AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Personality and Social Psychology:
Attitudes and Social Cognition

© 2023 American Psychological Association
ISSN: 0022-3514

https://doi.org/10.1037/pspa0000370

(When) Do Counterattitudinal Exemplars Shift Implicit Racial Evaluations? Replications and Extensions of Dasgupta and Greenwald (2001)

Benedek Kurdi[1], Alex Sanchez[2, 3], Nilanjana Dasgupta[4], and Mahzarin R. Banaji[2]
[1] Department of Psychology, University of Illinois Urbana–Champaign
[2] Department of Psychology, Harvard University
[3] Department of Psychology, Princeton University
[4] Department of Psychological and Brain Sciences, University of Massachusetts Amherst

Dasgupta and Greenwald (2001) demonstrated that exposure to positive Black exemplars (e.g., Colin Powell) and negative White exemplars (e.g., Jeffrey Dahmer) can reduce implicit pro-White/anti-Black evaluations, as measured by an Implicit Association Test. Here, we report seven preregistered online experiments conducted with volunteer U.S. participants ($N = 6{,}953$) that sought to replicate and probe the boundary conditions of this finding. Contrary to expectations, we found no shift in implicit racial evaluations in two close replication attempts (Experiments 1–2). Experiments 3–4 ruled out the possibility of insufficiently strong exemplar valence and subtyping as explanations for the failures to replicate. In Experiment 5, implicit racial evaluations did exhibit malleability in response to two different procedures relying on repeated evaluative pairings and evaluative statements, suggesting that they are capable of change. With insight from these studies, Experiments 6–7 were mounted with modifications to the Dasgupta and Greenwald (2001) procedure. Significant reductions in implicit pro-White/anti-Black evaluations were now observed when race, valence, and the contingency between the two were highlighted. In addition, across all experiments, the magnitude of shift in implicit racial evaluations was significantly predicted by participants' ability to recall the Black–positive and White–negative contingencies experienced during the exemplar exposure task. Together, these data suggest that exposure to counterattitudinal exemplars can shift implicit racial evaluations toward neutrality, but such malleability strongly depends on contingency awareness. We discuss implications for social cognitive theory, theoretically informed debiasing interventions, and different paths toward resolving initial replication failures.

Keywords: associative learning, attitudes, contingency awareness, implicit evaluations, replication

In a special section of the *Journal of Personality and Social Psychology* devoted to studies that examined change in implicit evaluations[1] and stereotypes, Dasgupta and Greenwald (2001) reported a result seen as surprising at the time: Exposure to

Benedek Kurdi https://orcid.org/0000-0001-5000-0584

Preregistrations, materials, data, and analysis scripts are available via the Open Science Framework (https://osf.io/rce3m/).

Benedek Kurdi is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) nonprofit organization and international collaborative of researchers who are interested in implicit social cognition.

Benedek Kurdi played a lead role in conceptualization, formal analysis, investigation, and writing–original draft. Alex Sanchez played a lead role in conceptualization, investigation, and writing–review and editing. Nilanjana Dasgupta played a lead role in conceptualization and writing–review and editing. Mahzarin R. Banaji played a lead role in conceptualization and writing–original draft.

Correspondence concerning this article should be addressed to Benedek Kurdi, Department of Psychology, University of Illinois Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820, United States. Email: kurdi@illinois.edu

[1] Throughout the article, we use the term "implicit racial evaluations" to refer to the construct measured by the race IAT, which is the main dependent measure used in the present studies. We do so to distinguish latent representations of evaluative knowledge (which we refer to as "attitudes") and observable behavioral manifestations of such evaluative knowledge ("evaluations"; see Cunningham et al., 2007) from each other. Of course, indirect measures (such as the IAT) can be affected by controlled processes of memory retrieval and direct measures (such as feeling thermometers or Likert-type scales) can be affected by automatic processes of memory retrieval (Jacoby, 1991). However, on balance, the IAT involves the relatively automatic (unintentional) and self-report measures involve the relatively controlled (intentional) retrieval of evaluative information (De Houwer et al., 2009). As such, to distinguish these two modes of information retrieval from each other, we refer to the construct measured by the former as "implicit evaluations" and the latter as "explicit evaluations" without assuming the existence of multiple memory systems, attitude representations in long-term memory, or a perfect mapping from measures (direct vs. indirect) onto attitude representations (explicit vs. implicit) or processes of retrieval (controlled vs. automatic).

counterattitudinal (positive Black and negative White) exemplars[2] was sufficient to shift[3] implicit pro-White/anti-Black[4] racial evaluations toward neutrality to a considerable degree. Along with two other reports published in the same section (Blair et al., 2001; Lowery et al., 2001), the article by Dasgupta and Greenwald (2001) set into motion what was soon to become a fundamental shift in our understanding of the nature of racial attitudes and, specifically, implicit racial evaluations (see Blair, 2002, for an early review). In the early days of implicit social cognition research, it had been widely believed that, given their automatic nature (Bargh et al., 1996; Devine, 1989; Fazio et al., 1986; Greenwald & Banaji, 1995), once acquired, implicit racial evaluations would be recalcitrant in the face of any attempt at even temporary modulation (Banaji, 2004). Contrary to this view, these early results suggested that malleability in implicit racial evaluations, and therefore perhaps even long-term change in the underlying attitudes, was possible.

In Experiment 1 of Dasgupta and Greenwald (2001), which is our focus in the present project, 33 White American and Asian American undergraduates from the University of Washington were assigned either to an experimental condition designed to shift implicit racial evaluations toward neutrality ($n = 18$) or a procedurally matched control condition ($n = 15$). A third condition with 15 participants relied on a manipulation designed to shift implicit racial evaluations toward an even stronger pro-White/anti-Black stance. Given the exploratory nature of this condition, along with the fact that it produced no shifts in implicit evaluations and is not of central theoretical interest, we have omitted it from the present experiments.

In the critical experimental condition of Dasgupta and Greenwald (2001), participants were told that they would complete a general knowledge test probing their familiarity with famous and infamous Americans. Although this feature of the paradigm was not made explicit to participants, all Black American exemplars presented during the ostensible general knowledge test were positive, including civil rights leaders such as Martin Luther King, Jr., political figures such as Colin Powell, entertainers such as Eddie Murphy, and athletes such as Michael Jordan. By contrast, all White American exemplars were negative, including serial killers, terrorists, gangsters, and other criminals such as Al Capone, Ted Kaczynski, Jeffrey Dahmer, and Timothy McVeigh.

