, and age.

In addition, we sought to test whether the lack of pro-White/anti-Black implicit racial bias observed in Study 1 might have occurred not due to our specific methods, but simply due to our samples harboring little pro-White/anti-Black implicit bias due to their demographic makeup (Studies 1a and 1b used mostly female, and mostly Asian and Asian-American college students). To test this, we measured participants not just on their responses to multiply categorizable targets, but also on their implicit racial bias as measured by a traditional two-category Race IAT (Greenwald et al., 1998).

Toward a Target-Level Analysis: The Target D Score

Studying intersectionality beyond two dimensions via ST-IATs as in Study 1 encounters pragmatic methodological limitations. For example, studying targets 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) requires 24 separate experimental conditions. Using ST-IATs in this way is inefficient, however, as it ignores systematic variation in implicit evaluations within target groups. In Study 2, we therefore adopted a more efficient approach, by quantifying implicit evaluations at the level of each individual target. To accomplish this, we developed the *Target D Score*. This measure relies on a similar logic to a standard ST-IAT D Score, but instead of providing a measure of an individual participant's responses to a specific target group in compatible vs. incompatible trials, a Target D Score provides a measure of an entire sample's responses to a specific target in compatible vs. incompatible trials. Via Target D Scores, we were able efficiently model the simultaneous effects of a greater number of target-level variables than is possible via target-group-based approaches.

A Data-Driven Approach to Person-Perception

In considering which target-level variables to model, we were interested in studying responses to targets varying in race, gender, social class, and age. However, we did not wish to presume in advance how participants would perceive and categorize such complex social targets. We therefore followed recent work by Koch, Imhoff, Dotsch, Unkelbach, and Alves (2016), who developed an innovative approach to studying the spontaneous perception of complex social targets. In contrast to traditional approaches, which have typically involved asking participants to rate targets on pre-chosen traits, Koch and colleagues asked participants to judge how similar or different they considered targets to be to each other. These subjective difference ratings were then subjected to Multidimensional Scaling (MDS, for a review, see Borg & Groenen, 2005) to identify the primary dimensions underlying participants' judgments. In Study 2, we relied on this data-driven method to ascertain whether indeed race, class, gender and age shape implicit bias. Study 2 was pre-registered at https://aspredicted.org/87gw6.pdf.[5]

[5] 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 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 Appendix B).

Target Photographs

We selected 54 images (18 Asian, 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) gathered as part of our lab's ongoing research into person perception. In addition to the images, the database contains 490,359 explicit ratings of the targets made by 3,311 US adults (1,875 female, 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. Unsurprisingly, perfect orthogonality between each target-level variable (race, gender, class, and age) was not possible, with small correlations persisting between racial categorizations and perceived social class (see Table 2). However, these correlations were relatively small (maximum r = 0.15). Moreover, our analyses were able to statistically control for such imbalances by modelling effects at the individual target level. For example, we were able to estimate effects of targets' race while controlling for their perceived social class, and *vice versa*. Thus, in contrast to Study 1, which employed traditional factorial experimental designs, Study 2 employed something closer to a conjoint design, in which multiple variables are simultaneously manipulated, and their independent effects are parsed out via multivariate analyses (Hainmueller, Hopkins, & Yamamoto, 2014).

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 female, 24 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.[6] 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 were tasked with categorizing positive words or White faces via a single computer key and categorizing negative words or Black faces via an alternative key (in compatible trials), or with categorizing positive words or Black faces via a single computer key, and negative words or White faces via an alternative key (in incompatible trials). For each participant we computed D scores according to Greenwald and colleagues' (2003) algorithm, with higher D scores indicating relatively faster responses in compatible versus incompatible trials (indicating anti-Black implicit bias). The split-half reliability of the Race IAT D Scores was 0.75. Participants were randomly assigned to complete either the three ST-IATs or the Race IAT first.

Difference Ratings

Following the IATs, participants were presented with 60 randomly selected pairs of the 54 targets. For each pair, participants were asked "how different or similar are these people?" and provided ratings on 0-100 sliders ranging from "Very Similar" to "Very Different." This resulted in an average of 14.8 ratings (*SD* = 3.57) made of each of the 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). We tested multiple MDS solutions ranging from one dimension to six, and ultimately chose the five-dimension solution as 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 Appendix A for more information).

Next, to assess what the five dimensions represented, we calculated correlations between targets' scores on each dimension with the previously collected explicit trait ratings of each target within our database (Table 3). The first dimension correlated strongly with targets' perceived subjective SES (r = 0.91),[7] the second with categorization as Asian (r = -0.81) and categorization as Black

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

[7] 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.

(r = 0.79),[8] the third with categorization as White (r = 0.78), the fourth with categorization as Female (r = 0.81), and the fifth with age (r = 0.91). These results suggested that targets were spontaneously perceived as varying 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, correlations are bolded according to the dimension traits correlate most strongly with.

	MDS Dimensions				
	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)

However, it remained possible that the emerging dimensions might have aligned even more closely with some other non-measured variable. To test this, we asked participants in an separate student sample (N = 281, 193 female, M_{age} = 20.95, SD_{age} = 3.2, 140 Asian, 61 White, 23 Latino, 6 Black, 28 other race) to nominate via open-ended response what they perceived each dimension to represent, based upon visualizations of targets arranged according to their dimension scores (see Figure 3). Responses generally comported with the explicit trait rating correlations. Dimension 1 was described primarily as representing social class (e.g., "class", "wealth"), Dimension 2 race (e.g., "race", "skin"), Dimension 4 gender (e.g., "gender", "women"), and Dimension 5 age (e.g., "age", "old"). Notably, for Dimension 3, the most frequently chosen word was "gender," despite it having been correlated most strongly with categorization as White. However, this appeared to be due to the variety of words used to describe race. Of Dimension 3's twelve most commonly chosen words, descriptors of race were more frequent ("white", "race", "skin", "whiteness", and "caucasian" were chosen 68 times), than descriptors of gender ("gender", and "men" were chosen 48 times). Therefore, in what follows we refer to Dimension 3 as Race, with the caveat that this dimension also correlated with gender (we will come back to this point below).

[8] The fact that two race dimensions emerged—one (Dimension 2) separating Asians and Blacks, and the other (Dimension 3) separating Whites from Asians and Blacks—is sensible given that two linear dimensions are necessary to separate the three racial groups represented.

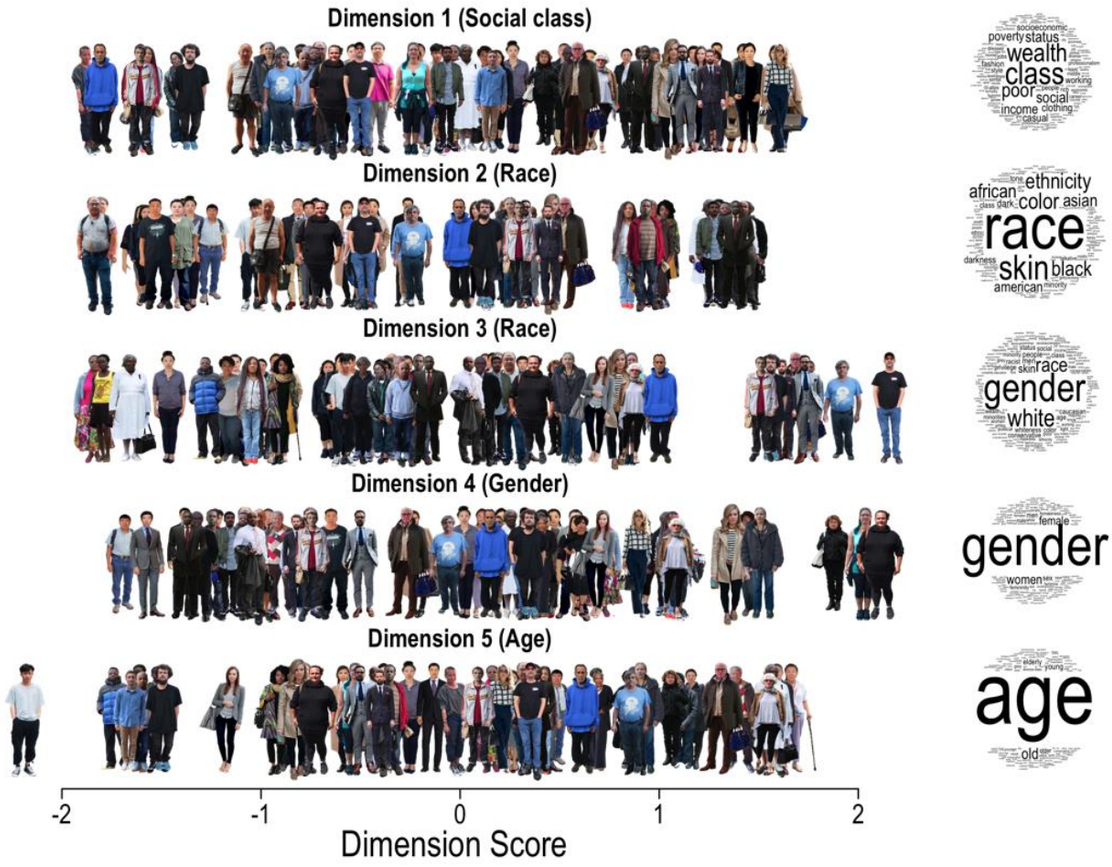

Figure 3. Study 2 targets arranged according to their scores on each of the 5 spontaneously emerging dimensions underlying relative similarity/dissimilarity judgments, alongside word clouds based upon the free-response text responses chosen to describe each dimension.

Calculating and Validating Target D Scores

To identify the optimal scoring algorithm for calculating Target D Scores, we undertook a data-driven process, testing different scoring procedures with regard to both their internal reliability (as indexed by split-half reliability estimates), and 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). This procedure is described in depth in Appendix K.

The scoring algorithm producing the greatest combined internal reliability and convergent validity[9] 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 (approximately 12% of all trials) and above 4000 milliseconds (approximately 0.02% of all trials), (c) penalizing error trials, in which the wrong computer key was pressed in response to a target (approximately 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 standardize these difference scores, they were then divided by the overall standard deviation of all logged response times between 100 and 4000 milliseconds. Like ST-IAT D Scores, Target D Scores above/below zero indicate that a sample evaluated a target relatively positively/negatively (i.e., responded relatively faster/slower to a target in compatible compared to incompatible trials).

This measurement algorithm resulted in an average split-half reliability of 0.57 across Target D Scores computed for every unique target used in this manuscript. This figure is lower than desirable, but close to the observed split-half reliability of the ST-IAT D Scores from Study 1

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

(0.66 and 0.68 for Studies 1a and 1b respectively), and higher than a recent meta-analytic estimate of the internal reliability of Fazio and colleagues' (1986) EPT (0.53, Greenwald & Lai, 2003).

As an initial test of the utility of Target D Scores, 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). Unsurprisingly, there was a significant positive correlation between Target D Scores and each target's mean income ratings, $r(67) = 0.35$, $p = .003$. However, targets' mean income ratings also remained a significant predictor of Target D Scores in a multiple regression controlling for targets' group membership, $\beta = 0.91(SE = 0.36)$, $t(58) = 2.58$, $p = 0.013$, $\eta^2 = 0.07$.[10] Thus, even within target groups, targets explicitly judged as appearing higher in income tended to be evaluated more positively than targets judged as appearing lower in income. This result provided initial validation of the idea of quantifying implicit evaluations at the individual target level, as this systematic within-group variation had previously been obscured by our reliance on traditional ST-IAT D Scores.[11]

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), $\hat{\beta}(SE_{\hat{\beta}}) = 0.06(0.02)$, $t(48) = 4.07$, $p < .001$, $\eta^2 = 0.19$, with bias favouring higher class over lower class targets. We also observed a significant effect of Dimension 3 (Race), $\hat{\beta}(SE_{\hat{\beta}}) = -0.04(0.02)$, $t(48) = -2.71$, $p = .01$, $\eta^2 = 0.08$, with bias favouring Asian and Black targets over White targets, and Dimension 4 (Gender), $\hat{\beta}(SE_{\hat{\beta}}) = 0.06(0.02)$, $t(48) = 3.89$, $p < .001$, $\eta^2 = 0.17$, with bias favouring female targets over male targets.

In a second model, we included two-way interactions between dimensions. Doing so 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, $\hat{\beta}(SE_{\hat{\beta}}) = 0.06(0.02)$, $t(39) = 4.29$, $p < .001$, $\eta^2 = 0.17$, 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 $\eta^2 = 0.10$ and two-way interaction effects of approximately $\eta^2 = 0.08$ (see Appendix L for details).

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

	Multi-Dimensional Scaling dimensions							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	η^2	r^2	$\hat{\beta}(SE_{\hat{\beta}})$	p	η^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**	**0.189**		**0.062(0.013)**	**<.001**	**0.175**	
Dimension 2 (Race[b])	0.002(0.015)	0.871	<.001		-0.001(0.014)	0.929	0.002	
Dimension 3 (Race[c])	**-0.041(0.015)**	**0.009**	**0.083**		**-0.037(0.014)**	**0.009**	**0.06**	
Dimension 4 (Gender[d])	**0.059(0.015)**	**<.001**	**0.172**		**0.059(0.013)**	**<.001**	**0.153**	
Dimension 5 (Age)	-0.008(0.015)	0.602	0.003		-0.013(0.013)	0.342	0.001	
Dimension 1 × Dimension 2					-0.023(0.017)	0.171	0.015	
Dimension 1 × Dimension 3					0.01(0.015)	0.526	0.003	
Dimension 1 × Dimension 4					**0.063(0.015)**	**<.001**	**0.144**	

[10] 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.
[11] Such within-target-group 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.

Dimension 1 × Dimension 5					-0.024(0.018)	0.173	0.015	
Dimension 2 × Dimension 4					-0.015(0.015)	0.335	0.007	
Dimension 2 × Dimension 5					0.014(0.013)	0.279	0.009	
Dimension 3 × Dimension 4					0.002(0.02)	0.928	<.001	
Dimension 3 × Dimension 5					-0.025(0.013)	0.056	0.03	
Dimension 4 × Dimension 5					0.002(0.015)	0.896	0	
				0.453				0.695
	Explicit target ratings							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	η^2	r^2	$\hat{\beta}(SE_{\hat{\beta}})$	p	η^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**	**0.07**		-0.028(0.025)	0.274	0.061	
Asian[f]	-0.041(0.038)	0.283	0.014		-0.063(0.049)	0.206	0.015	
White[f]	-0.054(0.038)	0.156	0.025		-0.069(0.048)	0.157	0.035	
Female[f]	**0.153(0.031)**	**<.001**	**0.296**		**0.127(0.048)**	**0.011**	**0.266**	
Age[g]	-0.023(0.016)	0.147	0.026		0.019(0.028)	0.498	0.018	
Social class × Asian					0.007(0.04)	0.865	<.001	
Social class × White					0.058(0.032)	0.072	0.032	
Social class × Female					**0.096(0.029)**	**0.002**	**0.107**	
Social class × Age					-0.016(0.016)	0.32	0.01	
Asian × Female					0.039(0.069)	0.579	0.003	
Asian × Age					-0.022(0.036)	0.534	0.004	
White × Female					0.023(0.067)	0.73	0.001	
White × Age					-0.068(0.035)	0.061	0.035	
Female × Age					-0.019(0.028)	0.503	0.004	
				0.423				0.632

Note: Statistically significant coefficients are bolded, 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

As a follow-up analysis, we probed whether the negative effect of Dimension 3 observed in our initial models— a bias favoring Asian and Black targets over White targets—might have occurred due to Dimension 3 capturing target gender as well as target race (see Table 3). To test this, we predicted Target D Scores from a *z*-scored social class composite measure averaging targets' *z*-scored mean ratings of subjective SES, occupational prestige, education, and income, (Cohen's α = 0.98), as well as binary indicators of Asian race, White race, and female gender[12], and *z*-scored mean ratings of targets' age. Results (Table 4) suggested significant effects of targets' perceived social class, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.04(0.02), t(48) = 2.45, p = .02, η^2= 0.07, with bias favouring higher class over lower class targets, and of targets' gender, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.15(0.03), t(48) = 4.96, p < .001, η^2= 0.30, with bias favouring female over male targets. Consistent with the idea that the previously observed effect of Dimension 3 had occurred due to its overlap with gender, there were no significant effects of target race. As with the MDS dimensions, there was no significant effect of target age.

In a second model, we included each two-way interaction between predictors (except between the two race indicators). This again significantly improved model fit, F(9,39) = 2.46, p = 0.02. Target gender was again 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, η^2= 0.11. The pattern of this interaction aligned with that observed for MDS Dimensions 1 and 4, and 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).

[12] 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.