On each trial, participants were presented with an image of a target individual, along with two descriptions, which both had the same valence but only one of which was accurate. For example, Al Capone would appear along with the correct description "American gangster who terrorized Chicago in the 1920s" and the foil "leader of an anti-government militia." Participants were asked to choose the accurate description and received feedback on the accuracy of their response. Subsequently, participants categorized the name (but not the image) of each exemplar by race to ensure that they were cognizant of the racial group membership of the famous and infamous individuals.

The control condition was structurally matched to the experimental condition, except that instead of individuals from the two social groups (Black Americans and White Americans), participants were exposed to pictures of flowers and insects, accompanied by descriptions that were always positive for flowers and negative for insects. Participants' task was to select the accurate description of each flower and insect and subsequently sort the exemplars by taxonomic categories. Subsequently, participants' racial attitudes were indexed using direct and indirect tests. Specifically, a version of the Implicit Association Test (IAT; Greenwald et al., 1998) was used to measure implicit racial evaluations in both conditions, with feeling thermometer and semantic differential items administered to measure explicit racial evaluations. A significant and large difference (corresponding to Cohen's $d = 0.94$) was obtained on the theoretically critical test of implicit racial evaluations. Explicit evaluations did not significantly differ from each other across the two conditions.

Since the publication of these influential early findings by Dasgupta and Greenwald (2001), a firm theoretical understanding has emerged that implicit evaluations, including implicit racial evaluations, can exhibit sizable temporary shifts toward neutrality in response to a wide range of interventions (see Karpinski & Hilton, 2001; Sinclair et al., 2005; Turner & Crisp, 2010; Wittenbrink et al., 2001, for early examples; see Ferguson et al., in press; Kurdi & Charlesworth, 2023; Morehouse & Banaji, in press, for recent reviews). For example, Lai et al. (2014) organized an intervention tournament in which all interested investigators were invited to submit the procedure that they thought would shift implicit racial evaluations among White Americans to the largest possible extent. Although not all submitted interventions had impact, implicit racial evaluations were found to shift in response to a broad range of experimental manipulations, including a vivid first-person narrative involving a positive Black and negative White protagonist (Marini et al., 2012), redefining group boundaries toward a shared identity (Dovidio et al., 2009), implementation intentions (Mendoza et al., 2010), and evaluative conditioning using Black–positive and White–negative stimulus pairings (Olson & Fazio, 2006).

If the theoretical consensus about the malleability of implicit racial evaluations is so robust today, why did we consider

[2] Dasgupta and Greenwald (2001) refer to the corresponding exemplars as "counterstereotypic" rather than "counterattitudinal." We opted for the latter terminology because most positive Black exemplars used in the present experiments represent areas that are often stereotypically associated with Black excellence in the United States, including entertainment, sports, and spiritual leadership. As such, we believe that it is an open empirical question whether exemplars that are both counterstereotypic and counterattitudinal would produce effects that are different from the ones observed here. In addition, in a departure from the original results, participants in the present studies reported relatively more positive explicit evaluations of Black Americans than of White Americans (although evaluations of both groups were positive in an absolute sense), which may make the use of the term "counterattitudinal" to refer to positive Black and negative White exemplars questionable. However, given that the present studies focus on implicit, rather than explicit, evaluations and participants exhibited clear pro-White/anti-Black implicit evaluations in the control conditions of all studies, we believe that such use is warranted.

[3] Throughout the article, we use the term "shift" or "(temporary) malleability" to refer to differences in IAT performance across the control and experimental conditions, for two reasons. First, in all present experiments, responding on the IAT was measured immediately following the intervention and, as such, it is unclear whether the observed effects would persist over longer periods of time. Second, given that both the exemplars used during learning and the categories used on the IAT were highly familiar to participants, any intervention effects may have been mediated by either genuine attitude change (i.e., an enduring modification to evaluative representations) or the selective retrieval of already represented evaluative information. We return to these possibilities and their theoretical and practical implications in more detail in the General Discussion.

[4] Throughout the article, we use the terms "pro-White/anti-Black" or "pro-Black/anti-White" to refer to a relative evaluation in favor of White (Black) over Black (White) Americans. Given that the IAT is an inherently relative measure, it does not allow for inferences about evaluative tendencies in the absolute sense.

conducting a replication of the Dasgupta and Greenwald (2001) findings in the early 2020s? First, the Dasgupta and Greenwald (2001) report has been highly influential, with over 1,800 citations as of September 2023 according to Google Scholar. As such, whether the shift in implicit racial evaluations documented by these authors replicates after more than 20 years is of inherent theoretical and practical interest.

Second, an independent replication today is timely given that a previous replication attempt published by Joy-Gaba and Nosek (2010) over a decade ago replicated the Dasgupta and Greenwald (2001) result with large samples but with considerably smaller effect sizes (Cohen's *d*s = 0.17 and 0.14). Given the large effect observed in the original experiment, this substantial decrease from 2001 to 2010 raises questions about the true magnitude of the underlying effect.

Third, the Dasgupta and Greenwald (2001) paradigm is unique in its reliance on exposure to counterattitudinal exemplars in shifting implicit evaluations and has not been emulated often over the past two decades. In fact, several interventions that have been shown to shift implicit racial evaluations toward neutrality, including many of the ones identified as effective by Lai et al. (2014), have limited generalizability beyond a laboratory context due to the effortfulness and artificiality of the learning task.

For example, in the most effective intervention in the Lai et al. (2014) article, participants were asked to imagine a vivid counter-attitudinal scenario in which they are viciously attacked by a White man and then heroically saved by a Black man. The formation of implementation intentions ("I will think 'good' when seeing Black faces and 'bad' when seeing White faces") and practicing counter-attitudinal (Black–good/White–bad) pairings hundreds of times is similarly effortful and lacks an obvious real-world analog. Against this backdrop, the appeal of the Dasgupta and Greenwald (2001) procedure lies in its simplicity, along with the fact that incidental exposure to counterattitudinal exemplars without any effortful processing involving race, valence, or the relationship between the two, has been thought to be sufficient to produce the effect. As such, the procedure is often seen as a laboratory model of simply encountering counter-attitudinal individuals in one's daily life.