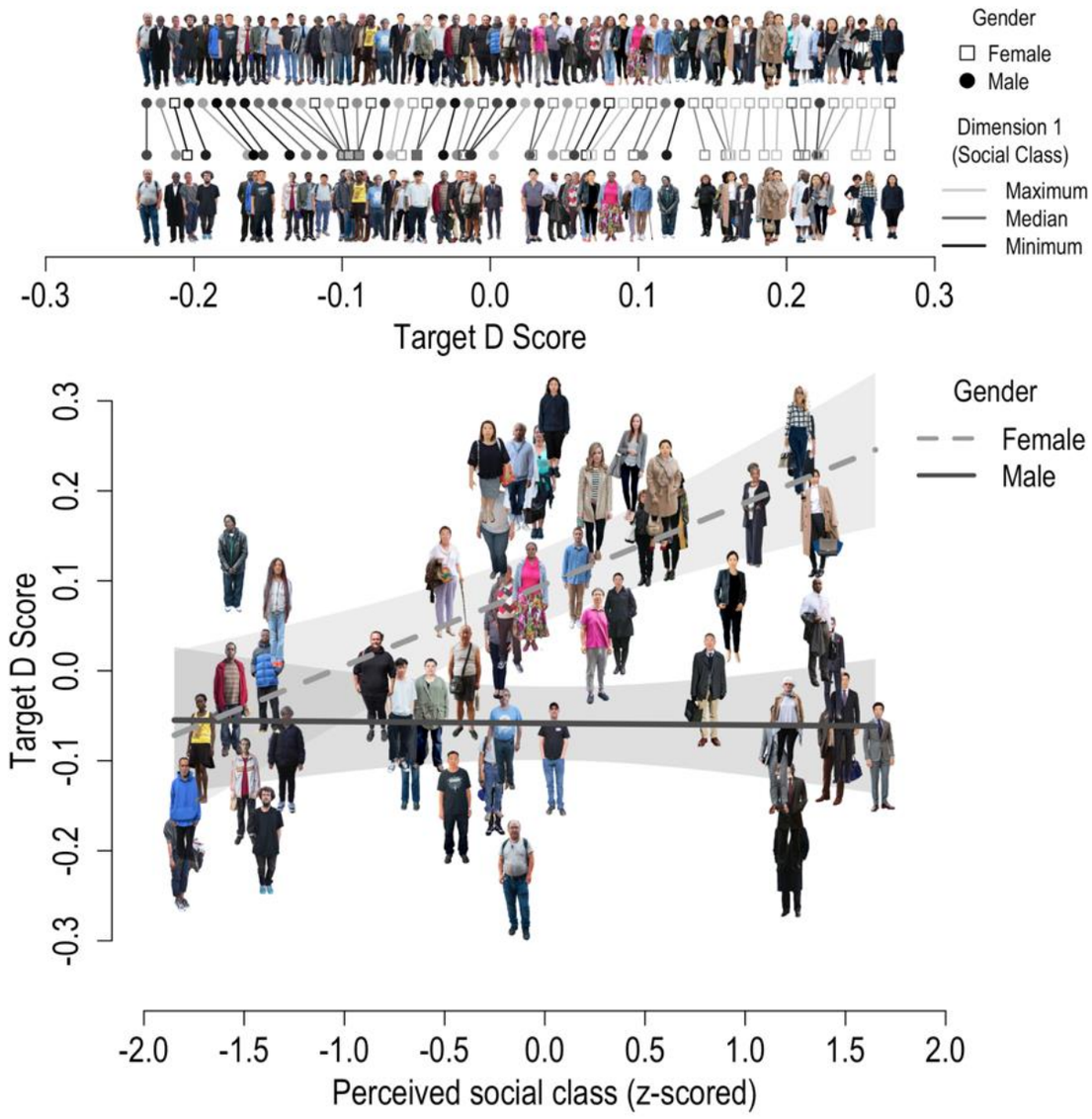

Figure 4. The top panel displays targets ordered by their Target D Scores (the row above) and arranged according to their exact Target D Scores (the row below). The bottom panel displays the interaction between targets' gender and perceived social class (a z-scored composite of targets' perceived income,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

Race IAT Results

Participants' responses on the traditional two-category Race IAT measure showed that the sample exhibited significant anti-Black/pro-White bias, with an average D Score of 0.30 (*SD* = 0.4), which was significantly above zero, $t(367) = 14.11$, $p < .001$, Cohen's $d = 0.75$, 95% CI = [0.26,0.34].

Discussion

In Study 2, we observed implicit evaluative biases toward the targets to be primarily driven by an interaction between gender and social class: upper-class female targets elicited especially positive evaluations. Further, this interaction emerged regardless of whether we predicted implicit evaluations from MDS dimension scores derived via Koch and colleagues (2016) data-driven procedure, or from explicit ratings of targets. By contrast, target race yielded more equivocal effects, with an apparent anti-White bias emerging within our MDS Dimension models failing to emerge when Target D Scores were predicted from targets' explicit race categorizations. We observed little effect of target age.

This pattern of results fails to align neatly with theories of compounding bias or the category dominance model. Although theories of compounding bias are consistent with especially positive evaluations of upper-class female targets, they offer little explanation as to why we observed

little evidence of anti-Black implicit racial bias in our ST-IATs (if anything, we observed weak evidence of anti-White bias). Second, although the category dominance model can make sense of equivocal race effects and absent age effects, as well as the relatively large effect of target gender within the explicit target ratings analyses, it does not provide an easy explanation of the interaction effects between categories, which require at least some participants to be sensitive to multiple categories at once.[13]

Also noteworthy is that despite showing little evidence of anti-Black bias within ST-IATs, our sample displayed a robust pro-White/anti-Black bias on the traditional two-category Race IAT. This suggests that the ST-IAT results cannot be explained as being simply a function of sampling bias. Rather, the explanation for the lack of anti-Black implicit racial bias in the ST-IATs must lie with the procedures used, in which participants responded to complex full-body target images, and in which participants were not explicitly focused upon any specific categorization dimension.

Study 3

In Study 3, we again measured implicit evaluations of targets varying in race, gender, social class, and age, but sought to improve our methodologies in a number of ways. First, we exerted tighter experimental control over our target stimuli to better guard against potential target-level confounding. In Studies 1 and 2, targets of different races appeared with different body shapes and in different clothes, and targets of different social classes exhibited different facial features. In Study 3, we swapped multiple target faces onto multiple different target bodies, thus holding body shape and clothing constant across target race categories, and holding faces constant across social class categories.

Second, in Studies 1 and 2, targets of different races were presented within separate ST-IAT tasks. This rendered it possible that participants may have selectively applied recoding strategies to ST-IATs containing targets of specific races, thereby perhaps suppressing implicit racial biases (e.g., Meissner & Rothermund, 2013). In Study 3, we avoided this by presenting race target groups together within ST-IAT tasks.

Third, we also sought to investigate 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 $p > 0.27$). There were also no significant differences in female or male categorization between race or age groups (all $F < 0.002$, all $p > 0.98$), no 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 $p > 0.97$).

[13] This is because if each participant's responses were dominated by a single category, participants whose responses were dominated by gender should have responded equally positively to both upper-class and lower-class female targets, and participants whose responses were dominated by social class should have responded equally positively to both female and male upper-class targets. Although such participants could collectively display main effects of both class and gender, they should not, in theory, display an interaction between the two categories

Bodies

We selected 24 unique bodies from the full-body photo database used in Study 2. Bodies were selected to vary in 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 rating data[14] in which each body was rated by an average of 84.1 raters (SD = 111.0), there were no significant differences in perceived attractiveness or racial prototypicality 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 within each age/gender subgroup to each of the 6 bodies within each age/gender subgroup. The 144 resulting stimuli were then assembled into six targets groups, each containing all 24 faces and 24 bodies, with each face attached to three low-SES and three high-SES bodies, and each face or body appearing in each target group only once. Each target 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 Appendix C for more details.

[14] It should be noted that ratings of each body were made with different, original faces attached to each body, rendering these data only a rough guide to the specific influence of the bodies themselves, rather than the original faces.

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 = 836, 590 female, 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 female, 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/blind.php?x=hy6if2,[15] Study 3b was pre-registered at https://aspredicted.org/blind.php?x=xb58b8.[16]

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

[15] After the original planned sample size was reached in Study 3a (N = 379), the split-half reliability of the Target D Scores remained low (0.37). We therefore decided to collect additional data, and re-pre-registered the study at https://aspredicted.org/blind.php?x=nz35zb. 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 added explicit ratings scales of targets' attractiveness, competence, political orientation, and photo blurriness. These changes had minor effects on the conclusions of the study (see Appendix F for more information).

[16] 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 Appendix F for more details.

following the procedures described above.[17] In Study 3a, participants viewed targets in upper-body presentation, in Study 3b, participants viewed targets in full-body presentation.

Difference Ratings

As in Study 2, 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, we considered this to be sufficiently established (see Appendix E for details), and so omitted the difference ratings from the additional data collected for Study 3a and from Study 3b.

Explicit Ratings of Targets

As a manipulation check, participants in Studies 3a and 3b rated each of 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

To ascertain whether we successfully manipulated the perceived race, gender, social class, and age of targets, we inspected correlations between participants' explicit ratings of the targets and targets' *a priori* categorizations 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). Additionally, there was relatively little non-orthogonality between these key variables; the highest inadvertent correlation was an association between targets' perceived social class and age, with ratings on the SES slider correlating weakly with ratings on the age slider (Study 3a r = 0.15, Study 3b r = 0.12). To control for this non-orthogonality, we again relied on target-level analyses, and modelled targets' social class and age as continuous variables, using z-scored mean explicit ratings.

Table 5

Correlations between a priori categorizations of targets 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

[17] 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.

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

To assess how targets' race, gender, social class, and age affected implicit evaluations, we calculated Target D Scores for each of the 144 unique targets (Study 3a split-half reliability = 0.54, Study 3b split-half reliability = 0.59). Because the same faces and bodies were shared by multiple targets, we fitted cross-classified hierarchical linear models (HLMs) predicting Target D Scores, 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 relied on the R packages lme4 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, Christensen, 2017).

Study 3a. First, we predicted Target D Scores from fixed effects of *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, $\hat{\beta}(SE_{\hat{\beta}}) = 0.10(0.02)$, t(18.85) = 4.30, $p < .001$, $\Delta r^2 = 0.06$[18], and White targets, $\hat{\beta}(SE_{\hat{\beta}}) = 0.09(0.02)$, t(18.69) = 4.07, $p < .001$, $\Delta r^2 = 0.05$, evaluated more positively than Black targets (for the simultaneous addition of both race dummies $\Delta r^2 = 0.07$). 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, $\hat{\beta}(SE_{\hat{\beta}}) = 0.2(0.02)$, t(13.36) = 8.71, $p < .001$, $\Delta r^2 = 0.43$. 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, $\chi^2(9) = 7.52$, $p = 0.58$, so we relegate these results to Appendix D. 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, $\hat{\beta}(SE_{\hat{\beta}}) = 0.2(0.02)$, t(26.64) = 6.23, $p < .001$, $\Delta r^2 = 0.26$, 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*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*

[18] Δr^2 refers to differences in r^2 values (Edwards, Muller, Wolfinger, Qaqish, & Schabenberger, 2008) between full models and models with each predictor removed.

Fixed effects								
(Intercept)	-0.129(0.021)	<.001			-0.107(0.032)	0.002		
Social class	0.007(0.011)	0.569	<.001		-0.026(0.045)	0.563	<.001	
Asian	**0.096(0.022)**	**<.001**	**0.058**		0.075(0.038)	0.059	0.014	
White	**0.091(0.022)**	**<.001**	**0.053**		0.041(0.062)	0.514	<.001	
Female	**0.2(0.023)**	**<.001**	**0.434**		**0.203(0.033)**	**<.001**	**0.26**	
Age	0.006(0.011)	0.598	<.001		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	<.001	
Competence					0.016(0.049)	0.739	<.001	
Liberal					-0.043(0.029)	0.143	0.004	
Blurry					0.016(0.013)	0.249	0.005	
			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*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects								
(Intercept)	-0.152(0.028)	<.001			-0.123(0.029)	<.001		
Social class	**0.044(0.016)**	**0.01**	**0.053**		0.022(0.042)	0.594	<.001	
Asian	**0.101(0.027)**	**0.001**	**0.044**		0.06(0.038)	0.117	<.001	
White	**0.092(0.026)**	**0.003**	**0.038**		0.037(0.061)	0.547	<.001	
Female	**0.232(0.033)**	**<.001**	**0.411**		**0.237(0.033)**	**<.001**	**0.342**	
Age	-0.009(0.016)	0.585	<.001		-0.006(0.017)	0.74	<.001	
Warmth					-0.016(0.022)	0.487	0.001	
Extroversion					-0.016(0.015)	0.297	0.003	
Attractiveness					0.035(0.026)	0.186	0.007	
Competence					-0.012(0.043)	0.772	<.001	
Liberal					-0.022(0.028)	0.425	0.002	
Blurry					**-0.044(0.012)**	**<.001**	**0.061**	
			0.556				0.617	
Random effects								
Face				0.026				0.012
Body				0.061				0.031
Residual				0.113				0.117

Note: Statistically significant coefficients are bolded, Black is the reference category for race contrasts

[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports 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 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, Δr^2 = 0.04, and White targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.09(0.03), t(18.41) = 3.46, p = 0.003, Δr^2 = 0.04, evaluated more positively than Black targets (for the simultaneous addition of both race dummies Δr^2 = 0.05), 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, Δr^2 = 0.41, 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, Δr^2 = 0.05. Targets' age did not significantly affect implicit evaluations.

As in Study 3a, adding two-way interactions did not significantly improve model fit, $\chi^2(9)$ = 11.99, p = 0.21, so we relegate these results to Appendix G. Also similar to Study 3a, target gender was again the only target-level demographic factor that remained a significant predictor after adding the control variables, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.24(0.03), t(23.46) = 7.31, p < .001, Δr^2 = 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, $p < .001$, $\Delta r^2 = 0.06$.

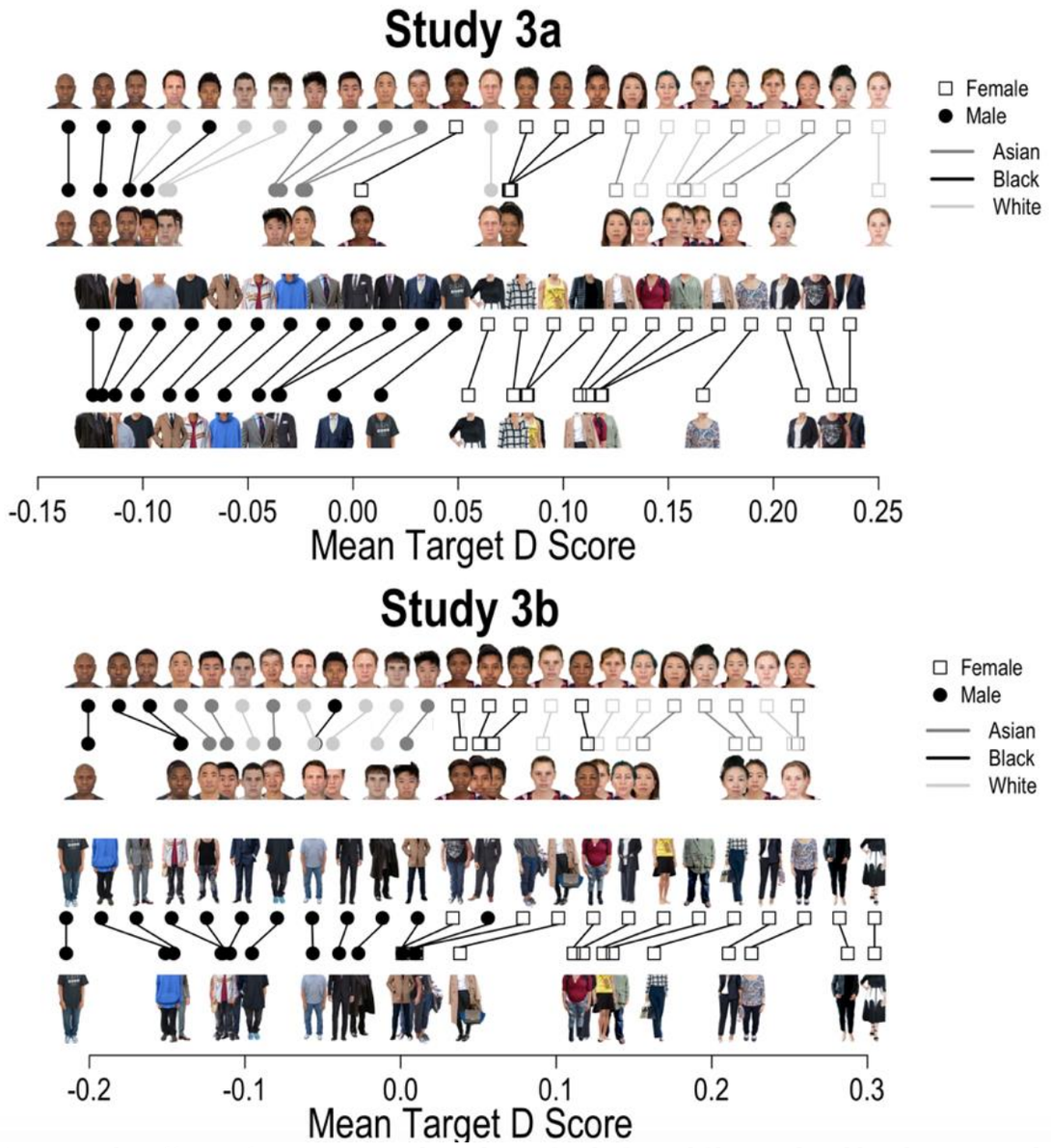

Figure 6. The effects of target race and gender in Study 3a and 3b visualized by showing each unique face and body arranged according to their mean Target D Scores (lower rows) and rank-ordered by their mean Target D Scores (upper rows).

Figure 6 displays each of the 24 unique faces and bodies according to their mean Target D Scores from Studies 3a and 3b. Particularly notable is the effect of gender, with faces and bodies nearly perfectly arranged according to gender. Also notable is that participants' gender bias was driven by both positive evaluations of female targets and negative evaluations of male targets: female faces and bodies typically elicited mean Target D Scores above zero, while male faces and bodies typically elicited mean Target D Scores below zero.

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. Study 3a achieved 80% power to detect main effects between approximately $\Delta r^2 = 0.05$ and $\Delta r^2 = 0.09$, and interaction effects between approximately $\Delta r^2 = 0.005$ and $\Delta r^2 = 0.035$. Study 3b achieved 80% power to detect main effects between approximately $\Delta r^2 = 0.04$ and $\Delta r^2 = 0.095$, and interaction effects between approximately $\Delta r^2 = 0.005$ and $\Delta r^2 = 0.03$ (for more details see Appendix L).

Discussion

In Study 3, we again measured implicit evaluations of targets varying in race, gender, social class, and age. To minimize target-level confounding, we used photo editing software to swap

different faces onto different bodies, and to assess the impact of target presentation, we presented targets as both upper-body (Study 3a) and full-body images (Study 3b).