Fourth, although anti-Black racism and myriad forms of anti-Black discrimination continue to persist in U.S. society (Banaji et al., 2021; Kraus et al., 2019; Skinner-Dorkenoo et al., 2023), the societal context in general and race relations in particular have shifted considerably since the Dasgupta and Greenwald (2001) experiments were conducted in the late 1990s. Among other things, media representation of Black Americans has become both more voluminous (Shor & van de Rijt, 2023) and less uniformly negative (Leonard & Robbins, 2021); societal interest in and awareness of anti-Black racism and discrimination has increased as a result of the Black Lives Matter movement (Barrie, 2020; Reny & Newman, 2021); and both direct and indirect tests have provided evidence for a decrease in anti-Black attitudes (Charlesworth & Banaji, 2019, 2022).

These changes in the broader societal context could be expected to modulate the effectiveness of the Dasgupta and Greenwald (2001) procedure in reducing implicit pro-White/anti-Black evaluations today in multiple ways. On the one hand, it is conceivable that a less hostile societal climate, including more positive Black media representations, higher levels of awareness of anti-Black discrimination, and more positive racial attitudes overall, could facilitate shifts in implicit racial evaluations toward neutrality in response to counter-attitudinal exemplars, for example, by making participants more likely to notice or more motivated to engage with the information conveyed through the experimental procedure (Klayman, 1995). Indeed, at an organizational level, debiasing interventions are more likely to succeed in hospitable social environments relative to those in which group-based discrimination goes unrecognized as a problem in the first place (Kalev et al., 2006). If this is the case, then the Dasgupta and Greenwald (2001) effect should be even larger today than it was in the late 1990s.

On the other hand, it is also conceivable that changes in the broader societal context may reduce the effectiveness of the Dasgupta and Greenwald (2001) intervention relative to the conditions under which the original research was conducted. A general principle of learning is that of expectancy violation, that is, that more surprising information should lead to more learning than less surprising information (Rescorla & Wagner, 1972)—a finding that has also been repeatedly obtained in the context of social learning (Heffner et al., 2021; Solié et al., 2022). Accordingly, more frequent exposure to positive information about Black Americans in today's social environment may make the positive Black exemplars presented in the study less noticeable or notable, thereby blunting the effects of the experimental manipulation. Indeed, consistent with the idea of expectancy violation, greater prior contact with outgroups has been found to reduce the effectiveness of counterattitudinal exemplar exposure in modulating implicit evaluations (see Dasgupta & Asgari, 2004, in the context of gender, and Dasgupta & Rivera, 2008, in the context of sexual orientation).

Moreover, research has now demonstrated that procedures explicitly drawing attention to and creating awareness of social groups, and often relying on verbal statements referring to the entire category rather than exposure to particular exemplars, can produce stronger effects on implicit evaluations than procedures relying on rote learning of co-occurrences (Kurdi & Banaji, 2017). As such, the unexpected nature of counterattitudinal exemplars may be more easily noticed more when the task draws attention to an exemplar's category membership and prototypical knowledge associated with that category. For all these reasons, we considered it imperative to conduct a replication of the influential experiment by Dasgupta and Greenwald (2001), the data of which were collected close to 25 years ago.

The Present Project

Driven by these considerations, in Experiments 1–2, we sought to replicate the reduction in implicit pro-White/anti-Black evaluations observed by Dasgupta and Greenwald (2001) in procedures that stayed as close as possible to the original paradigm. To our surprise, exposure to counterattitudinal exemplars did not produce any appreciable shift in implicit evaluations in the two initial experiments. In Experiments 3–4, we tested and eliminated two potential explanations for the lack of replication: insufficiently strong exemplar valence and subtyping of famous individuals.

In Experiment 5, we used two manipulations borrowed from Kurdi and Banaji (2017), known to produce sizable shifts in implicit age and nationality evaluations, to probe the malleability of implicit racial evaluations using a different set of learning tasks. We did observe shifts in implicit evaluations in both conditions of this experiment relative to baseline, which led us to conclude that the lack of shifts observed in Experiments 1–2 must have been a function of some aspect(s) of the Dasgupta and Greenwald (2001) paradigm (either in isolation or in combination with societal factors) rather than the recalcitrance of implicit racial evaluations more broadly.

In Experiments 6–7, we hypothesized that the lack of shifts in implicit racial evaluations observed in Experiments 1–4 may have been due to the incidental learning conditions created by the Dasgupta and Greenwald (2001) procedure. Specifically, the original "general knowledge test" framing and the subsequent exposure task did not direct participants' attention to (and may even have directed participants' attention away from) the racial group membership of the exemplars, the valence of the biographical descriptions, and the relationship between race and valence. As such, in the two final experiments, we sought to increase participants' awareness of race–valence contingencies using three procedures that varied in the explicitness of instructions provided prior to exemplar exposure and found a significant reduction of implicit pro-White/anti-Black evaluations in both experiments.

Experiment 1: Close Replication I

Experiment 1 was conducted as a close replication of Experiment 1 from Dasgupta and Greenwald (2001), which had demonstrated that exposure to counterattitudinal (positive Black and negative White) race exemplars can shift implicit pro-White/anti-Black evaluations toward neutrality to a sizable degree (Cohen's $d = 0.94$).

The experiment consisted of a learning phase and a test phase. In the learning phase, participants were randomly assigned to an experimental or control condition. Participants in the experimental condition completed an exposure task described to them as a general knowledge test. On each trial of the exposure task, they were shown the picture of a positive Black or negative White exemplar and were asked to select the correct description of the exemplar from two options. For example, Jeffrey Dahmer would appear along with the correct description "serial killer who cannibalized his victims" and the valence-matched foil "bombed the World Trade Center in NYC." Subsequently, participants were provided the names of the exemplars and asked to categorize each by race. Participants in the control condition completed a structurally matched task involving stimuli irrelevant to racial attitudes (specifically, insects and flowers). Then, in the test phase of both conditions, implicit and explicit evaluations were measured in the same fixed order, followed by some newly added exploratory items probing memory of the contingency between race and valence categories in the experimental condition.

Given that the goal of Experiment 1 was close replication, we followed the procedure of Dasgupta and Greenwald (2001) as faithfully as possible. Any deviations from the original paradigm of which we are aware are explicitly mentioned below. If no deviation is noted, it should be assumed that the procedure was in line with that of the original experiment.

Method

Open Science Practices

We report all measures, manipulations, and exclusions for this and all remaining experiments. The hypothesis, design, sample size, and participant exclusions were preregistered. All preregistrations (https://osf.io/rce3m/), raw data files (https://osf.io/ckdzt/), analysis scripts (https://osf.io/dy56j/), and materials (https://osf.io/spzx9/) used in this and all remaining experiments are available for download from the Open Science Framework (OSF).

In Experiments 1–4, we recruited participants irrespective of country of origin. In Experiments 5–7, only participants from the United States were recruited. To match the original Dasgupta and Greenwald's (2001) study protocol and because knowledge of famous and infamous U.S. exemplars is necessary to complete the task, our main analyses focus on U.S. participants; however, data for non-U.S. participants can be downloaded from OSF. Analyses including non-U.S. participants are also available in the analysis script and yield the same substantive conclusions as those reported in the article. All U.S. participants were included in the analyses irrespective of their racial group membership, with effects of participant race explicitly tested in moderation analyses.

Participants and Design

Participants were 1,533 adult volunteers recruited via the Project Implicit educational website (https://implicit.harvard.edu/). In line with standard recommendations (Greenwald et al., 2003) and as preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT (Greenwald et al., 1998), which served as the main dependent measure ($n = 30$), or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 26$). These exclusions left 1,477 participants in the sample, of which we focus on the final sample of 1,108 U.S. participants below.

In the final sample, 762 participants were female, 310 participants male, and 28 participants of other genders. Mean participant age was 38 years ($SD = 17$ years). 734 participants identified as White, 115 participants as Black, 100 participants as Hispanic, 77 participants as Asian, 71 participants as multiracial, five participants as Middle Eastern, two participants as Pacific Islander, and one participant as Native American.

The experiment consisted of a learning phase and a test phase. In the learning phase, participants were randomly assigned to an experimental condition ($n = 593$) involving exposure to positive Black and negative White exemplars or a procedurally matched control condition ($n = 515$) involving exposure to exemplars of flowers and insects. In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory items probing memory of the contingency between racial and valence categories.

Materials

Images and Descriptions of Flowers and Insects. Ten grayscale images of flowers and 10 grayscale images of insects, along with the corresponding accurate and inaccurate descriptions, were retained for use from Dasgupta and Greenwald (2001).

Images and Descriptions of Positive Black and Negative White Exemplars. Eight grayscale images of positive Black exemplars and nine grayscale images of negative White exemplars, along with the corresponding accurate and inaccurate descriptions, were retained for use from Dasgupta and Greenwald (2001). All original exemplars (and, therefore, all exemplars used in the present replication) were male. Two positive Black exemplars and one negative White exemplar from the original set, along with the corresponding accurate and inaccurate descriptions, were replaced because a pretest had revealed substantial shifts in their societal evaluations.

The pretest was conducted on the same online platform as the remaining experiments, in a sample of 244 U.S. participants. As part of the pretest, participants were presented with the 10 positive Black and the 10 negative White exemplars used in Dasgupta and Greenwald (2001). In addition, 10 new positive Black and 10 new negative White exemplars were also included, which we reasoned could serve as appropriate replacements if necessary, given cultural shifts in some targets' reputations (e.g., Bill Cosby served as a positive Black exemplar in 2001).

Participants were presented with all 40 exemplars in individually randomized order. Participants first indicated whether they were familiar with the target or not. If the target was not familiar, the program proceeded to the next target. If the target was familiar, participants were asked to indicate, using a 201-point sliding scale, how coldly or warmly they felt toward the target. Based on these responses, we calculated a weighted liking index (WLI) for each of the 40 exemplars by multiplying the mean feeling thermometer score by the proportion of participants who indicated that they were familiar with the individual ($M = -1.00$, $SD = 47.47$). The theoretical range of this measure is −100 (universally known and maximally disliked individual) to +100 (universally known and maximally liked individual).

Scores on the WLI led to the replacement of three exemplars used by Dasgupta and Greenwald (2001): Among Black exemplars, Tiger Woods (WLI = 7.81) and Bill Cosby (WLI = −48.00), whose reputations had suffered considerably since the original experiments were conducted, were replaced with Morgan Freeman (WLI = 68.04) and Stevie Wonder (WLI = 56.48), respectively. Among White exemplars, Howard Stern's (WLI = −17.75) reputation had improved considerably since the original Dasgupta and Greenwald (2001) experiments were conducted. As such, he was replaced with Jeffrey Epstein (WLI = −70.29). In the final set, the mean WLI of Black exemplars was 47.94 ($SD = 20.31$) and the mean WLI of White exemplars was −54.29 ($SD = 22.47$), supplying a useable set of positive Black and negative White exemplars.

Procedure and Measures

The experiment consisted of a learning phase and a test phase. For the purposes of the learning phase, participants were randomly assigned to an experimental or a control condition. All participants completed the same test phase. With some specific exceptions listed below, the experiment followed all elements of the procedure of Experiment 1 from Dasgupta and Greenwald (2001).

Learning Phase.

Experimental Condition. The learning phase in the experimental condition consisted of an exposure task (described to participants as a knowledge test) and a categorization task. The purpose of the exposure task was to have participants engage with positive Black and negative White exemplars. The purpose of the categorization task was to remind participants of each exemplar's racial group membership.

Leading up to the exposure phase, participants were informed that they would be tested on their knowledge of famous American individuals. They were told that on each trial, they would see an individual and would be asked to select the correct description applicable to that individual by using the "E" or "I" key on their keyboard. Participants were further told that incorrect responses would be indicated by a red X on the screen; in order to proceed, participants were asked to enter the correct response.

Following these instructions, participants completed the exposure phase, which included a total of 40 trials. Each trial consisted of the presentation of a grayscale photograph of a target, including his name, in the center of the screen, with one description displayed in the bottom left and the other description in the bottom right corner. Each target, along with an accurate and inaccurate description, was presented once over the course of the first block of 20 trials and then once again over the course of the second block of 20 trials.

The order of targets in the first and second blocks was independently randomized for each participant. The left or right positioning of the correct and incorrect descriptions was also randomized on each trial. Importantly, correct and incorrect descriptions were matched on valence to avoid inadvertent association of the unintended (opposite) valence with any exemplar in the learning phase. For example, the correct description for Colin Powell (a positive Black exemplar) was "former Chairman, Joint Chiefs of Staff for the U.S. Department of Defense" and the incorrect description was "U.S. Ambassador to the United Nations."