Across both methods we observed a dominant effect of target gender. Gender uniquely explained approximately 43% and 41% of variance in Target D Scores in Studies 3a and 3b, respectively. By contrast, the next largest effect—target race—accounted for just 7% and 5% of variance in Target D Scores across the two studies. These findings are consistent with the category dominance model, which posits that responses to multiply categorizable targets will be driven by single specific categories. However, the category dominance model does not predict which category will dominate when participants are not primed or manipulated in specific ways, and beyond the single experiment discussed above conducted by Jones and Fazio (2010), little prior scholarship would have predicted target gender to drive implicit evaluations of multiply categorizable targets to such an extent.

Despite its dominant effect, however, we did not observe implicit evaluations to be driven solely by targets' gender. Also notable in Study 3's results were 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 provide at least some level of support for theories of compounding bias, as they suggest that implicit biases do combine additively, at least to some extent, across multiple social categories.

A number of other results of Study 3 were also noteworthy. First, the presence of anti-Black implicit racial bias in both studies was consistent with the idea that such biases may have been suppressed in Studies 1 and 2, possibly as a result of recoding effects (Meissner & Rothermund, 2013). Second, the significant effect of social class only for the full-body targets in Study 3b aligns with the idea that full-body target images may increase the relative salience of social class. Finally, it was notable that we did not replicate the interaction between target gender and social class observed in Study 2. Given that Study 3 incorporated a more tightly controlled experimental design than our previous studies, we believe the interaction in Study 2 likely emerged due to the idiosyncratic nature of the targets within social class/gender subgroups, and is therefore unlikely to reliably generalize to other contexts or stimuli.

Study 4

The results of Study 3 suggest that implicit evaluations of multiply categorizable social targets varying in race, gender, social class, and age may be primarily driven by targets' gender. However, two limitations of Study 3 motivated our final study. First, like our previous studies, Study 3 relied on non-representative samples of university students (71% and 75% female and 49% and 55% Asian, respectively). Second, like our previous studies, Study 3 relied solely on ST-IATs to measure implicit evaluations. Previous researchers have argued that different measurement procedures might produce different patterns of implicit biases toward multiply categorizable targets (Gawronski, Cunningham, LeBel, & Deutsch, 2010). This suggests the need to measure bias in different ways.

In Study 4 we sought to address both of these concerns by (a) recruiting a nationally representative sample of American adults, and (b) measuring implicit evaluations via both ST-IATs and Fazio and colleagues' EPT (Fazio et al., 1986). In addition, building on the suggestive results of Study 3, we tested the impact of viewing either full-body or upper-body target images.

Participants and Procedure

Participants were a sample of 1620 American adults nationally representative on gender, age, and race via Prolific (803 female, M_{age} = 38.58, SD_{age} = 14.2, 140 Asian, 1167 White, 103 Latino, 155 Black, 38 other race, 17 no race reported). Study 4 was pre-registered at https://aspredicted.org/blind.php?x=jv7549. As pre-registered, we excluded ST-IAT data from 9 participants and EPT data from 6 participants for having mean response times greater than 3000ms.[19]

[19] 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

ST-IATs

We used the same set of targets as Study 3. Participants were randomly assigned to one of the six target groups, and to view either full-body or upper-body presentation. Participants completed two consecutive ST-IATs as described above using their target group as stimuli.

Evaluative Priming Task

Participants also performed an EPT (Fazio et al., 1986). EPTs 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 2500ms gap between the presentation of each prime/target pairing. Participants took breaks after the 32nd and 64th trials, and proceeded when ready.[20] Participants were randomly assigned to complete either their ST-IATs prior to their EPT, or *vice versa*.

Explicit Ratings of Targets

Participants were asked to rate each of their 24 targets via 0-100 sliders on targets' 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

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 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 EPT data requires a different scoring algorithm compared to ST-IAT data, as applying the ST-IAT algorithm to the EPT data yielded Target D Scores with virtually zero internal reliability (see Appendix K for details). For the EPT data, 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, (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

---

100ms and 4000ms instead of 100ms and 6000ms, and by penalizing error trials. As reported in Appendix J, these deviations had little effect on our results.

[20] . 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).

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.

The internal reliabilities of the ST-IAT and EPT Target D Scores were both relatively low compared to our previous studies. This may have been due to careless responding, participants tiring across the two separate implicit bias tasks, or there being greater variability in individuals' idiosyncratic implicit biases among Prolific participants compared with student samples. Nonetheless, the raw correlation between ST-IAT and EPT Target D Scores was $r = 0.25$, which when corrected for attenuation via Spearman's formula (Murphy & Davidshofer, 1988), suggests an estimated true correlation between the two measures of $r = 0.76$. Thus, despite the unreliability of each Target D Score, each can arguably be seen as representing noisy indicators of a closely related construct. In light of this, we decided to diverge from our pre-registered analysis plan and averaged ST-IAT and EPT Target D Scores to create a composite Target D Score measure. This composite exhibited a higher internal reliability than the ST-IAT and EPT Target D Scores (0.48). In what follows we present results separately for ST-IAT, EPT, and also the composite Target D Scores.

Predicting Target D Scores

For each Target D Score (ST-IAT, EPT, and composite), we fitted a separate series of cross-classified HLMs. To test for differences between full-body and upper-body presentation, full-body and upper-body Target D Scores were included separately for each target in each model. As in Study 3, we included in each model random intercepts for targets' faces and bodies. 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. 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 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, $p < .001$, $\Delta r^2 = 0.06$. 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, $p < .001$, $\Delta r^2 = 0.37$, and target race, with both Asian targets, $\hat{\beta}(SE_{\hat{\beta}}) = 0.06(0.01)$, t(266.43) = 3.91, $p < .001$, $\Delta r^2 = 0.03$, and White targets, $\hat{\beta}(SE_{\hat{\beta}}) = 0.05(0.01)$, t(263.94) = 3.78, $p < .001$, $\Delta r^2 = 0.03$, evaluated more positively than Black targets (for the simultaneous addition of both race dummies $\Delta r^2 = 0.04$). 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, $p < .001$, $\Delta r^2 = 0.04$. 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 Appendix I).

EPT Target D Scores. For EPT 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, $\hat{\beta}(SE_{\hat{\beta}}) = 0.02(0.01)$, t(23.1) = 3.5, $p = .002$, $\Delta r^2 = 0.06$. 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, $p < .001$, $\Delta r^2 = 0.08$, 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, Δr^2 = 0.02, and White targets,[21] $\hat{\beta}(SE_{\hat{\beta}})$ = -0.04(0.02), t(273.54) = -2.34, p = .02, Δr^2 = 0.02 (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. 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 (for full results see Appendix I).

Composite Target D Scores. For the composite 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(22.51) = 6.13, p < .001, Δr^2 = 0.09. We also observed significant effects of target gender, with female targets evaluated more positively than male targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.10(0.01), t(19.84) = 10.95, p < .001, Δr^2 = 0.38, and target race, with both Asian targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.04(0.01), t(265.90) = 4.035, p < .001, Δr^2 = 0.04, and White targets, $\hat{\beta}(SE_{\hat{\beta}})$ = 0.03(0.01), t(263.53) = 2.34, p = 0.02, Δr^2 = 0.02, 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(272.54) = -1.69, p = 0.09. Targets' age also had no significant effect on implicit evaluations. 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, $\hat{\beta}(SE_{\hat{\beta}})$ = -0.03(0.01), t(260.73) = -3.07, p < .001, Δr^2 = 0.02. Model fit was not significantly improved by adding two-way interactions between the full-body target indicator and each of the target-level factors, $\chi^2(5)$ = 6.95, p = 0.22, or by adding two-way interactions between each of the target-level factors, $\chi^2(9)$ = 1.96, p = 0.99. All significant fixed effects from previous models remained significant after controlling for attractiveness and photo blurriness (see Table 7).

Simulation-based power sensitivity analyses suggested that Study 4 achieved 80% power to detect main effects of between approximately Δr^2 = 0.04 and Δr^2 = 0.07, and interaction effects of approximately Δr^2 = 0.025 (see Appendix L for details).

Table 7
Results from hierarchical linear models in Study 4

	ST-IAT Target D Scores							
	Model 1				Model 5			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD
Fixed effects								
(Intercept)	-0.119(0.012)	<.001			-0.09(0.013)	<.001		
Social class	**0.032(0.006)**	**<.001**	**0.064**		**0.032(0.011)**	**0.006**	**0.019**	
Asian	**0.056(0.014)**	**<.001**	**0.033**		**0.056(0.015)**	**<.001**	**0.029**	
White	**0.054(0.014)**	**<.001**	**0.031**		**0.054(0.016)**	**0.001**	**0.022**	
Female	**0.144(0.013)**	**<.001**	**0.368**		**0.144(0.015)**	**<.001**	**0.363**	
Age	-0.01(0.006)	0.137	0.005		-0.009(0.007)	0.197	0.002	
Full-body target					**-0.057(0.013)**	**<.001**	**0.038**	
Attractiveness					-0.002(0.013)	0.887	<.001	
Blurry					-0.008(0.007)	0.254	0.003	
			0.493				0.534	
Random effects								
Face				<.001				<.001
Body				0.011				0.01
Residual				0.1				0.096
	EPT Target D Scores							
	Model 1				Model 5			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD

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

Fixed effects								
(Intercept)	0.108(0.013)	<.001			0.109(0.014)	<.001		
Social class	**0.024(0.007)**	**0.002**	0.058		0.004(0.012)	0.756	<.001	
Asian	**0.032(0.015)**	**0.037**	0.018		**0.041(0.016)**	**0.011**	**0.024**	
White	-0.004(0.015)	0.813	<.001		0.014(0.018)	0.441	0.002	
Female	**0.054(0.013)**	**<.001**	0.081		**0.035(0.016)**	**0.039**	**0.027**	
Age	-0.003(0.007)	0.639	<.001		0.004(0.008)	0.566	<.001	
Full-body target					0.0002(0.014)	0.989	<.001	
Attractiveness					0.025(0.014)	0.07	0.012	
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

	Composite Target D Scores							
	Model 1				Model 2			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects								
(Intercept)	-0.005(0.009)	0.55			0.01(0.01)	0.329		
Social class	**0.028(0.005)**	**<.001**	**0.091**		**0.018(0.008)**	**0.036**	**0.009**	
Asian	**0.044(0.011)**	**<.001**	**0.038**		**0.048(0.011)**	**<.001**	**0.042**	
White	**0.025(0.011)**	**0.02**	**0.012**		**0.034(0.012)**	**0.007**	**0.015**	
Female	**0.099(0.009)**	**<.001**	**0.33**		**0.089(0.011)**	**<.001**	**0.302**	
Age	-0.006(0.005)	0.188	0.003		-0.002(0.005)	0.663	<.001	
Full-body target					**-0.029(0.01)**	**0.004**	**0.017**	
Attractiveness					0.012(0.01)	0.222	0.002	
Blurry					-0.008(0.005)	0.124	0.004	
			0.47				0.499	
Random effects								
Face				<.001				<.001
Body				0.004				<.001
Residual				0.075				0.074

Note: Statistically significant coefficients are bolded

[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports 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 both ST-IAT and EPT methods, plus a composite Target D Score measure comprised using both methods. Target gender again emerged as the dominant predictor of implicit evaluations, with female targets evaluated more positively than males, and target gender explaining the bulk of the explainable variation in Target D Scores. We also observed smaller but robust effects of target social class, with upper-class targets evaluated more positively than lower-class targets, and target race, with Asian targets evaluated more positively than Black targets across all Target D Scores, and White targets evaluated more positively than Black targets for ST-IAT and composite Target D Scores. We observed no significant effects of target age, no significant interactions between target-level factors, and no significant differences between upper-body and full-body target presentation, except for overall more positive evaluations overall of upper-body than full-body targets for the ST-IAT and composite Target D Scores.

Most notably, these results suggest that the dominance of gender in Study 3 was not due to non-representative sampling. In Study 4, we used a representative US sample with regard to race, gender, and age, and found target gender to uniquely explain approximately 37% of variance in ST-IAT Target D Scores. In the identical task used in Studies 3a and 3b, this figure had been similar (43% and 41% for upper-body and full-body targets, respectively).

Notably, however, the dominance of gender in implicit evaluations was most apparent within the ST-IAT and composite Target D Scores, and was less pronounced within the EPT results. This may suggest that different measurement techniques tend to elicit different results as to which target-level characteristics drive implicit evaluations, as has been previously argued (Gawronski

et al., 2010). However, given the unreliability of the EPT Target D Scores (split-half reliability = 0.28), and the general concordance of effects across the two tasks (preferences for females over males, Asians over Blacks, and the higher-class over the lower-class emerged via each method), we believe it is too soon to make any confident conclusions in this regard.

General Discussion

Implicit bias is central to the study of social cognition. Given that people are multiply categorizable, understanding the influences upon 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 social targets in naturalistic modes of presentation and categorizable in numerous ways, testing two competing theories about intersectionality. 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: upper-class targets were evaluated more positively than lower-class targets. In Study 2, we measured implicit evaluations of targets varying in race, gender, social class, and observed an interaction effect indicative of a specific positive bias toward upper-class females. In Study 3, with similarly intersectional targets, we explored the impact of portraying targets in full-body versus upper body photographs. 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. Finally, in Study 4 we replicated the results of Study 3 using a representative US sample with both ST-IAT and EPT measures of implicit evaluations. Across both measures, we observed Asian targets to be evaluated more positively than Black targets, upper-class targets to be evaluated more positively than lower-class targets, and again observed target gender to be the most important predictor of implicit evaluations, with female targets evaluated more positively than males.

We believe the present work makes a number of 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 to intersectionality. Consistent with category dominance (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 all align with the notion that when faced with complex social stimuli, social perceivers act as 'cognitive misers,' and implicit evaluations largely respond largely to a single dimension of social categorization.

However, our results are also consistent with the notion of that implicit biases compound—at least to some extent—across multiple categories. In Studies 3 and 4, which used the most tightly controlled set of targets, we observed relatively consistent effects of three separate target-level factors: gender, race, and social class. So, while we found little evidence for the kind of multiplicative interaction effects suggested by the multiple jeopardy-advantage hypothesis (Ransford, 1980), we did find the most negative implicit evaluations to be made toward individuals displaying multiple intersecting stigmatized social identities (in this case, lower SES Black males), and the most positive implicit evaluations to be made toward individuals displaying multiple intersecting positively-valued social identities (in this case, upper SES Asian and White 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 also appear to do so according to a category dominance hierarchy, with a single 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 yet another rationale for embracing the concept of intersectionality in psychological science. Often, arguments in the field highlight the importance of centering upon the experiences of individuals possessing multiple marginalized social identities, or the idea that social categories are likely to interact in unpredictable ways (e.g., Cole, 2009; Goff, & Kahn, 2013; Kang & Bodenhausen, 2015). Yet when responses to multiply categorizable targets are driven by a category dominance hierarchy, then this too may only be discoverable via intersectional research programs. For example, in past research on implicit evaluative bias, results have suggested that social class produces stronger effects on binary IAT tasks than race, gender, social class, or age (Nosek, 2005). However, our results suggests that unidimensional results such as these provide little guidance regarding the relative influence of each category when they are displayed simultaneously by social targets. Given that intersectionality is a fact of everyday social encounters, advancing understanding of how implicit bias operates in real-world contexts is likely to be severely limited by the absence of studying responses to such complex intersectional targets.

On an empirical level, we believe it is striking that gender emerged as the dominant driver of implicit evaluations. This finding was unexpected, but appears robust across student samples and a representative US sample, and has some precedent, with gender emerging as the sole significant predictor of categorization errors in a prior weapon identification task incorporating multiply categorizable targets (Jones & Fazio, 2010). However, this prior work involved both a relatively small and non-representative sample (79 college students), as well 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 provide a more robust demonstration of this dominant gender bias.

One explanation for this result is that 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 targets' faces and clothing. This may have made gender the most visually salient social category overall. But even if this is the underlying mechanism behind our results, this would not preclude gender's dominance from generalizing to real-world interactions, as in most everyday contexts individuals' faces and bodies/clothing are both visible.

It has long been established that individuals tend to 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 pattern of replicated results has attracted relatively little attention. However, its dominance in the present results suggests the greater attention to gender-based implicit evaluative bias might have an important role to play in building our understanding of the causes and consequences of implicit evaluative bias.

Finally, from a methodological perspective, we suggest 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. Most commonly, past work has measured and modelled implicit attitudes at the level of target groups, either by calculating stand-alone measures of evaluations of target groups (e.g., Jones & Fazio, 2010; Mitchell et al., 2003, Studies 4 & 5; Moore-Berg et al., 2017; Perszyk et al., 2019), or by quantifying one or more binary relative preferences between target groups (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 traditional target-group-based approaches, as well as allow for greater statistical control of target-level confounds.

A second approach used in prior research 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, Target D Scores provide additional advantages over these methods. First, 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. Second, unlike response time-level analyses, Target D Scores allow researchers to assess measurement reliability. This is important, as it allows researchers to distinguish between ranges of response times that contribute reliable information regarding implicit evaluations, and ranges of response times that contribute only unhelpful random noise.[22]

Some limitations regarding the present research should be noted. The first regards the question of why anti-Black bias was absent in Studies 1 and 2, but was present in Studies 3 and 4. As discussed above, one possibility is that because targets of different race were presented in separate ST-IATs in Studies 1 and 2, participants may have been able to use recoding strategies (Meissner & Rothermund, 2013) to suppress anti-Black bias in these studies. However, another possibility is that our method in Studies 1 and 2 of matching targets of different races on explicit ratings of perceived social class may have inadvertently created confounds between races. According to the causal attribution principle of *augmentation* (Kelley, 1973), the perceived importance of causes for specific outcomes is increased by the absence of other perceived causes of the same outcomes. A majority of Americans report believing that being White has a positive causal effect on the attainment of social class status (Pew Research Centre, 2019). Therefore, when Black and White targets are matched on explicit ratings of perceived social class, the Black targets may be judged as higher on other traits perceived as causal effects of social class status, such as competence, or industriousness. If so, such a mismatch could also have suppressed anti-Black bias in Studies 1 and 2 by globally increasing the relative positivity of responses to Black targets. Further research is needed to adjudicate between these competing explanations.