Once they had completed the exposure task, participants proceeded to the categorization task. They were informed that the upcoming task would test whether they remembered the racial group membership of the famous individuals from the previous task. They were told that, on each trial, the name of an individual would appear on the screen and that they would be asked to use the "E" and "I" keys to indicate whether the individual was White or Black. Error correction was the same as on the exposure task. Similar to the exposure task, the categorization task consisted of two blocks of 20 trials over the course of which each target appeared once, in individually randomized order.

Control Condition. The control condition was procedurally matched to the experimental condition. However, instead of the Black and White exemplars used in the experimental condition, participants were asked to select the correct description for flowers and insects in the exposure task and categorized targets as flowers or insects in the categorization task.

Test Phase. The test phase consisted of measurement of implicit racial evaluations using the IAT, measurement of explicit racial evaluations using self-report items, and exploratory measures probing memory of the contingency between racial and valence categories (in the experimental condition only). Given our theoretical focus on implicit evaluations, the IAT was always administered first, and self-report items were always administered second.

Implicit Racial Evaluations. Implicit racial evaluations were measured using a version of the IAT (Greenwald et al., 1998).

Category labels were *White* and *Black*. Category stimuli for the former category included *JOSH*, *BRANDON*, *JUSTIN*, *IAN*, and *ANDREW*, and for the latter category included *LAMAR*, *JAMAL*, *LIONEL*, *TORRANCE*, and *MALIK*. At the time the original experiment was conducted, it was technically difficult to present image stimuli in computerized experiments. As such, although names were regarded as a less appropriate way to represent race (stereotypical Black names represent a subtype of Black Americans, not the group as a whole), Dasgupta and Greenwald (2001) used names as category stimuli. Therefore, we did the same in closely replicating the experiment.

Attribute labels were *Pleasant* and *Unpleasant*. Attribute stimuli for the former attribute included *rainbow*, *gift*, *joy*, *paradise*, and *laughter*, and for the latter attribute included *sickness*, *cancer*, *vomit*, *war*, and *poison*. For ease of discriminability, category stimuli were presented in green color and all-caps font, and attribute stimuli in blue color and all lowercase font.

The IAT consisted of five blocks: (a) category practice (White vs. Black; 20 trials); (b) attribute practice (pleasant vs. unpleasant; 20 trials); (c) first critical block (White/pleasant vs. Black/unpleasant or Black/pleasant vs. White/unpleasant; 40 trials); (d) reverse category practice (Black vs. White; 20 trials); and (e) reverse critical block (White/unpleasant vs. Black/pleasant or Black/unpleasant vs. White/pleasant, depending on the order in the first combined block; 40 trials). The placement of category and attribute labels to the left and right side of the screen and the order of the two critical blocks was randomized. Participants used the "E" and "I" keys for categorization and were required to correct inaccurate responses before proceeding.

IAT D scores were calculated using the improved scoring algorithm (Greenwald et al., 2003). Higher D scores indicate stronger pro-White/anti-Black evaluations based on the relative speed and accuracy of responding across the two critical blocks of the IAT.

Explicit Racial Evaluations. Explicit racial evaluations were measured using (a) a feeling thermometer item and (b) semantic differential items along the dimensions ugly–beautiful, bad–good, unpleasant–pleasant, dishonest–honest, and awful–nice. Participants used 201-point sliding scales to enter their responses. They first responded to the feeling thermometer items with White Americans and Black Americans as the targets (in randomized order). Afterward, they completed the semantic differential items, first for one racial group and then for the other racial group. The order of the two racial groups as well as the order of items within each racial group was randomized.

The explicit evaluation items with White Americans (Cronbach's $\alpha = .91$) and Black Americans ($\alpha = .91$) as the target categories were highly reliable and were therefore used to create an index of overall evaluation. The index for evaluations of Black Americans was then subtracted from the index for evaluations of White Americans to derive an explicit evaluation difference score paralleling the IAT D score, with higher scores indicating higher levels of pro-White/anti-Black evaluation.

Exploratory Measures of Contingency Memory. Participants in the experimental condition were administered four exploratory measures of contingency memory, in decreasing order of stringency (Moran et al., 2021). First, they were asked to indicate whether they noticed anything out of the ordinary with the famous American individuals presented during the initial task. Second, they were asked to report whether they noticed anything systematic about the famous individuals who were generally admired versus the famous individuals who were generally disliked. Third, they were asked to answer yes or no to the question of whether all the admired famous individuals in the learning task were of one racial group and all the disliked famous individuals were of another racial group. Finally, they were asked to choose which of the two target groups consisted of admired individuals, with the response options including White Americans, Black Americans, neither, and I don't know. We return to these items in an analysis collapsing across all experiments below.

Deviations From the Dasgupta and Greenwald (2001) Procedure. Although the present experiment was a close replication of Experiment 1 from Dasgupta and Greenwald (2001), here, we note all deviations from the original procedure of which we are aware.

First, the original experiment was conducted with a sample of undergraduate students in a university lab, whereas the present experiment was conducted online with a diverse sample of U.S. volunteers. Second, the instructions used to introduce the exposure and categorization tasks during the learning phase of the present experiment were slightly different from the ones used in the original experiment (for the specific deviations in wording, see online materials). We address this difference empirically in Experiment 2 below. Third, as described above, three exemplars used in the learning phase were replaced. Fourth, the IAT in the original experiment consisted of seven blocks, whereas in the present experiment, it consisted of five blocks. Fifth, participants in the original experiment completed the explicit evaluation measures using pen and paper, whereas participants in the present experiment completed them online. Finally, participants in the experimental condition of the present experiment completed four newly added exploratory measures of contingency memory.

Analytic Strategy

Statistical analyses were performed in the R statistical computing environment (Version 4.2.1).

Results

Implicit Racial Evaluations

Implicit racial evaluations by condition in this and all remaining experiments are shown in Figure 1.

Participants exhibited pro-White/anti-Black implicit evaluations both in the control condition (mean IAT D score = 0.33, $SD = 0.43$) and in the experimental condition ($M = 0.32$, $SD = 0.45$). Critically, the two conditions did not significantly differ from each other, with the Bayes Factor providing strong support for the null hypothesis, $t(1099.76) = 0.33$, $p = .744$, $\text{BF}_{01} = 14.07$, Cohen's $d = 0.02$. As such, the sizable condition difference obtained by Dasgupta and Greenwald (2001), corresponding to a Cohen's d effect size of 0.94, did not replicate in the present experiment.