A second limitation is ambiguity regarding how to interpret discrepancies between the ST-IAT and EPT results in Study 4. Our ST-IAT data suggested a specific anti-Black bias compared with both Asian and White targets, but our EPT data suggested a specific pro-Asian bias compared with both White and Black targets, who were evaluated equivalently. Additionally, we found target gender to play a much more dominant role in the ST-IAT compared to the EPT. It is hard to know whether these inconsistencies represent real, reliable differences in how people respond to the same targets via these two different tasks, or whether they stem from the noisiness of Target D Scores in Study 4. More data are needed to answer this question, as well as to compare results from both the ST-IAT and EPT with other implicit methods, such as the Affect Misattribution Procedure (Payne et al., 2005).

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. However, real-world social targets vary on far more than just these four variables; modelling such complexity will require the study of other social variables, such as 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), and facial expressions (Steele et al., 2018).

Additionally, 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.

---

[22] 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 internal reliable measurement.

Finally, we have chosen to collapse across differences between participants and assess aggregated implicit evaluative biases toward targets as displayed by our participants as a whole. Yet we have little doubt that individuals also vary in important ways regarding the specific categories and sub-categories that most influence their implicit evaluations. 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*.

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.

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.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin, 21*, 1139–1150. doi: 10.1177/01461672952111002

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

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164-172.

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.

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

Marcus-Newhall, A., Miller, N., Holtz, R., & Brewer, M. B. (1993). Cross-cutting category membership with role assignment: A means of reducing intergroup bias. *British journal of social psychology*, *32*(2), 125-146.

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.

Murphy, K. R., & Davidshofer, C. (1988). *Psychological Testing: Principles and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

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

Quillian, L., & Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology*, 107, 717–767. doi: 10.1086/338938

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.

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*.

Appendix A: Study 2 Multi-Dimensional Scaling Results

As described in our manuscript, following the approach of Koch and Imhoff (2016), we subjected the distance matrix containing the mean perceived difference between each of the 1,431 unique target pairs to Multi-Dimensional Scaling using the majorization approach assuming an interval scale (SMACOF; De Leeuw & Mair, 2009). We ultimately chose the five-dimension solution as 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). Figure S1 displays the scaling stress and r^2 values of MDS solutions ranging between one and six dimensions.

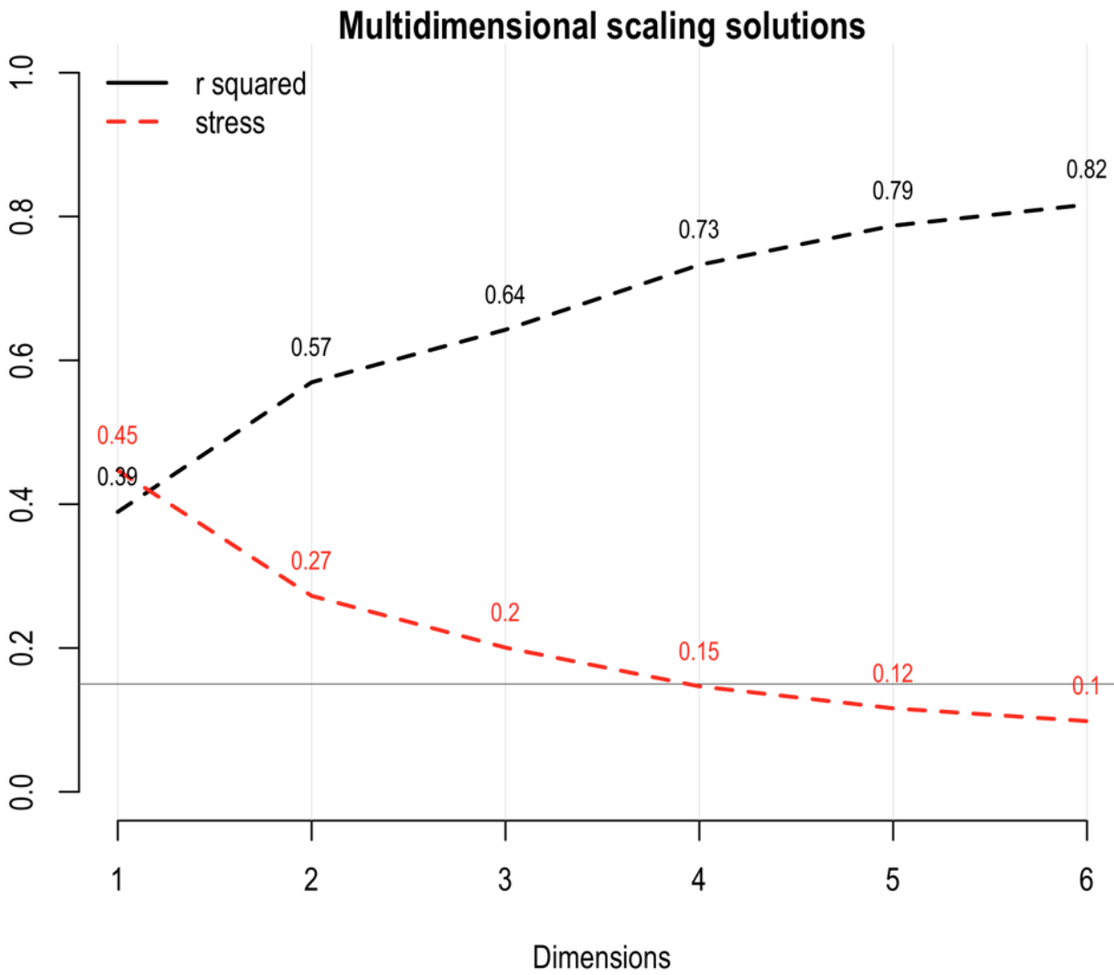

Figure S1. Scaling stress and r^2 values from MDS solutions fit to the distance ratings data of Study 2 ranging between one and six dimensions.

Appendix B: Study 2 Pre-Registered Analyses

At the time we pre-registered Study 2, we had not yet fully developed the Target D Score method, and so planned to analyze data at the level of individual logged response times. We pre-registered the following paragraph:

> …we will first identify the key dimensions underlying difference judgements following the technique of Koch, Imhoff, Dotsch, Unkelbach, and Alves (2016). We will then score each target on each identified dimension, and use those scores to predict (logged) response times in the single category IAT tests, with the key test being the interaction between dimension scores and a compatible/incompatible trial indicator. If participants tendency to exhibit relatively faster reaction times in compatible trials depends on the dimension scores, this will be interpreted to mean that dimension is associated with implicit bias.

We also pre-registered that "…we will exclude reaction times below 300 and above 10,000 ms" As these are the lengths of responses included in traditional two-category IAT D Scores. The model described in the pre-registration is therefore a cross-classified multilevel model predicting logged reaction times, in which reaction times are nested within both participants and targets, and the key fixed effects of interest are interaction terms between targets' scores on each multidimensional scaling dimension and a dummy indicating if a reaction time occurred in a compatible or incompatible trial (0 = compatible, 1 = incompatible). We did not pre-register a random effects structure, but we also included random intercepts for participants and targets, and random slopes on the incompatible dummy for both participants and targets, to account for random variation in (a) the overall positivity of participants toward all targets (b) the overall positivity of responses to each individual target. This results in the following model:

$$
\begin{aligned}
\boldsymbol{y_{ij} =\ } & \boldsymbol{\beta_0 + \beta_1 dimension\_1_j + \beta_2 dimension\_2_j + \beta_3 dimension\_3_j} \\
& \boldsymbol{+ \beta_4 dimension\_4_j + \beta_5 dimension\_5_j + \beta_6 incompatible_{ij}} \\
& \boldsymbol{+\ \beta_7 dimension\_1_j incompatible_{ij}} \\
& \boldsymbol{+\ \beta_8 dimension\_2_j incompatible_{ij}} \\
& \boldsymbol{+\ \beta_9 dimension\_3_j incompatible_{ij}} \\
& \boldsymbol{+\ \beta_{10} dimension\_4_j incompatible_{ij}} \\
& \boldsymbol{+\ \beta_8 dimension\_5_j incompatible_{ij} + \zeta_i incompatible_{ij}} \\
& \boldsymbol{+\ \zeta_j incompatible_{ij} + \varepsilon_i + \varepsilon_i + \varepsilon_{ij}}
\end{aligned}
$$

where *i* indexes participants and *j* indexes targets, y_{ij} is a logged response time of participant *i* toward target *j*, $incompatible_{ij}$ is a dummy variable indicating whether the response time occurred in a compatible or incompatible ST-IAT trial, $dimension\_1_j \ldots dimension\_5_j$ are target *j*'s score on dimensions 1 through 5, ζ_i and ζ_j are random slopes on $incompatible_{ij}$ at the participant and target levels, respectively, ε_i and ε_j are random intercepts for participants and targets, respectively, and ε_{ij} is the residual term. The results of this model are presented in Table S1 below.

Table S1
Hierarchical Linear Model predicting logged reaction times between 300 and 10,000ms in Study 2 ST-IATs

	$\hat{\beta}(SE_{\hat{\beta}})$	*p*
Fixed effects		
(Intercept)	6.26(0.009)	<.001
Dimension 1	-0.012(0.006)	0.045
Dimension 2	-0.001(0.006)	0.895
Dimension 3	-0.002(0.006)	0.772
Dimension 4	-0.01(0.006)	0.108
Dimension 5	-0.003(0.006)	0.568
Incompatible	0.026(0.01)	0.009
Dimension 1 × Incompatible	0.027(0.008)	0.002
Dimension 2 × Incompatible	0.002(0.009)	0.793
Dimension 3 × Incompatible	0.007(0.009)	0.423

Dimension 4 × Incompatible	0.025(0.009)	0.004
Dimension 5 × Incompatible	0.005(0.008)	0.571
	SD	*r*[a]
Random effects		
Participant (Random Intercept)	0.113	
Participant (Incompatible Random Slope)	0.088	-0.204
Target (Random Intercept)	0.000	
Target (Incompatible Random Slope)	0.006	*NA*[b]
Residual	0.491	

[a] *r* indicates the correlation between random slopes and intercepts at the participant and target level

[b] No correlation was computable between target random slopes and intercepts due to the lack of variation in random intercepts.

As shown in Table S1, when the data is analysed in this way, there are significant interaction effects between Dimensions 1 and 4 (the social class dimension, on which higher scores are associated with higher social class, and the gender dimension, on which higher scores are associated with female targets) and the incompatible dummy. These effects mirror the effects of social class and gender using the Target D Score method in our main manuscript. By contrast, there is no significant interaction between Dimension 3 (on which higher scores indicated White targets and lower scores indicated Black and Asian targets) and the incompatible dummy in the pre-registered analysis, but we did observe a significant effect of Dimension 3 using the Target D Score method. However, as we discuss in our manuscript, due to the overlap of Dimension 3 with both race and gender, we do not think the Dimension 3 effect should be interpreted as representing a reliable effect of race. This analysis therefore results in effectively the same conclusions as the target-level analysis based on the updated Target D Score algorithm presented in the main manuscript.

Appendix C: Study 3 Stimulus Creation and Pre-Testing

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. Figure S3 displays mean CFD norming data for the faces each race/age/gender subgroup on perceived attractiveness, perceived racial prototypicality, categorization as female, male, Asian, Black, and White, and perceived age.

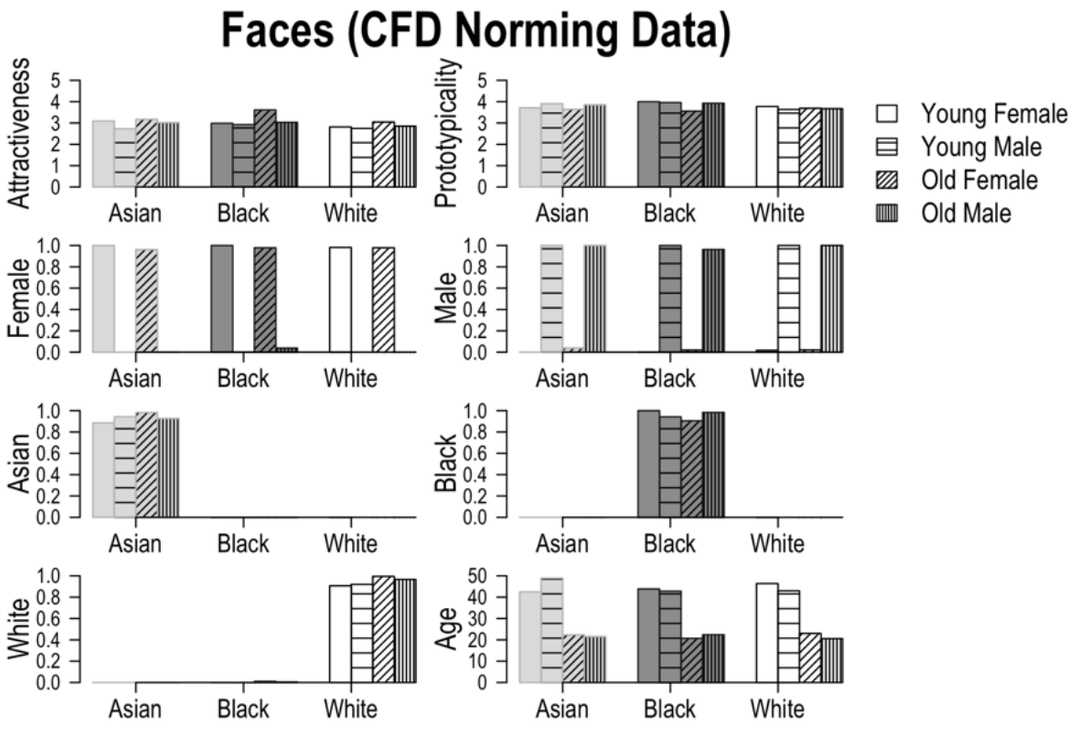

Figure S3. CFD norming data on chosen faces by race/gender/age subgroups. Female, Male, Asian, Black, and White, refer, respectively, to proportions of raters' binary categorizations of targets into each category.

Bodies

We selected 24 unique bodies from a large database of full-body photographs developed by our research lab for previous projects. Bodies were selected to vary in 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. Figure S4 presents data previously collected by our lab[23] for the bodies in each class/age/gender subgroup on perceived attractiveness, perceived age, perceived income, and perceived SES.

[23] It should be noted that ratings of each body were made with different, original faces attached to each body, rendering these data only a rough guide to the specific influence of the bodies themselves, rather than the original faces.

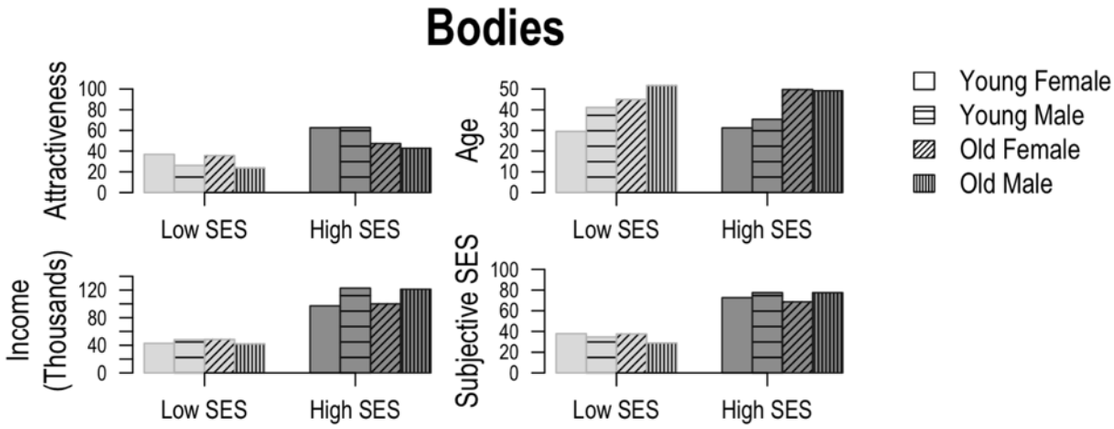

Figure S4. Explicit ratings data on chosen bodies by race/gender/age subgroups.

Appendix D: Study 3a Two-Way Interaction Model

Table S2 shows the results of the hierarchical linear model fit in Study 3a which included two-way interaction terms between target-level factors (race, gender, social class, and age). No two-way interactions were significant, and this model did not improve fit compared to a simpler model including only main effects of each factor, so these results were relegated to the Appendices to save space in the main manuscript.

Table S2
Results from hierarchical linear model including two-way interaction terms in Study 3a

	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD
Fixed effects				
(Intercept)	-0.113(0.026)	<.001	NA	
Social class	0.001(0.019)	0.948	<.001	
Asian	0.085(0.034)	0.026	0.023	
White	0.066(0.034)	0.075	0.014	
Female	0.171(0.036)	<.001	0.112	
Age	0.006(0.022)	0.78	<.001	
Social class × Asian	0.006(0.023)	0.806	<.001	
Social class × White	0.021(0.022)	0.347	0.001	
Social class × Female	-0.01(0.023)	0.669	<.001	
Social class × Age	-0.017(0.011)	0.138	0.012	
Asian × Female	0.026(0.048)	0.594	<.001	
Asian × Age	-0.002(0.024)	0.934	<.001	
White × Female	0.057(0.048)	0.262	0.003	
White × Age	0.008(0.026)	0.757	<.001	
Female × Age	-0.002(0.024)	0.944	<.001	
Model			0.542	
Random effects				
Face				0.019
Body				0.03
Residual				0.109

Note: Statistically significant coefficients are bolded
[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Appendix E: Study 3a Multi-Dimensional Scaling Results

During the initial data collection phase of Study 3a, we collected subjective difference ratings judgements for each of the 276 unique pairs of targets within each of the 24-target target groups. We did not obtain difference judgements for each possible unique pair of the 144 total targets used in the Study, as this would have required too much data (there are 10,296 unique target pairs among 144 targets), and would have led to participants viewing the same faces on multiple bodies, and the same bodies attached to multiple faces. The relative similarity/difference of each pair of targets was judged by an average of 13.21 participants (*SD* = 3.8).

Based on these difference ratings, we constructed six separate distance matrices for each of the target groups, and conducted six separate Multi-Dimensional Scaling (MDS) analyses on these. Similar to Study 2, these analyses were intended to assess the primary dimensions on which targets were perceived as differing. Unlike Study 2, we did not intend to use Dimension scores as regressors, but simply for the MDS to confirm that the targets were primarily perceived as differing on race, gender, class, and age. We pre-registered this in the following paragraph:

> We will first identify the key dimensions underlying difference judgements in each condition following the technique of Koch, Imhoff, Dotsch, Unkelbach, and Alves (2016). We predict that with some possible variability between conditions, the primary dimensions to emerge will be based on (1) targets' perceived SES (2) targets' perceived race (3) targets' perceived gender (4) targets perceived age, and that this will be demonstrated by correlating targets' scores on the identified dimensions with targets' mean scores on the explicit trait ratings. Assuming that the spontaneously used dimensions that emerge line up as expected with the explicit trait ratings, we will then use the targets' mean trait rating scores to predict (logged) response times in the single category IAT tests, with the key test being the interaction between the trait scores and a compatible/incompatible trial indicator.

Figure S2 displays the scaling stress and r^2 values of MDS solutions ranging between one and seven dimensions for each of the six target groups. This figure shows that for most target groups, MDS solutions of 4 or 5 dimensions resulted in acceptable fit.

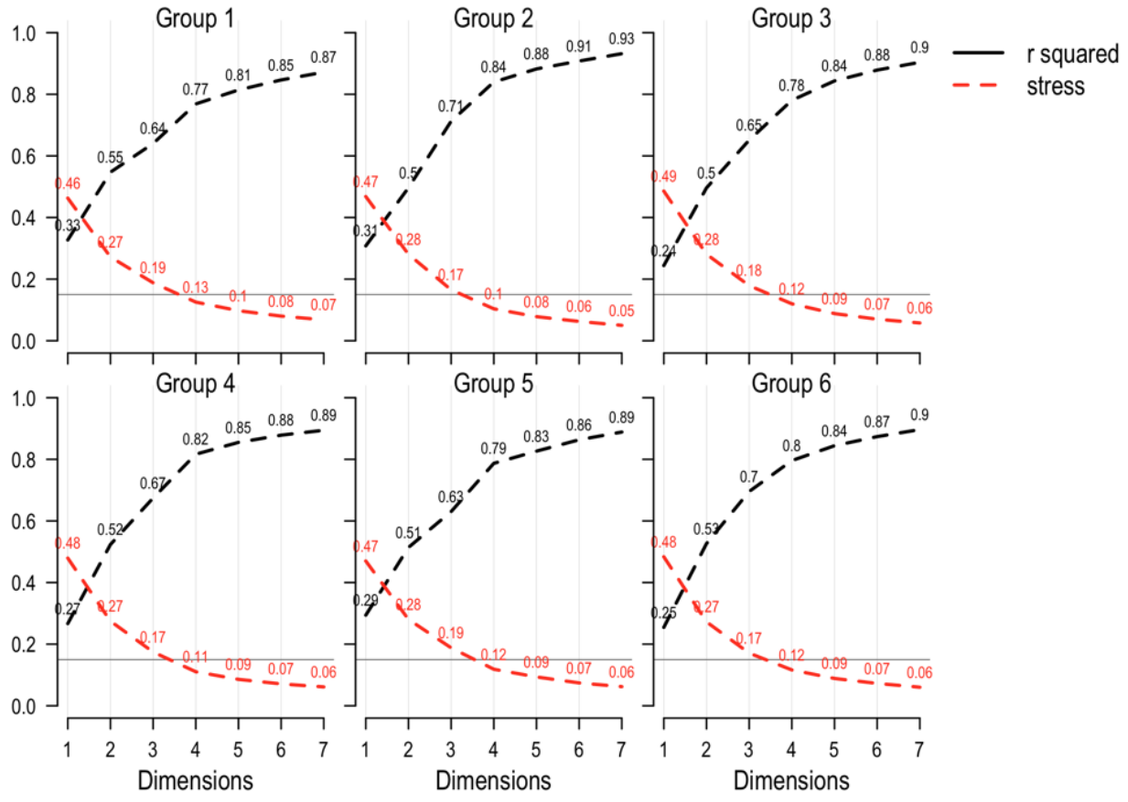

Figure S2. Scaling stress and r^2 values from MDS solutions fit to the distance ratings data of Study 3a ranging between one and seven dimensions.

To assess what the dimensions represented, we computed correlations between targets' dimension scores and targets' mean ratings on the measured explicit traits (male/female gender, Asian appearance, Black appearance, White appearance, SES, warmth, age, extroversion). These correlations are reported in Table S3.

Table S3
Target-level correlations between Multi-Dimensional Scaling dimension scores and mean explicit trait ratings from Study 3a

	Target Group 1						Target Group 2				
Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Black	-0.9					Black	0.92				
Warmth	-0.74		-0.44			Extroversion	0.52	-0.36			
Extroversion	-0.73					Male		0.87	-0.33		
White	0.61	0.51	0.47			Warmth	0.52	-0.57			
Asian	0.31	-0.82	-0.42			Asian	-0.57		-0.71	-0.3	
SES	-0.47	-0.56	0.53	0.36		White	-0.42	0.47	0.53	0.52	
Age						SES	0.34		-0.61	0.7	
Male			0.53	-0.78		Age					-0.8
	Target Group 3						Target Group 4				
Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Male	0.81	-0.56				Black	0.95				
Black	0.53	0.8				Extroversion	0.74		0.3		
Warmth		0.59		-0.31		Warmth	0.47	-0.44	0.35		
Extroversion		0.58	0.42			White	-0.44	0.73		0.41	
SES			0.93			Male		0.37	-0.87		
White		-0.31	0.41	0.77		SES		0.64		-0.68	
Asian	-0.32	-0.52	-0.31	-0.62		Asian	-0.53	-0.51	-0.32	-0.55	
Age				-0.31	0.86	Age					0.45
	Target Group 5						Target Group 6				
Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Rating	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Black	0.96					Black	0.94				
Extroversion	0.61			-0.33		Male		-0.95			
White	-0.37	0.81	-0.35			White	-0.37		-0.84		
Asian	-0.63	-0.7				Asian	-0.61		0.64	-0.4	
SES		-0.36	-0.76	-0.49		Warm	0.31	0.39	0.59		
Warmth	0.35	-0.45	0.62			SES	-0.42		0.48	0.67	
Male			-0.4	0.83		Extroversion	0.37			0.48	
Age					-0.8	Age					0.58

Note: correlations below $r = 0.3$ are suppressed. Bolded figures indicate the dimension on which each explicit trait loaded most strongly.

A similar dimension structure appeared for five of the six target groups (Target Groups 2-6): two dimensions most strongly correlated with targets perceived race, and the remaining three dimensions correlated most strongly with targets' perceived gender, SES, and age. The one exception to this general pattern was Target Group 1, for whom two race dimensions and a gender dimension clearly emerged, but no clear SES or age dimensions (although SES correlated strongly with Dimension 3, gender also correlated with Dimension 3 just as strongly, and SES in fact correlated more strongly with Dimension 2). However, we believe that this exception was more likely a result of noise in the data, rather than a systematic difference between Target Group 1 and the other groups. Overall, the consistency of the results for Target Groups 2-6, as well as their consistency with our previous MDS results from Study 2, suggested that as we assumed, race, gender, social class, and age were very likely the primary target-level variables spontaneously perceived in our Study 3a targets.

Appendix F: Study 3a Pre-Registered Analyses

In our original Study 3a pre-registration we stated the following with regard to our planned analyses:

> …Assuming that the spontaneously used dimensions that emerge line up as expected with the explicit trait ratings, we will then use the targets' mean trait rating scores to predict (logged) response times in the single category IAT tests, with the key test being the interaction between the trait scores and a compatible/incompatible trial indicator. If participants' tendency to exhibit relatively faster or slower reaction times in compatible compared to incompatible trials depends on the trait scores, this will be interpreted to mean that the trait is associated with implicit bias. Based on previous results, we predict that SES and gender will emerge as the key significant predictors of implicit bias against the targets, with responses more positive toward higher SES and more female targets, while race, age, warmth, and extraversion will show little relationship with implicit bias. We also specified that "we will exclude reaction times below 300 and above 10,000 ms."

The model described in the pre-registration is therefore a cross-classified multilevel model predicting logged reaction times, in which reaction times are nested within participants, target faces, and target bodies, and the key fixed effects of interest are interaction terms between targets' personal characteristics (SES, race, gender, age) and a dummy indicating if a reaction time occurred in a compatible or incompatible trial (0 = compatible, 1 = incompatible). We did not pre-register a random effects structure, but began by including random intercepts for participants, targets' faces, and targets' bodies, as well as random slopes on the incompatible dummy for participants, targets' faces, and targets' bodies, to account for random variation in (a) the overall positivity of participants toward all targets, (b) the overall positivity of responses to each individual face, and (c) the overall positivity of responses to each individual body. This results in the following model:

$$\begin{aligned} \boldsymbol{y_{ijk} =\ \beta_0 +\ \beta_1 SES_{jk} + \beta_2 Asian_{jk} + \beta_3 White_{jk} + \beta_4 Female_{jk} + \beta_5 Age_{jk}} \\ \boldsymbol{+ \beta_6 incompatible_{ijk} +\ \beta_7 SES_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_8 Asian_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_9 White_{jk} incompatible_{ijk} +\ \beta_{10} Female_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_8 dimension\_5_{jk} incompatible_{ijk} + \zeta_i incompatible_{ijk}} \\ \boldsymbol{+\ \zeta_j incompatible_{ijk} + \zeta_k incompatible_{ijk} +\ \varepsilon_i\ +\ \varepsilon_j\ + \varepsilon_k\ +\ \varepsilon_{ij}} \end{aligned}$$

where *i* indexes participants, *j* indexes targets' faces, and *k* indexes targets' bodies, y_{ijk} is a logged response time of participant *i* toward target *jk*, $incompatible_{ijk}$ is a dummy variable indicating whether the response time occurred in a compatible or incompatible ST-IAT trial, SES_{jk}, $Asian_{jk}$, $White_{jk}$, $Female_{jk}$, and Age_{jk} are targets' individual mean ratings on the SES slider (z-scored), dummies indicating targets' Asian race, White race, and female gender, and mean ratings on the age slider (z-scored), respectively, ζ_i, ζ_j and ζ_k are random slopes on $incompatible_{ijk}$ at the participant, target face, and target body levels, respectively, ε_i , ε_j and ε_k are random intercepts for participants, target faces, and target bodies, respectively, and ε_{ijk} is the residual term. However, this model failed to converge, necessitating a simpler random effects structure. Consequently, we removed ζ_j and ζ_k, and fit the following reduced model:

$$\begin{aligned} \boldsymbol{y_{ijk} =\ \beta_0 +\ \beta_1 SES_{jk} + \beta_2 Asian_{jk} + \beta_3 White_{jk} + \beta_4 Female_{jk} + \beta_5 Age_{jk}} \\ \boldsymbol{+ \beta_6 incompatible_{ijk} +\ \beta_7 SES_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_8 Asian_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_9 White_{jk} incompatible_{ijk} +\ \beta_{10} Female_{jk} incompatible_{ijk}} \\ \boldsymbol{+\ \beta_8 dimension\_5_{jk} incompatible_{ijk} + \zeta_i incompatible_{ijk} +\ \ \varepsilon_i} \\ \boldsymbol{+\ \varepsilon_j + \varepsilon_k\ +\ \varepsilon_{ij}} \end{aligned}$$

Fitting this model to the initial data collected for Study 3a (N = 379) results in the estimates presented in Table S4. As shown in the Table, at this stage the only significant effect was an interaction between target gender and the incompatible dummy, indicating a pro-female/anti-male bias.

Table S4
Hierarchical Linear Model predicting logged reaction times between 300 and 10,000ms in Study 3a initial data collection ST-IAT data (N = 379)

	$\hat{\beta}(SE_{\hat{\beta}})$	*p*
Fixed effects		
(Intercept)	6.262(0.017)	<.001
SES	-0.011(0.007)	0.115
Asian	0.003(0.017)	0.849
White	-0.006(0.017)	0.713
Female	-0.024(0.014)	0.088
Age	-0.009(0.007)	0.192
Incompatible	-0.031(0.02)	0.135
SES × Incompatible	0.009(0.01)	0.362
Asian × Incompatible	0.022(0.023)	0.343
White × Incompatible	0.036(0.023)	0.125
Female × Incompatible	0.048(0.019)	0.012
Age × Incompatible	0.006(0.01)	0.552
	SD	*r*[a]
Random effects		
Participant (Random Intercept)	0.184	
Participant (Incompatible random slope)	0.125	-0.399
Target face (Random Intercept)	0.009	
Target Body (Random Intercept)	0.00	
Residual	0.477	

[a] *r* indicates the correlation between random slopes and intercepts at the participant and target level

As discussed in the manuscript, at this point we decided to collect more data and re-pre-register Study 3a. This was chiefly due to our development of the Target D Score method, and the discoveries that (a) the response latency window of 300ms-10,000ms produced substantially less reliable Target D Scores than a response latency window of 100ms-6,000ms, and (b) even with the superior 100ms-6,000ms latency response window, the internal reliability of the Target D Scores from the initial data collected for Study 3a (r_{ab} = 0.23/r_{xx} = 0.37) remained substantially lower than the internal reliability achieved for Target D Scores in Study 2 (r_{xx} = 0.71). We considered this to be likely due to Study 3a using more targets per ST-IAT (24) than we had used in previous studies (8 per ST-IAT in Study 1 and 18 per ST-IAT in Study 2), and, perhaps, to our stimuli producing automatic valence associations with less overall variance as a result of holding facial attractiveness constant across social classes, and having less variation in body size and shape than Study 2's targets. We therefore felt it necessary to collect more data to reduce measurement error in Target D Scores. In our re-pre-registration we stated the following regarding our reasons for re-pre-registering:

> This pre-registration is actually an extension of a previous pre-registration (#35583). We are collecting more data for this project primarily due to issues with measurement accuracy. Specifically, for the Target D Scores, internal reliability is currently too low (r ~ 0.23). So even though the data is showing effects already of target-level characteristics (especially gender, with a strong anti-male bias emerging), we wish to try to increase the measurement accuracy of the target D scores to provide a more sensitive test of the other dimensions (race, class, age, etc).

The phrase "…the data is showing effects already of target-level characteristics (especially gender, with a strong anti-male bias emerging)" referred to models we had fit using Target D Scores based on the 100ms-6,000ms latency response window, not the model results presented above in Table S3. Regarding our modelling strategy, we stated the following:

> There will be two main kinds of analyses run: target-level analyses (which identify the target-level variables that predict implicit bias toward targets across the sample as a whole), and response-time level analyses (which enable tests of target x participant interactions). For the target-level analyses, we will predict Target D Scores (the unique level of implicit bias shown toward each of the 144 targets) in a series of linear models of increasing complexity. First, we will predict target D Scores from main effects of targets' race, perceived social class, gender, and age. Then, we will add 2-way and 3-way

interactions between these predictors. Finally, we will test if results hold controlling for perceived warmth, competence, attractiveness, political affiliation, and photo blurriness. For the response-time level analyses, we will explore whether participants' race, gender, social class, and political affiliation moderates effects of target-level variables.

Due to our evolving understanding of the optimal algorithm for calculating Target D Scores, we also specified that "…we have found that including response times between 100 and 6000ms leads to the greatest internal reliability of target D scores, so we plan to do the same in this analysis." This re-pre-registration therefore signalled our growing awareness that target-level analyses using Target D Scores are sufficient when all that is being examined are the effects of target-level variables, but that more complex response-time models are necessary to study more complex interactions between target-level characteristics and characteristics of participants. In the present project, we have omitted these analyses for simplicity, but hope to publish further research specifically aimed at exploring such participant-level moderators.

Yet while the re-pre-registration signalled our intention to run target-level models using Target D Scores, at that point our planned algorithm was different to the one we eventually used in the paper. This produced only minor differences in results. To display this, Table S5 presents results of the same models displayed in Table 6 of our manuscript, but predicting D Scores calculated via the pre-registered response latency window of 1000ms-6000ms, and including, rather than penalizing error trials. As shown in the model, the results are highly similar to the improved algorithm, though effect sizes are slightly smaller, which is consistent with using a slightly noisier measure.

Table S5
Results from hierarchical linear models in Study 3a using Target D Scores calculated using trials from 1000ms-6000ms and including error trials.