Explicit Racial Evaluations

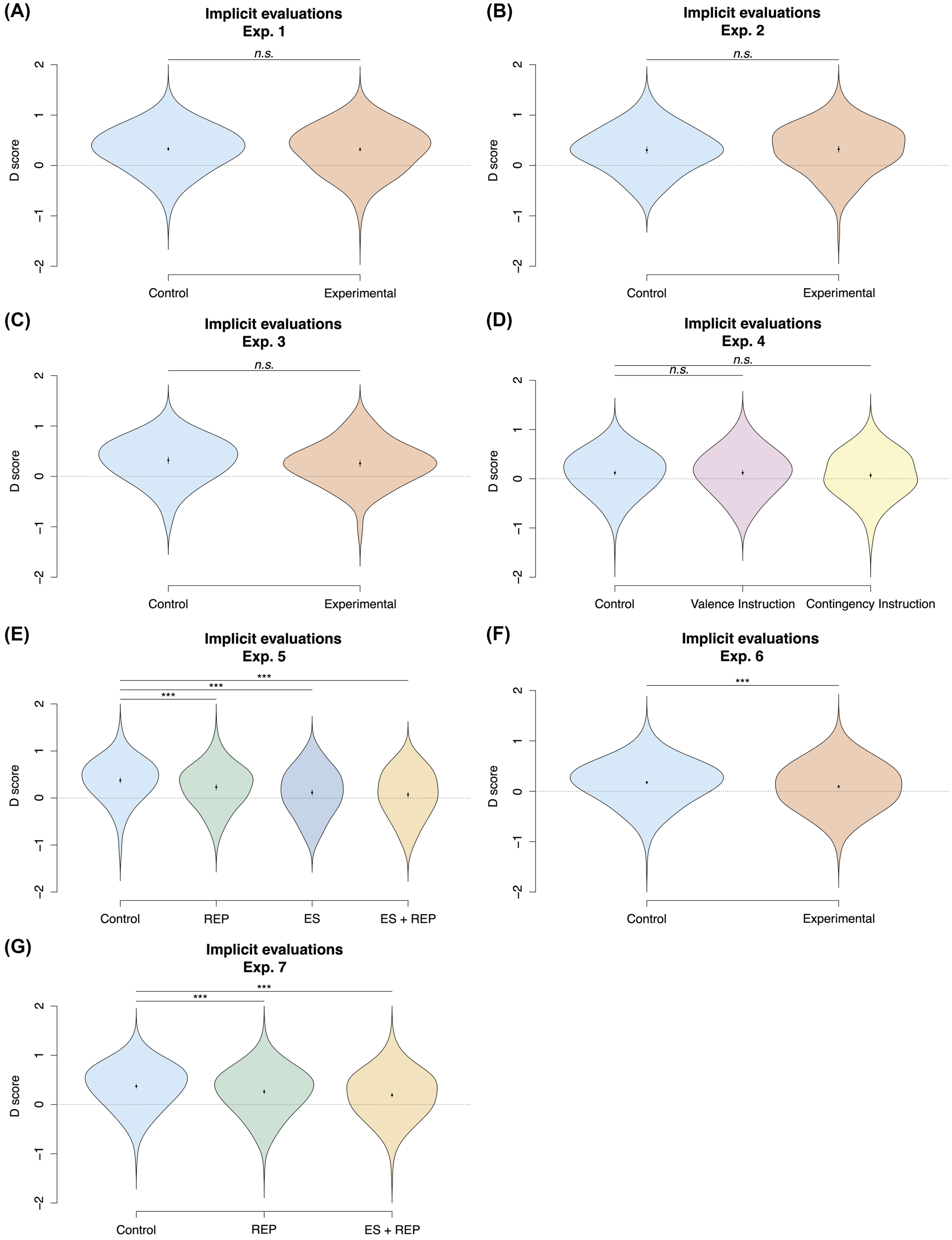
In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations both in the control condition (mean difference score = −12.61, $SD = 30.75$) and in the experimental condition ($M = -14.22$, $SD = 28.97$). Similar to implicit evaluations, the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(1053.20) = 0.89$, $p = .372$, $\text{BF}_{01} = 9.91$, Cohen's $d = 0.05$. This result is consistent with the one obtained by Dasgupta and Greenwald (2001) who also did not find a condition difference in explicit racial evaluations.

Moderation by Participant Race

Given that baseline implicit (and explicit) racial evaluations are known to vary by participant race (Nosek et al., 2007; Ratliff et al.,

Figure 1

The Results of Experiments 1–7



Note. Each panel, labeled (A) through (G), shows the results of an individual experiment. The *x*-axis shows experimental conditions, and the *y*-axis shows IAT *D* scores, with higher scores corresponding to higher levels of implicit pro-White/anti-Black evaluations. REP = repeated evaluative pairings; ES = evaluative statements; n.s. = not significant; IAT = Implicit Association Test. See the online article for the color version of this figure.

*** $p < .001$.

2020) and because the original Dasgupta and Greenwald (2001) experiment was conducted in a sample of only White American and Asian American participants, we sought to ascertain that participant race was not associated with any heterogeneity in intervention effectiveness in the present experiment.

Participant race did not moderate condition effects on implicit evaluations, $F(5, 1091) = 0.95$, $p = .450$, $\text{BF}_{01} = 262.98$, partial $\eta^2 < 0.01$. Although the frequentist analysis suggested that participant race moderated condition effects on explicit evaluations, the Bayesian analysis provided strong evidence for the null hypothesis and the effect size was small, $F(5, 1081) = 3.89$, $p = .002$, $\text{BF}_{01} = 10.12$, partial $\eta^2 = 0.02$. As such, we refrain from interpreting this effect.

Discussion

In an experiment that closely matched the procedural details of Experiment 1 from Dasgupta and Greenwald (2001), we found strong evidence against a difference between the control and experimental conditions. This result implies that, unlike in the original experiment, exposure to counterattitudinal (positive Black and negative White) exemplars did not shift implicit racial evaluations toward neutrality to any appreciable degree. In the remaining experiments, we first sought to ascertain the robustness of this finding and then probed whether appropriate psychological conditions could be created under which exposure to counterattitudinal exemplars shifts implicit racial evaluations toward neutrality.