	Study 3a (upper-body targets)							
	Model 1				Model 3			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects								
(Intercept)	-0.121(0.023)	<.001			-0.101(0.034)	0.005		
Social class	0.014(0.012)	0.262	0.005		-0.026(0.045)	0.563	<.001	
Asian	0.081(0.024)	0.003	0.043		0.057(0.041)	0.172	0.007	
White	0.071(0.024)	0.008	0.033		0.029(0.066)	0.666	<.001	
Female	0.194(0.025)	<.001	0.415		0.197(0.034)	<.001	0.249	
Age	0.001(0.013)	0.95	<.001		-0.012(0.018)	0.517	<.001	
Warmth					0.009(0.023)	0.697	<.001	
Extroversion					0(0.018)	0.996	<.001	
Attractiveness					0.013(0.025)	0.589	<.001	
Competence					0.032(0.049)	0.517	<.001	
Liberal					-0.051(0.03)	0.089	0.007	
Blurry					0.02(0.014)	0.154	0.01	
			0.499				0.51	
Random effects								
Face				0.02				0.026
Body				0.04				0.043
Residual				0.107				0.105

Note: Statistically significant coefficients are bolded
[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Appendix G: Study 3b Two-Way Interaction Model

Table S6 shows the results of the hierarchical linear model fit in Study 3b which included two-way interaction terms between target-level factors (race, gender, social class, and age). No two-way interactions were significant, and this model did not improve fit compared to a simpler model including only main effects of each factor, so these results were relegated to the Appendices to save space in the main manuscript.

Table S6
Results from hierarchical linear model including two-way interaction terms in Study 3b

	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD
Fixed effects				
(Intercept)	-0.139(0.031)	<.001	NA	
Social class	0.041(0.023)	0.089	0.018	
Asian	0.058(0.037)	0.133	0.007	
White	0.088(0.037)	0.032	0.017	
Female	0.202(0.044)	<.001	0.155	
Age	-0.019(0.028)	0.512	<.001	
Social class × Asian	0.004(0.025)	0.872	<.001	
Social class × White	-0.024(0.023)	0.292	0.001	
Social class × Female	0.025(0.032)	0.435	<.001	
Social class × Age	0.015(0.015)	0.325	0.006	
Asian × Female	0.083(0.052)	0.133	0.006	
Asian × Age	-0.011(0.025)	0.683	<.001	
White × Female	0.006(0.052)	0.908	<.001	
White × Age	0.03(0.029)	0.31	0.002	
Female × Age	0.018(0.031)	0.562	<.001	
Model			0.577	
Random effects				
Face				0.023
Body				0.059
Residual				0.113

Note: Statistically significant coefficients are bolded
[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Appendix H: Study 3b Pre-Registered Analyses

In Study 3b, we originally pre-registered the following:

> There will be two main kinds of analyses run: target-level analyses (which identify the target-level variables that predict implicit bias toward targets across the sample as a whole), and response-time level analyses (which enable tests of target x participant interactions). For the target-level analyses, we will predict Target D Scores (the unique level of implicit bias shown toward each of the 144 targets) in a series of linear models of increasing complexity. First, we will predict target D Scores from main effects of targets' race, perceived social class, gender, and age. Then, we will add 2-way and 3-way interactions between these predictors. Finally, we will test if results hold controlling for perceived warmth, competence, attractiveness, political affiliation, and photo bluriness.

We also specified that "In previous work, we have found that including response times between 100 and 6000 ms leads to the greatest internal reliability of target D scores, so we plan to do the same in this analysis." This is only slightly different from the procedures we used. In our manuscript, we used a response latency of 100ms-4000ms, and penalize error trials, as based on further testing, we find this to provide the highest combined internal reliability and convergent validity (see Appendix K). Table S7 presents results using Target D Scores calculated using the pre-registered response latency window, and including, rather than penalizing, error trials. These results are extremely close to those presented in our manuscript for Study 3b in Table 6.

Table S7
Results from hierarchical linear models in Study 3b using Target D Scores calculated using trials from 1000ms-6000ms and including error trials.

	Study 3b (upper-body targets)							
	Model 1				Model 3			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects								
(Intercept)	-0.128(0.025)	<.001			-0.097(0.027)	<.001		
Social class	0.041(0.014)	0.01	0.049		0.029(0.04)	0.476	<.001	
Asian	0.08(0.025)	0.004	0.032		0.038(0.036)	0.287	<.001	
White	0.078(0.025)	0.005	0.031		0.018(0.058)	0.76	<.001	
Female	0.224(0.029)	<.001	0.421		0.231(0.03)	<.001	0.366	
Age	-0.012(0.014)	0.414	<.001		-0.005(0.016)	0.74	<.001	
Warmth					-0.019(0.021)	0.367	0.002	
Extroversion					-0.015(0.014)	0.273	0.002	
Attractiveness					0.039(0.025)	0.117	0.008	
Competence					-0.027(0.041)	0.518	<.001	
Liberal					-0.018(0.027)	0.494	<.001	
Blurry					-0.042(0.011)	<.001	0.059	
			0.55				0.613	
Random effects								
Face				0.019				0
Body				0.053				0.024
Residual				0.111				0.113

Note: Statistically significant coefficients are bolded
[a] Δr^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Appendix I: Study 4 Models 2-4

In our manuscript we describe three kinds of models fit for Study 4 that are relegated to Appendices to save space. These are models 2, 3 and 4 of our series of hierarchical linear models. Model 2 extends on Model 1, which included main effects of each target level characteristic: Social class (targets' *z*-scored mean rating on the subjective SES slider), race (two dummies indicating targets' Asian and White race, respectively, with Black as reference level), gender (a dummy indicating targets' female gender) and age (targets' *z*-scored mean rating on the age SES slider). Model 2 extended on this initial model by adding an indicator of whether a target was presented in upper-body or full-body presentation (0 = upper-body, 1 = full-body). Model 3 then added interaction terms between the full-body indicator and each target-level factor. As discussed in our manuscript, these interaction terms did not improve model fit for any of the Target D Scores: ST-IAT, EPT, or composite, so were removed. Model 4 added two-way interactions between each target-level factor. These interaction terms also did not improve model fit for any variety of Target D Scores, so were also removed. Table S8, S9, and S10 present results of Model 2, 3, and 4 for the ST-IAT Target D Scores, the EPT Target D Scores, and the composite Target D Scores, respectively.

Table S8
Results from hierarchical linear models using the ST-IAT Target D Scores in Study 4

	ST-IAT Target D Scores											
	Model 2				Model 3				Model 4			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects												
(Intercept)	-0.09(0.01)	<.001			-0.08(0.02)	<.001			-0.09(0.02)	<.001		
Social class	0.03(0.01)	<.001	0.06		0.03(0.01)	0.001	0.033		0.04(0.01)	<.001	0.036	
Asian	0.06(0.01)	<.001	0.032		0.06(0.02)	0.003	0.016		0.06(0.02)	0.002	0.021	
White	0.05(0.01)	<.001	0.03		0.06(0.02)	0.003	0.017		0.04(0.02)	0.028	0.01	
Female	0.14(0.01)	<.001	0.361		0.12(0.02)	<.001	0.269		0.14(0.02)	<.001	0.097	
Age	-0.01(0.01)	0.152	0.004		-0.01(0.01)	0.302	0.001		-0.02(0.01)	0.228	0.003	
Full-body	-0.05(0.01)	<.001	0.039		-0.07(0.02)	0.003	0.017		-0.05(0.01)	<.001	0.038	
Full-body × Social class					0(0.01)	0.961	<.001					
Full-body × Asian					0(0.03)	0.868	<.001					
Full-body × White					-0.01(0.03)	0.697	<.001					
Full-body × Female					0.05(0.02)	0.047	0.007					
Full-body × Age					0(0.01)	0.828	<.001					
Social class × Asian									-0.02(0.01)	0.289	0.001	
Social class × White									-0.02(0.01)	0.112	0.004	
Social class × Female									0.01(0.01)	0.529	<.001	
Social class × Age									0(0.01)	0.834	<.001	
Asian × Female									-0.01(0.03)	0.686	<.001	
Asian × Age									0.01(0.01)	0.607	<.001	
White × Female									0.02(0.03)	0.454	<.001	
White × Age									0.01(0.02)	0.644	<.001	
Female × Age									0(0.01)	0.762	<.001	
Model			0.532				0.536				0.532	
Random effects												
Face				0.00				0.00				0.00
Body				0.01				0.01				0.02
Residual				0.10				0.10				0.10

Note: Statistically significant coefficients are bolded
[a] Δ r^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Table S9
Results from hierarchical linear models using the EPT Target D Scores in Study 4

	EPT Target D Scores											
	Model 2				Model 3				Model 4			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects												
(Intercept)	0.11(0.01)	<.001			0.13(0.02)	<.001			0.11(0.02)	<.001		
Social class	0.02(0.01)	0.002	0.057		0.02(0.01)	0.027	0.028		0.02(0.01)	0.167	0.013	
Asian	0.03(0.02)	0.037	0.017		0.01(0.02)	0.631	<.001		0.03(0.02)	0.234	0.004	
White	0(0.02)	0.813	<.001		-0.02(0.02)	0.456	0.001		0.01(0.02)	0.529	0.001	
Female	0.05(0.01)	<.001	0.081		0.04(0.02)	0.054	0.025		0.06(0.02)	0.01	0.042	
Age	0(0.01)	0.643	<.001		-0.01(0.01)	0.324	0.003		0.01(0.01)	0.423	0.002	
Full-body	0(0.01)	0.847	<.001		-0.04(0.02)	0.083	0.011		0(0.01)	0.879	<.001	
Full-body × Social class					0(0.01)	0.938	<.001					
Full-body × Asian					0.04(0.03)	0.165	0.007					
Full-body × White					0.02(0.03)	0.417	0.002					
Full-body × Female					0.04(0.02)	0.139	0.008					
Full-body × Age					0.01(0.01)	0.363	0.003					
Social class × Asian									0.01(0.02)	0.649	<.001	
Social class × White									0.01(0.02)	0.717	<.001	
Social class × Female									0(0.02)	0.866	<.001	
Social class × Age									0(0.01)	0.796	<.001	
Asian × Female									0.02(0.03)	0.589	<.001	
Asian × Age									-0.02(0.02)	0.128	0.009	
White × Female									-0.03(0.03)	0.284	0.004	
White × Age									-0.01(0.02)	0.721	<.001	
Female × Age									-0.01(0.01)	0.558	<.001	
Model			0.167				0.186				0.181	
Random effects												
Face				0.00				0.00				0.00
Body				0.01				0.01				0.02
Residual				0.10				0.10				0.10

Note: Statistically significant coefficients are bolded
[a] Δ r^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Table S10
Results from hierarchical linear models using the Composite Target D Scores in Study 4

	Composite Target D Scores											
	Model 2				Model 3				Model 4			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects												
(Intercept)	0.01(0.01)	0.42			0.02(0.01)	0.071			0.01(0.01)	0.521		
Social class	0.03(0)	<.001	0.088		0.03(0.01)	<.001	0.048		0.03(0.01)	0.001	0.038	
Asian	0.04(0.01)	<.001	0.038		0.03(0.02)	0.023	0.01		0.04(0.02)	0.005	0.019	
White	0.03(0.01)	0.018	0.012		0.02(0.02)	0.147	0.004		0.03(0.02)	0.064	0.007	
Female	0.1(0.01)	<.001	0.329		0.08(0.01)	<.001	0.205		0.1(0.02)	<.001	0.098	
Age	-0.01(0)	0.201	0.003		-0.01(0.01)	0.17	0.003		0(0.01)	0.816	<.001	
Full-body	-0.03(0.01)	0.002	0.02		-0.06(0.02)	0.002	0.021		-0.03(0.01)	0.003	0.019	
Full-body × Social class					0(0.01)	0.994	<.001					
Full-body × Asian					0.02(0.02)	0.382	<.001					
Full-body × White					0.01(0.02)	0.749	<.001					
Full-body × Female					0.04(0.02)	0.02	0.011					
Full-body × Age					0(0.01)	0.612	<.001					
Social class × Asian									0(0.01)	0.708	<.001	
Social class × White									-0.01(0.01)	0.445	<.001	
Social class × Female									0.01(0.01)	0.572	<.001	
Social class × Age									0(0.01)	0.793	<.001	
Asian × Female									0(0.02)	0.914	<.001	
Asian × Age									-0.01(0.01)	0.456	<.001	
White × Female									-0.01(0.02)	0.794	<.001	
White × Age									0(0.01)	0.962	<.001	
Female × Age									0(0.01)	0.874	<.001	
Model			0.49				0.5				0.484	
Random effects												
Face				0				0				0
Body				0.01				0.01				0.01
Residual				0.07				0.07				0.07

Note: Statistically significant coefficients are bolded
[a] Δ r^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Appendix J: Study 4 Pre-Registered Analyses

When planning Study 4, we had a quite clear understanding of what analyses we would run for the ST-IAT data, and stuck closely to the pre-registered analysis plan. The only discrepancies

between what we pre-registered and the analyses we report in our manuscript are that based on our continued efforts to identify the optimal Target D Score scoring algorithm, we switched the response latency window from 100ms-6000ms to 100ms-4000ms, and penalized, rather than included, error trials. We also divided by a single standard deviation for all targets, rather than a target-specific standard deviation. Due to the low measurement reliability of the ST-IAT and EPT Target D Scores, we also averaged these two kinds of Target D Scores into a composite Target D Score measure, but make it clear in our manuscript that this was not planned ahead of time. We pre-registered the following concerning ST-IAT Target D Scores:

> For the ST-IAT data, we plan to calculate Target D Scores for each of the 288 unique targets (mean logged response time toward a target in incompatible trials minus mean logged response times toward a target in compatible trials, with this difference divided by the standard deviation of all logged responses to the target), and then predict these Target D Scores in a series of hierarchical linear models of increasing complexity. First, we will include main effects of targets' race, perceived social class, gender, and age. Then, we will add an indicator of stimulus modality (full-body vs upper-body), and its interactions with each of the target-level variables (race, gender, social class, and age). Then, we will add 2-way interactions between each of the target-level variables (e.g., race x gender, social class x age), then 3-way interactions between these 2-way interactions and the stimulus modality indicator. We may then explore higher-order interactions, but we are not well-powered to detect these and do not expect to find robust evidence for anything beyond 2-way interactions. All models will be cross- classified, including random intercepts for the 24 unique target faces and 24 unique target bodies used in the targets. We will also test if results hold controlling for targets' perceived attractiveness and photo bluriness.

We also specified the response latency window, and a decision rule for excluding participants, based on some feedback we had had in Studies 3a and 3b that Inquisit Online was experiencing technical issues for some participants, preventing them responding in a timely fashion. We pre-registered:

> In previous work, we have found that including response times between 100 and 6000 ms leads to the greatest internal reliability of target D scores, so we plan to do the same in this analysis. Also recently some participants have been having some technical difficulties with the Inquisit online platform, which has not been responding when they press the 'E' or 'I' keys on their keyboard. We will therefore also exclude participants whose average response time is above 3 seconds.

Tables S11 and S12 present results from all five models fit in the paper using the precise pre-registered Target D Score algorithm for the ST-IAT data. Table S11 presents Models 1 and 5, which were reported in Table 7 in our manuscript, and Table 12 presents Models 2, 3, and 4, which were relegated to Appendicese. These tables show that no conclusions were altered by using the improved version of the algorithm.

Table S11
Results from hierarchical linear models in Study 4 using the precise pre-registered scoring algorithm

	ST-IAT Target D Scores							
	Model 1				Model 5			
	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD	$\hat{\beta}(SE_{\hat{\beta}})$	p	Δr^{2a}	SD
Fixed effects								
(Intercept)	-0.115(0.012)	<.001			-0.087(0.013)	<.001		
Social class	0.032(0.006)	<.001	0.063		0.038(0.011)	<.001	0.026	
Asian	0.058(0.014)	<.001	0.036		0.055(0.015)	<.001	0.028	
White	0.05(0.014)	<.001	0.026		0.044(0.016)	0.007	0.014	
Female	0.147(0.012)	<.001	0.347		0.153(0.014)	<.001	0.325	
Age	-0.008(0.006)	0.196	0.003		-0.01(0.007)	0.138	0.003	
Full-body target					-0.054(0.013)	<.001	0.034	
Attractiveness					-0.011(0.012)	0.377	<.001	
Blurry					-0.009(0.006)	0.176	0.003	
			0.508				0.541	
Random effects								
Face				0.00				0.00
Body				0.00				0.00
Residual				0.098				0.096

Note: Statistically significant coefficients are bolded

[a] Δ r^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

Table S12
Results from hierarchical linear models in Study 4 using the precise pre-registered scoring algorithm

	ST-IAT Target D Scores											
	Model 2				Model 3				Model 4			
	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*	$\hat{\beta}(SE_{\hat{\beta}})$	*p*	Δr^{2a}	*SD*
Fixed effects												
(Intercept)	-0.09(0.01)	<.001			-0.08(0.02)	<.001			-0.09(0.02)	<.001		
Social class	0.03(0.01)	<.001	0.06		0.03(0.01)	<.001	0.034		0.04(0.01)	0.002	0.025	
Asian	0.06(0.01)	<.001	0.035		0.06(0.02)	0.003	0.016		0.07(0.02)	0.001	0.024	
White	0.05(0.01)	<.001	0.025		0.06(0.02)	0.002	0.017		0.04(0.02)	0.032	0.009	
Female	0.15(0.01)	<.001	0.328		0.12(0.02)	<.001	0.24		0.15(0.02)	<.001	0.103	
Age	-0.01(0.01)	0.223	0.002		-0.01(0.01)	0.211	0.002		-0.01(0.01)	0.294	0.002	
Full-body	-0.04(0.01)	<.001	0.03		-0.06(0.02)	0.004	0.015		-0.05(0.01)	<.001	0.03	
Full-body × Social class					0(0.01)	0.74	<.001					
Full-body × Asian					0(0.03)	0.992	<.001					
Full-body × White					-0.02(0.03)	0.498	<.001					
Full-body × Female					0.05(0.02)	0.02	0.01					
Full-body × Age					0(0.01)	0.687	<.001					
Social class × Asian									-0.01(0.01)	0.671	<.001	
Social class × White									-0.01(0.01)	0.413	<.001	
Social class × Female									0.01(0.01)	0.622	<.001	
Social class × Age									0(0.01)	0.768	<.001	
Asian × Female									-0.02(0.03)	0.565	<.001	
Asian × Age									0.01(0.01)	0.64	<.001	
White × Female									0.02(0.03)	0.58	<.001	
White × Age									0.01(0.02)	0.68	<.001	
Female × Age									0(0.01)	0.777	<.001	
Model			0.538				0.547				0.535	
Random effects												
Face				0.00				0.00				0.00
Body				0.00				0.01				0.01
Residual				0.1				0.10				0.10

Note: Statistically significant coefficients are bolded
[a] Δ r^2 differences in r^2 values between full models and models with each predictor removed, except the lowest value, which reports r^2 for the full model.

By contrast, we did not pre-register precisely how we would analyze the EPT data, because we had not used EPTs before, and were unsure if they would produce internally reliable Target D Scores. We pre-registered the following:

> For the EPT data, we are less certain of how we will proceed, but ideally, we will be able to calculate Target D Scores in a similar manner to the ST-IAT data, and fit a similar series of models to the ST-IAT data, in order to compare the results. We do not predict overall conclusions regarding the effects of target-level variables or presentation modality to be different from the EPT data as compared to the ST-IAT data, but we also have not used EPT data before in this context, so we feel less confident about its measurement properties (e.g., if the EPT Target D Scores are much more noisy than the ST-IAT Target D Scores, we will obviously not be able to detect the same effects with them).

As reported in the manuscript and below in the section ‘Assessing the Measurement Accuracy of Target D Scores, our data-driven approach suggested that the EPT data produced un-useable Target D Scores if we following the ST-IAT algorithm (nearly zero internal reliability), so we were forced to explore and ultimately use a different procedure for the EPT data. Our EPT analyses should therefore be considered exploratory.

Appendix K: Assessing the Measurement Accuracy of Target D Scores

Single-Target IATs

To identify the optimal algorithm for computing Target D Scores from ST-IATs, we used a data-driven approach aimed at maximizing the combined internal reliability and convergent validity of the Target D Scores. Our procedure for testing the internal reliability of different scoring algorithms was as follows. First, we separated data by Study. This left us with 6 separate sets of Target D Scores: Study 1a (48 full-body male targets varying in race and social class), Study 1b (32 full-body male targets varying in race and social class), Study 2 (54 full-body targets varying in race, gender, social class and age), Study 3a (144 upper-body targets varying in race, gender, social class and age), Study 3b (144 full-body targets varying in race, gender, social class and age), and Study 4 (288 upper-body and full-body targets varying in race, gender, social class and age). As described in our manuscript, each study used a slightly different measurement procedure. E.g., in Study 1, participants completed six ST-IATs containing 8 targets each with each ST-IAT containing targets from one specific race/class sub-group (e.g., lower-class White males). In Study 4, participants completed two ST-IATs containing a single target group consisting of 24 targets varying in race, gender, social class, and age.

We tested the split-half reliability and convergent validity of Target D Scores calculated via each of the possible combinations of the following algorithm parameters: (a) including error trials, in which participants pressed the incorrect response key, excluding error trials, or penalizing error trials by replacing their response time latency with participants' individual mean response latency in compatible/incompatible trials plus 600ms; (b) using logged or raw response times; (a) including error trials, in which participants pressed the incorrect response key, excluding error trials, or penalizing error trials by replacing their response time latency with participants' individual mean response latency in compatible/incompatible trials plus 600ms. E.g., if participant X committed an error in a compatible trial, the response time of the error was replaced with their mean response time in compatible trials plus 600ms; (c) setting the lower limit for inclusion of a response time to 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, or 400 milliseconds; (d) setting the upper limit for inclusion of a response time to 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, or 10,000 milliseconds; (e) standardizing response times by dividing difference scores by a target-specific standard deviation of response times, or an overall standard deviation of response times to all targets; (f) ignoring the block 1/2 distinction and computing a single Target D Score from all trials, or calculating an average between a Target D Score based on trials in Compatible block 1 and Incompatible block 1, and a Target D Score based on trials in Compatible block 2 and Incompatible block 2. .

Internal Reliability

To test the split-half reliability of Target D Scores, we randomly split the raw ST-IAT data in half, computed Target D Scores for each half of the data, and then computed and saved their bivariate correlations. For each of the combinations of algorithm parameters, six separate split-half correlations were calculated for each of the six separate sets of Target D Scores (Studies 1a, 1b, 2, 3a, 3b, and 4). This procedure was repeated 100 times, and the resulting split-half correlation figures were averaged for each Study/parameters combination. E.g., one-hundred split-half correlations were computed for the Study 2 Target D Scores excluding error trials, using logged response times, including response times between 200 and 3000 milliseconds, using a single standard deviation, and ignoring the block 1/2 distinction. These 100 split-half correlations were then averaged. Finally, for each combination of parameters (e.g., logged response times/excluded error trials/200ms minimum/3000ms maximum/single *SD*/ignoring blocks), we computed an overall average split-half correlation across the six different sets of Target D Scores, and converted this average estimate to an estimated split-half reliability via the Spearman-Brown prophecy formula (Revelle & Condon, 2019).

$$r_{xx} = \frac{2r_{ab}}{1 + r_{ab}}$$

In this formula, r_{xx}= the estimated split-half reliability and r_{ab}= the observed split-half correlation. The average split-half reliability estimates of parameter combinations are depicted in Figure S1. The highest average split-half reliability (r_{xx} = 0.56) was produced by (a) penalizing error trials; (b) using logged response times; (c) setting a minimum response time value of 100ms; (d) setting a maximum response time value of 4000ms; (e) using a single *SD* for all targets; (f) ignoring the block 1/2 distinction. However, as Figure S5 shows, some parameters were more important than others. A minimum response time of 100ms, for example, produced a far more internally reliable measure than the cut-off of 300ms used in calculating D Scores for standard two-category IAT D Scores (Greenwald, Nosek, & Banaji, 2003). By contrast, including response times higher than 4000ms made little difference when response times were logged (see the three plots on the left of Figure S5), and there was relatively little difference between penalizing and including error trials (see the overall similarity between the top and middle rows of plots in Figure S5).

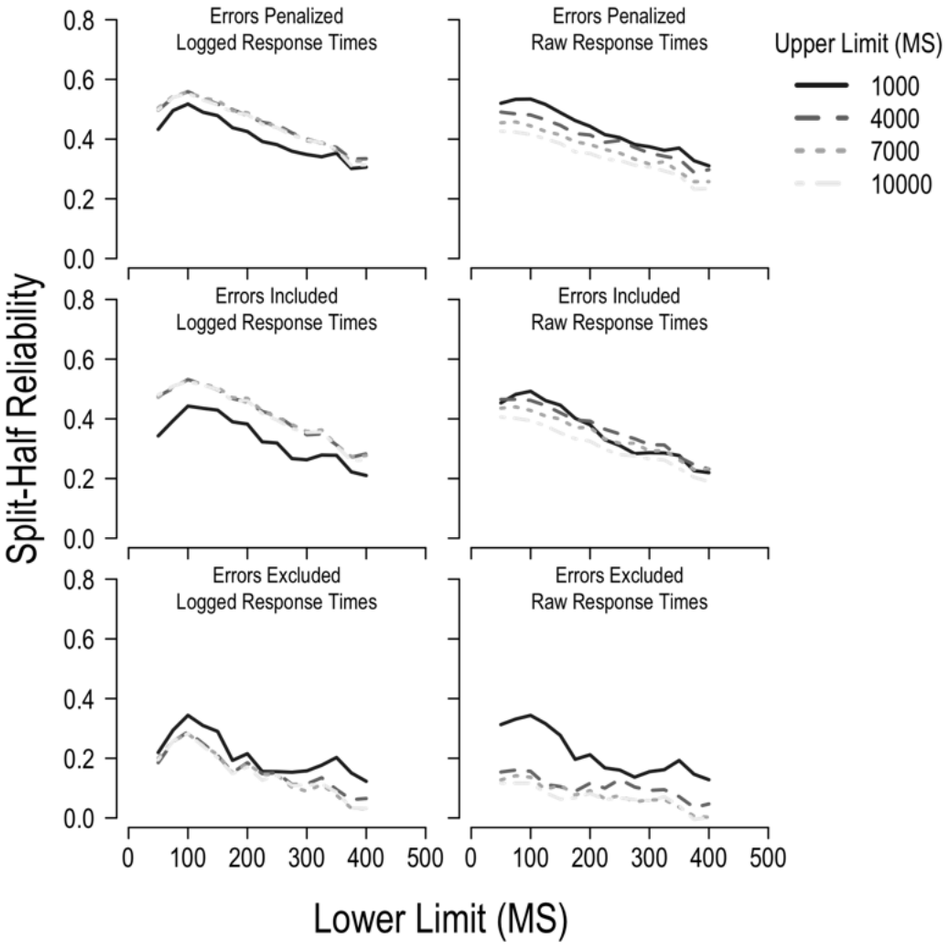

Figure S5. The average split-half reliabilities of Target D Scores using ST-IAT data computed using various combinations of algorithm parameters. Note: in this figure we display only figures for Target D Scores using a single standard deviation and ignoring the block 1/2 distinction.

Convergent Validity

To test the convergent validity of the ST-IAT Target D Scores, we computed Target D Scores for each Study and each set of parameters from the full available ST-IAT data. Based on our evidence that implicit evaluations of our targets were related to their gender, race, and social class, we then fit linear models predicting each computed set of Target D Scores from explicit ratings of targets: targets' race (three regressors: mean classification as Asian, mean classification as Black, and mean classification as White), targets' SES (mean ratings on the subjective SES ladder) and gender (mean classification as female). For each model, the r^2 was saved. Finally, for each combination of parameters, we computed an overall average r^2 across the six different sets of Target D Scores. The square roots of the average r^2 estimates of each parameter combination are depicted in Figure S6. The highest average convergent validity (r = 0.66) was produced by (a) penalizing error trials; (b) using logged response times; (c) setting a

minimum response time value of 150ms; (d) setting a maximum response time value of 4000ms; (e) using a target-specific *SD* for all targets; (f) ignoring the block 1/2 distinction.

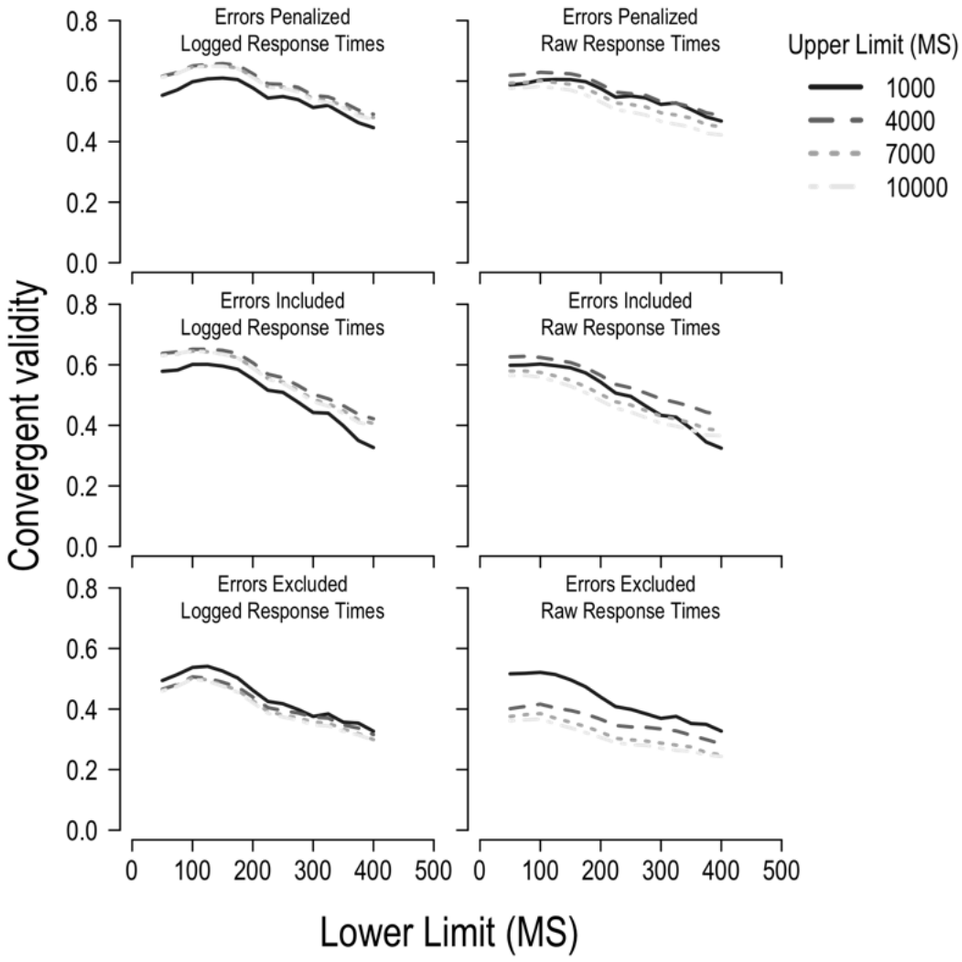

Figure S6. The average convergent validity estimates (correlations between Target D Scores and predictions of models using targets' perceived race, gender, and social class as predictors) computed using various combinations of algorithm parameters for ST-IAT Target D Scores. Note: in this figure we display only figures for Target D Scores using a single standard deviation and ignoring the block 1/2 distinction.

Selecting an Algorithm

To select the scoring algorithm producing the greatest combined internal reliability and convergent validity, we summed the average r_{xx} and r^2 values of each parameter combination, and selected the parameter combination producing the greatest $r_{xx} + r^2$ sum. Unsurprisingly, the greatest sum ($r_{xx} + r^2 = 0.81$) was achieved by the same algorithm that produced the greatest internal reliability: (a) penalizing error trials; (b) using logged response times; (c) setting a minimum response time value of 100ms; (d) setting a maximum response time value of 4000ms; (e) using a single *SD* for all targets; (f) ignoring the block 1/2 distinction. This is the algorithm we rely upon for the ST-IAT Target D Scores reported in our manuscript.

Evaluative Priming Task

To ascertain the optimal scoring algorithm for Target D Scores produced via our Evaluative Priming Task (EPT) data, we used the same procedure as used for ST-IAT data, with the exceptions that (a) we did not split data up by Study, because we only used the EPT task in Study 4, and (b) it was not an option to ignore/attend to the block 1/2 distinction, as EPT tasks do not use discrete blocks in the same way as ST-IATs or IATs (see our manuscript for a detailed description of the EPT method).

Internal Reliability

The average split-half reliability estimates of parameter combinations are depicted in Figure S7. The algorithm producing the greatest internal reliability for the EPT data (r_{xx} = 0.28) involved (a) excluding error trials; (b) using logged response times; (c) setting a minimum response time value of 50ms; (d) setting a maximum response time value of 1000ms; (e) using a single *SD* for all targets.

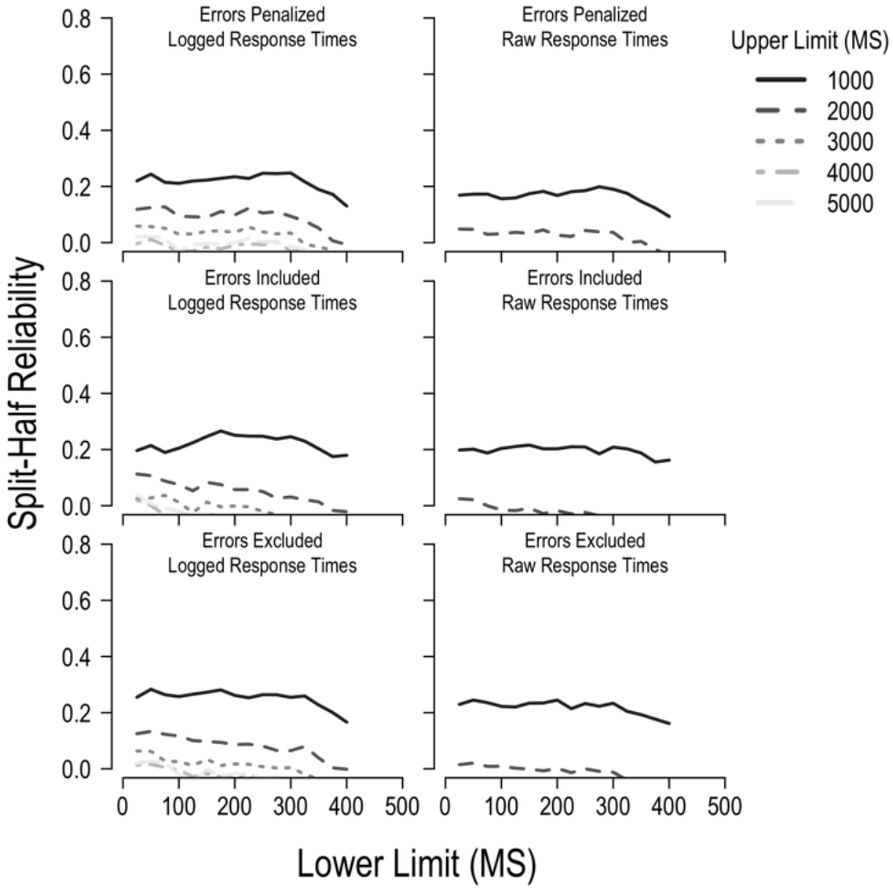

Figure S7. The average split-half reliabilities of Target D Scores for the Study 4 EPT data computed using various combinations of algorithm parameters. Note: in this figure we display only figures for Target D Scores using a single standard deviation.

Convergent Validity

The square roots of the average r^2 estimates of each parameter combination are depicted in Figure S8. The algorithm producing the greatest convergent validity for the EPT data (r = 0.39) involved (a) penalizing error trials; (b) using logged response times; (c) setting a minimum response time value of 25ms; (d) setting a maximum response time value of 1000ms; (e) using a single *SD* for all targets.