Experiment 2: Close Replication II

When conducting Experiment 1, we erroneously believed that the original instructions introducing the exposure and categorization tasks in the learning phase had been lost. As such, we constructed new instructions to reflect the original experiment as best we could. Following completion of Experiment 1, we were unexpectedly able to locate the verbatim text of the instructions used by Dasgupta and Greenwald (2001). Although the instructions created for Experiment 1 were highly similar to the ones originally used (see online materials), given the lack of replication, we deemed it necessary to repeat the experiment using the original instructions. Otherwise, all procedural details of Experiment 2 were identical to those of Experiment 1.

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiment 1.

Participants and Design

Participants were 504 adult volunteers recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT ($n = 8$), or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 6$). These exclusions left 490 participants in the sample, of which we focus on the final sample of 360 U.S. participants below.

In the final sample, 242 participants were female and 118 participants male. Mean participant age was 40 years ($SD = 15$ years). 247 participants identified as White, 48 participants as Black, 29 participants as Hispanic, 13 participants as Asian, 13 participants as multiracial, four participants as Middle Eastern, four participants as Pacific Islander, and one participant as Native American.

Similar to Experiment 1, the experiment consisted of a learning phase and a test phase. In the learning phase, participants were assigned to an experimental condition ($n = 191$) involving exposure to positive Black and negative White exemplars or a procedurally matched control condition ($n = 169$). In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory items.

Procedure and Measures

The learning phase was highly similar to the learning phase of Experiment 1, with the crucial exception that the original instructions created by Dasgupta and Greenwald (2001) were used to introduce the exposure and categorization tasks.

The test phase was also highly similar to Experiment 1. As in Experiment 1, the explicit evaluation items had high internal consistency ($\alpha = .90$ for evaluations of White Americans and $\alpha = .91$ for evaluations of Black Americans) and were therefore aggregated to form a single index.

At the end of the experiment, two additional exploratory contingency memory items were collected, one measuring memory for the racial group membership of negative exemplars, and one asking participants to report when they became aware of the valence–racial group contingency (provided that they did). Response options for the latter item included not at all, while answering the questions at the end of the experiment, or during the learning task. Retrospective contingency memory items are an imperfect measure of contingency awareness for multiple reasons. Notably, completing contingency memory items can create post hoc contingency awareness without participants having developed contingency awareness during the learning phase itself (Gawronski & Walther, 2012; Kurdi et al., 2022). As such, the final item was included to improve the validity of the contingency memory items as measures of contingency awareness by addressing this possibility.

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations both in the control condition ($M = 0.30$, $SD = 0.43$) and in the experimental condition ($M = 0.32$, $SD = 0.47$). Critically, the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(357.48) = -0.42$, $p = .676$, $\text{BF}_{01} = 7.88$, Cohen's $d = -0.04$. As such, the present experiment, similar to Experiment 1, failed to replicate the condition difference originally obtained by Dasgupta and Greenwald (2001).

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations both in the control condition ($M = -15.38$, $SD = 34.00$) and in the experimental condition ($M = -16.13$, $SD = 34.43$). In line with the original findings of Dasgupta and Greenwald (2001), the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(345.59) = 0.21$, $p = .838$, $\mathrm{BF}_{01} = 8.31$, Cohen's $d = 0.02$.

Moderation by Participant Race

Participant race did not moderate condition effects on implicit evaluations, $F(5, 345) = 0.98$, $p = .428$, $\mathrm{BF}_{01} = 27.42$, partial $\eta^2 = 0.01$, or on explicit evaluations, $F(5, 337) = 0.78$, $p = .567$, $\mathrm{BF}_{01} = 90.14$, partial $\eta^2 = 0.01$.

Discussion

Experiment 2 was an even closer replication of Experiment 1 from Dasgupta and Greenwald (2001) than Experiment 1 given that it used the verbatim text of the original instructions to introduce the exposure and categorization tasks in the learning phase. Nevertheless, similar to Experiment 1 above, we obtained reliable evidence against any condition differences. That is, unlike in Dasgupta and Greenwald (2001), implicit evaluations once again did not shift toward neutrality as a result of exposure to counterattitudinal (positive Black and negative White) exemplars.

Experiment 3: Using More Strongly Valenced Exemplars

The failures to replicate observed in Experiments 1–2 led us to consider reasons external to the experimental procedure that may have caused implicit evaluations to remain recalcitrant in the Dasgupta and Greenwald (2001) paradigm. The social environment of the United States in general, and race relations in particular, have changed considerably since publication of the Dasgupta and Greenwald experiments in 2001. For instance, in the pretest for Experiment 1, we ourselves found that the positive Black and negative White exemplars used in the original experiment have changed in familiarity and valence.

Critically, it is possible that the shift in implicit evaluations observed by Dasgupta and Greenwald (2001) cannot be obtained using exemplars (and stimuli representing those exemplars) that were highly positive or negative 25 years ago but may not elicit the same psychological response today. As such, Experiment 3 closely followed the original Dasgupta and Greenwald (2001) procedure but used the most positive Black and negative White exemplars identified in the pretest. Moreover, the original, low-resolution grayscale images used by Dasgupta and Greenwald (2001) to represent the positive Black and negative White exemplars were replaced by more contemporary color images of higher quality.

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiment 2.

Participants and Design

Participants were 522 adult volunteers recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (*a*) did not complete the IAT ($n = 10$), or (*b*) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 15$). These exclusions left 497 participants in the sample, of which we focus on the final sample of 334 U.S. participants below.

In the final sample, 217 participants were female, 103 participants male, and 10 participants of other genders. Mean participant age was 34 years (*SD* = 16 years). Two hundred one participants identified as White, 35 participants as Black, 34 participants as Hispanic, 33 participants as multiracial, 22 participants as Asian, five participants as Middle Eastern, two participants as Pacific Islander, and one participant as Native American.

Similar to Experiments 1–2, this experiment consisted of a learning phase and a test phase. In the learning phase, participants were assigned to an experimental condition ($n = 172$) involving exposure to positive Black and negative White exemplars or a procedurally matched control condition ($n = 162$). In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory contingency memory items.