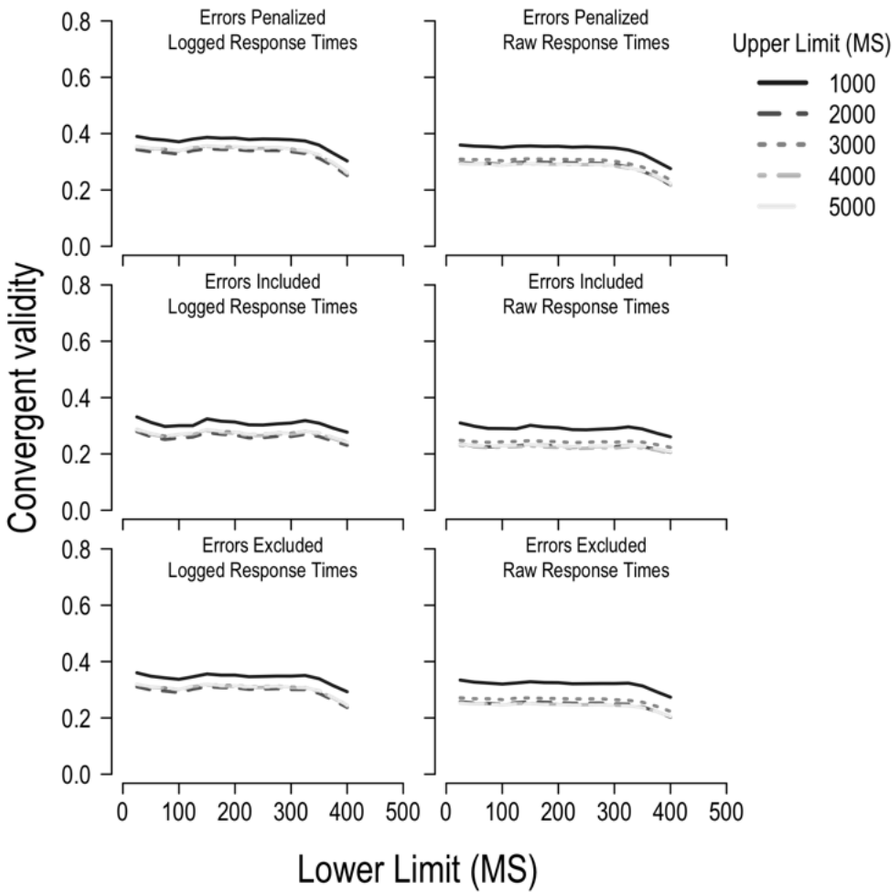

Figure S8. The average convergent validity estimates (correlations between Target D Scores and predictions of models using targets' perceived race, gender, and social class as predictors) computed using various combinations of algorithm parameters for EPT Target D Scores. Note: in this figure we display only figures for Target D Scores using a single standard deviation.

Selecting an Algorithm

The parameter combination producing the greatest r_{xx} + r^2 sum (r_{xx} + r^2 = 0.40) was achieved by the same algorithm that produced the greatest internal reliability: (a) excluding error trials; (b) using logged response times; (c) setting a minimum response time value of 175ms; (d) setting a maximum response time value of 1000ms; (e) using a single *SD* for all targets. This is the algorithm we rely upon for the EPT Target D Scores reported in our manuscript.

Appendix L: Power Analyses

Sample sizes for each study were chosen to maximize power within pragmatic constraints imposed by time and available human resources. However, given growing and justified interest in issues of statistical power within the psychological sciences (e.g., Fraley & Vazire, 2014), we performed power sensitivity analyses to assess the power of each study's sample size to detect effects of various sizes.

To achieve this, we used an approach incorporating a mixture of boot-strapping and simulation. To understand this approach, consider the case of a simple *post hoc* power analysis performed via boot-strapping. To carry this out, a researcher who has collected *N* cases and observed an effect of size θ needs only to repeatedly re-sample *N* cases with replacement from their data, and re-run their test of the effect in each of the boot-strapped samples. By recording the proportion of the boot-strapped samples in which the effect is statistically significant, the researcher can thereby produce an estimate of the power of their sample size *N* to detect the effect of interest at its observed effect size θ (Efron & Tibshirani, 1993). For example, if they find that the effect is significant in 80% of the boot-strapped samples, it would suggest 80% power at sample size *N* to detect the observed effect.

Alone, *post hoc* estimates yield little new information, because they are more or less re-statements of *p* values. However, if researchers alter parameters, they can use this procedure to yield estimates of power at different sample sizes or different effect sizes, which can provide valuable new information. For example, if the aforementioned researcher wished to estimate power at a different *N* to detect their effect θ, they would simply need to take smaller or larger boot-strapped samples from the observed data.

Altering effect sizes is also relatively easy. This is where the "simulation" part of the process comes into play, though we are here using the term 'simulation' relatively loosely. In fact, researchers can change the size of effects of interest within our data—and thus, within the 'populations' they draw bootstrapped samples from—via relatively minor adjustments of outcome variables that leave their datasets virtually unchanged. For example, in our case, a key effect of interest in Study 1b was the effect of target race on participants' D scores. To systematically alter the size of this effect, we first computed the following:

$$\acute{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{\beta race}_{\boldsymbol{i}}$$

where *i* indexes participants, y_i is the observed D Score of participant *i*, $\acute{y}_i$ is the adjusted D Score of participant *i*, β is the observed effect of a target race dummy variable $race_i$ in a linear model predicting y_i. In a new model predicting the altered outcome score $\acute{y}_i$, the effect of $race_i$ becomes zero. In a second step, we can add effects of $race_i$ of different sizes back to the dataset by choosing new values of the β term (δ), and creating a newly altered version of the outcome, like so:

$$\ddot{\boldsymbol{y}}_{\boldsymbol{i}} = \acute{\boldsymbol{y}}_{\boldsymbol{i}} + \boldsymbol{\delta race}_{\boldsymbol{i}}$$

where $\ddot{y}_i$ is a newly altered version of the outcome, and re-running the original model predicting $\ddot{y}_i$ will result in the estimated slope of $race_i$ being equal to δ. At this point, the dataset has still been changed very little; all that has changed is the size of the estimated effect of $race_i$, and with it, its effect size. At this point, we can estimate this effect size in the data, and estimate power to detect that effect at our *N* using the bootstrapping approach described above.

Studies 1a and 1b

Using this approach, we estimated the power sensitivity of our observed sample sizes to detect effects of various sizes in Studies 1a and Study 1b. For each separate analysis (the valence ST-IATs in Study 1a, the valence and wealth ST-IATs in Study 1b), we estimated the power sensitivity of our sample sizes to detect both a main and an interaction effect. Results (Figure S9) suggested that we had 80% power to detect effects of approximately η^2 = 0.005 in Study 1a, and η^2 = 0.015 in Study 1b.

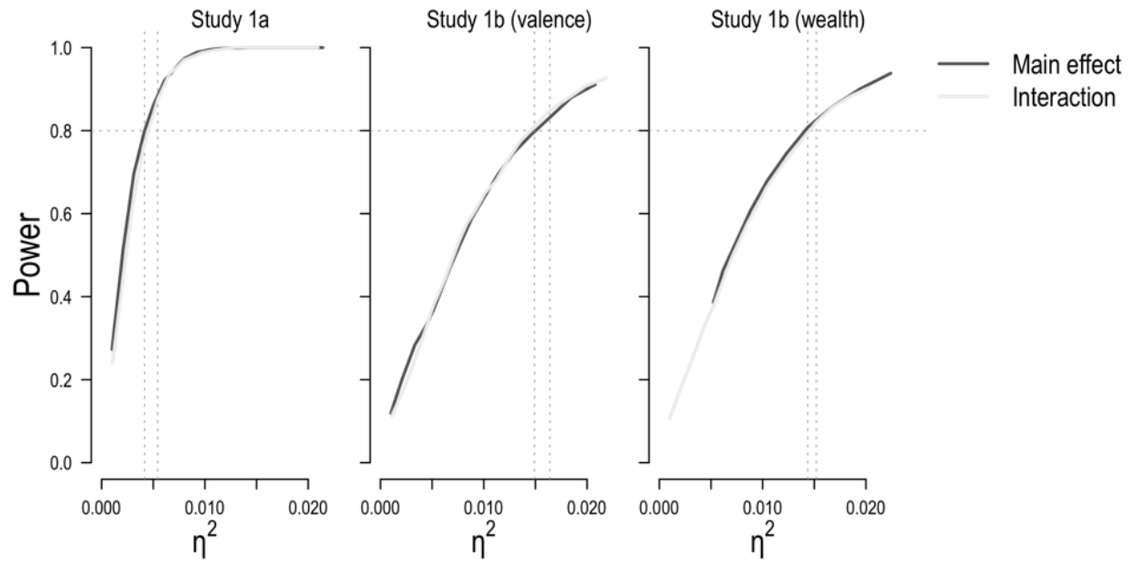

Figure S9. Power sensitivity curves for power to detect different-sized effects with our achieved sample sizes in Study 1a (N = 298, within-subjects design) and Study 1b (N = 533, between-subjects design).

Study 2

For Study 2's target-level multiple regression analysis, we estimated power sensitivity separately for a main effects of continuous MDS dimensions, and interactions between dimensions. Results suggested that based on our N of 54 unique targets we had 80% power to detect main effects of approximately η^2 = 0.1 and interaction effects of approximately η^2 = 0.08 (see Figure S10). Although these are relatively large effect sizes to power for compared to most psychological research, it is important to remember that by aggregating to the target level, effect sizes are greatly increased compared to analyses such as those in Study 1 that are conducted at the participant level.

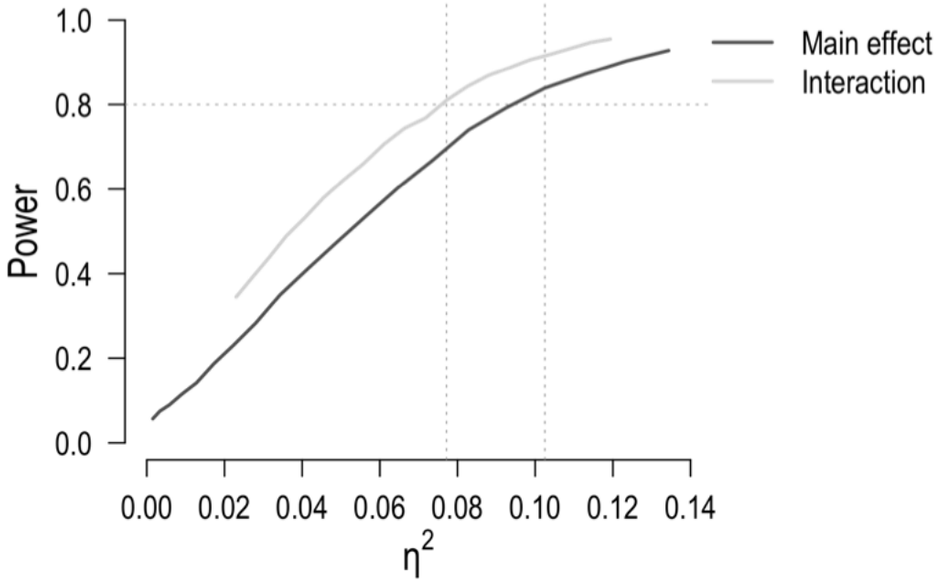

Figure S10. Power sensitivity curves for power to detect main effects of MDS dimensions and 2-way interaction effects between dimensions in Study 2 (N = 54 targets).

Studies 3a and 3b

For Studies 3a and 3b, due to we estimated separate power curves for main effects of each target-level factor (two race dummies, a gender dummy, and z-scored mean ratings of SES and age), and each two-way interaction between target-level factors. Due to the package lmerTest's (Kuznetsova, Brockhoff, Christensen, 2017) use of the Satterthwaite degrees of freedom method, degrees of freedom—and therefore statistical power—varied between effects. Figure S11 shows that with N = 144 targets made up of 24 unique faces and 24 unique bodies, Study 3a achieved

80% power to detect main effects between approximately $\Delta r^2 = 0.05$ (for the race dummies) and $\Delta r^2 = 0.09$ (for the gender dummy) and interaction effects between approximately $\Delta r^2 = 0.005$ (SES × Asian & SES × Black) and $\Delta r^2 = 0.035$ (Gender × Age). Study 3b achieved 80% power to detect main effects between approximately $\Delta r^2 = 0.04$ (the Asian race dummy) and $\Delta r^2 = 0.095$ (for the gender dummy) and interaction effects between approximately $\Delta r^2 = 0.005$ (SES × Asian & SES × Black) and $\Delta r^2 = 0.03$ (SES × Gender).

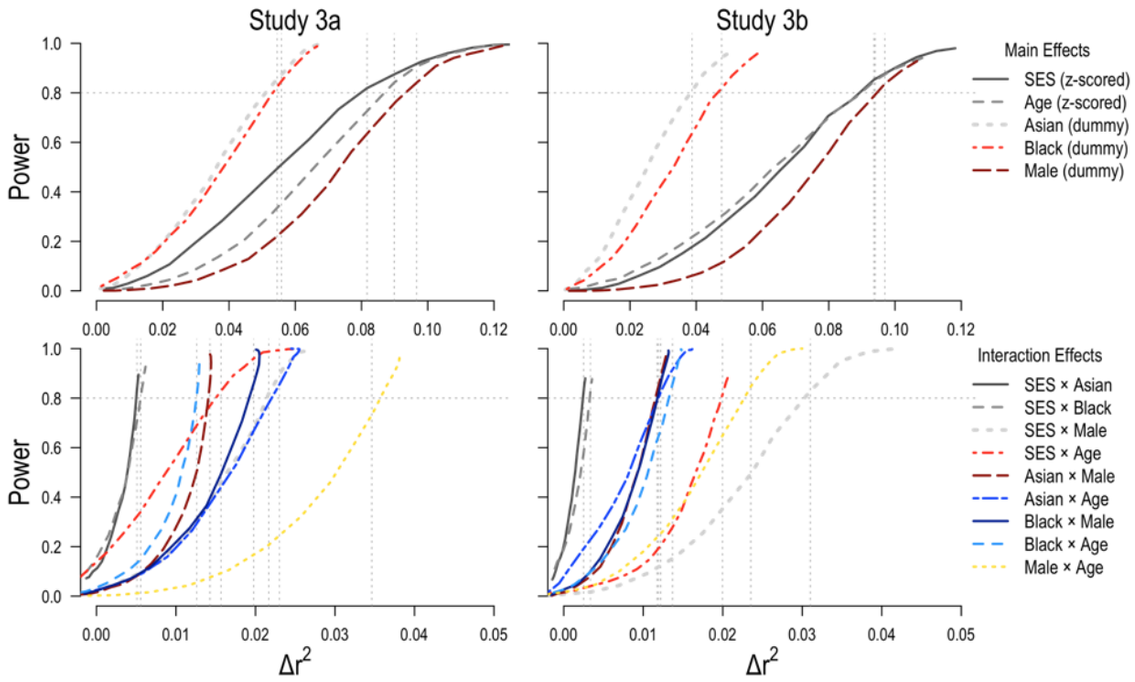

Figure S11. Power sensitivity curves for power to detect main effects of target-level factors (race, gender, social class, and age) and 2-way interaction effects between dimensions in Studies 3a and 3b (N = 144 targets made up of 24 unique faces and 24 unique bodies).

Study 4

For Study 4, we estimated separate power sensitivity curves for main effects of each target-level factor (two race dummies, a gender dummy, and z-scored mean ratings of SES and age), but just one power sensitivity curve for two-way interaction terms between target-level factors, as degrees of freedom were approximately equal between each interaction term. Figure S12 shows that with N = 288 targets made up of 24 unique faces and 24 unique bodies (each unique target had a full-body and upper-body Target D Score entered in models), Study 4 achieved 80% power to detect main effects between approximately $\Delta r^2 = 0.04$ (for the race dummies) and $\Delta r^2 = 0.07$ (for SES), and interaction effects of approximately $\Delta r^2 = 0.025$.

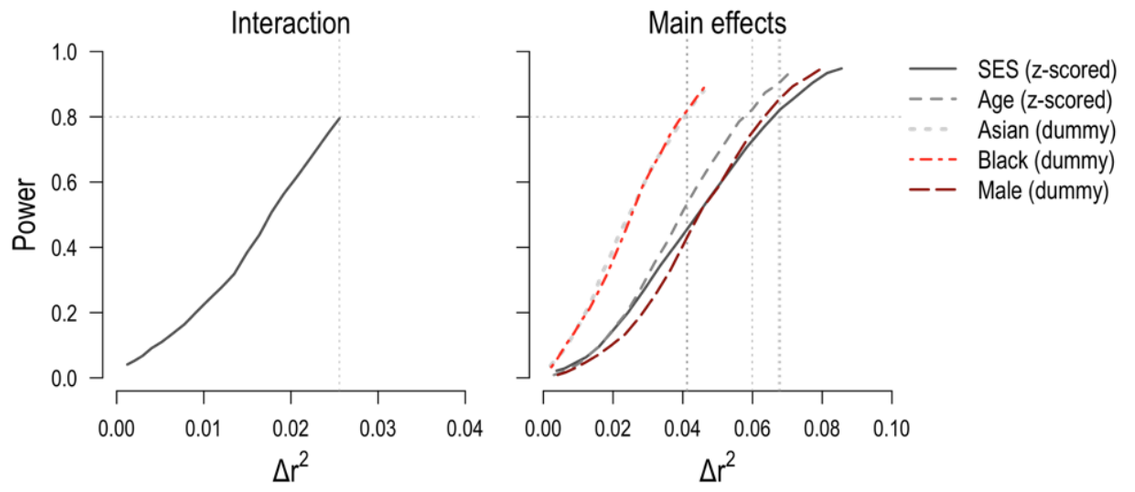

Figure S12. Power sensitivity curves for power to detect main effects of target-level factors (race, gender, social class, and age) and 2-way interaction effects between dimensions in Study 4 (N = 288 targets made up of 24 unique faces and 24 unique bodies).