Procedure and Measures

The learning phase was highly similar to the learning phase of Experiment 2, with two crucial exceptions. First, we used the most highly valenced (most positive Black and most negative White) exemplars identified in the pretest reported in the Method section of Experiment 1. In this set, the mean WLI of Black exemplars was 54.42 ($SD = 11.29$; compared with mean = 47.94 and $SD = 20.31$ in Experiments 1–2) and the mean WLI of White exemplars was −63.01 ($SD = 14.12$; compared with mean = −54.29 and $SD = 22.47$ in Experiments 1–2). Second, in both conditions, the low-resolution grayscale images used by Dasgupta and Greenwald (2001) to represent the exemplars were replaced with more naturalistic high-resolution color images.

The test phase was identical to the test phase of Experiment 2. As in Experiment 2, the explicit evaluation items showed high internal consistency ($\alpha = .91$ for evaluations of White Americans and $\alpha = .90$ for evaluations of Black Americans) and were therefore aggregated to form a single index.

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations both in the control condition ($M = 0.32$, $SD = 0.43$) and in the experimental condition ($M = 0.26$, $SD = 0.42$). Critically, the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(329.86) = 1.41$, $p = .160$, $\mathrm{BF}_{01} = 3.20$, Cohen's $d = 0.15$. As such, the present experiment, similar to Experiments 1–2, failed to replicate the condition difference originally obtained by Dasgupta and Greenwald (2001).

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations both in the control condition ($M = -15.70$, $SD = 35.26$) and in the experimental condition ($M = -15.98$, $SD = 27.04$). In line with the findings of Dasgupta and Greenwald (2001), the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(296.19) = 0.08$, $p = .936$, $\text{BF}_{01} = 8.15$, Cohen's $d = 0.01$.

Moderation by Participant Race

Participant race did not moderate condition effects on implicit evaluations, $F(4, 320) = 1.70$, $p = .150$, $\text{BF}_{01} = 14.47$, partial $\eta^2 = 0.02$. Although the frequentist analysis suggested that participant race moderated condition effects on explicit evaluations, the Bayesian analysis provided anecdotal evidence for the null hypothesis and the effect size was small, $F(4, 311) = 3.23$, $p = .013$, $\text{BF}_{01} = 2.59$, partial $\eta^2 = 0.04$. As such, we refrain from interpreting this effect.

Discussion

Experiment 3 again failed to replicate the focal finding of Dasgupta and Greenwald (2001). That is, we observed no shift in implicit racial evaluations toward neutrality following exposure to counterattitudinal exemplars. Critically, the present experiment relied on contemporary images of the most strongly valenced exemplars identified in a pretest; as such, we conclude that the lack of replication is unlikely to have been due to insufficient familiarity with or insufficiently strong normative evaluations of the exemplars used.

Experiment 4: Testing a Subtyping Account

At its core, the Dasgupta and Greenwald (2001) finding demonstrating malleability in implicit evaluations following exposure to counterattitudinal exemplars is a finding of psychological generalization (Ranganath & Nosek, 2008; Staats et al., 1959): In the learning phase, participants engage with famous Black and infamous White individuals, whereas unfamiliar Black and White individuals are presented at test. The question of interest is whether exposure to famous and infamous individuals attitudinally transfers to new instances of the broader social categories.

Accordingly, one potential explanation of why a condition difference failed to emerge in Experiments 1–3 relies on the idea of subtyping (Hewstone & Hamberger, 2000; Kunda & Oleson, 1995). Specifically, participants may have perceived the Black and White exemplars to whom they had been exposed during the learning phase as not (or not sufficiently) representative of the broader racial categories. In other words, the Black and White exemplars—including politicians, gangsters, serial killers, and entertainers—may have been judged to be so remarkable in their atypicality that they were subtyped into a separate category (perhaps that of celebrities), and therefore learning failed to generalize to evaluations of the Black and White racial categories.

This account may be seen as unlikely to explain why we failed to replicate the Dasgupta and Greenwald (2001) finding given that subtyping did not seem to have operated in that study. However, changes in race representations and race relations over the past decades may have activated a form of subtyping that did not occur in the late 1990s. Specifically, given more broad-based presence of positive Black exemplars in media content beyond a few celebrities (Leonard & Robbins, 2021), individuals such as Colin Powell or Eddie Murphy may have come to be seen as too atypical to shift evaluations of Black Americans as a racial category. Moreover, starting with Experiment 4, our attention turned away from simple replication of Dasgupta and Greenwald (2001) and toward creating maximally advantageous conditions to produce an effect of counterattitudinal exemplars on implicit evaluations, if it exists.

In Experiment 4, we started doing so by ruling out the possibility of subtyping infamous and famous individuals by creating exposure to one set of unfamiliar Black (good) and White (bad) individuals in the learning phase and measuring implicit evaluations using a new set of unfamiliar Black and White individuals at test. As such, the present experiment still tests whether generalization occurs from exposure at the learning phase to the test phase, but avoids concerns about subtyping of famous exemplars by drawing the exemplars used at learning and test from the same pool of novel targets.

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiment 3.

Participants and Design

Participants were 1,560 adult volunteers recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT ($n = 30$) or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 18$). These exclusions left 1,512 participants in the sample, of which we focus on the final sample of 1,055 U.S. participants below.

In the final sample, 733 participants were female, 294 participants male, and 22 participants of other genders. Mean participant age was 39 years ($SD = 16$ years). Seven hundred four participants identified as White, 135 participants as Black, 83 participants as Hispanic, 68 participants as multiracial, 41 participants as Asian, 11 participants as Middle Eastern, nine participants as Native American, and two participants as Pacific Islander.

Similar to Experiments 1–3, the experiment consisted of a learning phase and a test phase. In the learning phase, participants were assigned to one of three conditions: (a) a valence instruction condition ($n = 324$); (b) a contingency instruction condition ($n = 348$); and (c) a control condition ($n = 383$). Whereas the two former conditions involved exposure to positive Black and negative White exemplars with different instructions preceding such exposure, the control condition was procedurally matched but did not present Black or White exemplars to participants. In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory items.

Materials

Behavioral Statements. Fifty pretested positive behavioral statements (e.g., "Lent a friend his new sleeping bag and tent to go camping.") and 50 pretested negative behavioral statements (e.g., "Sells vacuum cleaners door-to-door for three times the department store price.") were adapted for use from work by Cone and colleagues (Cone & Calanchini, 2021; Cone & Ferguson, 2015). These statements have been shown to shift implicit evaluations of single individuals in impression formation tasks.

Facial Images. Twenty Black male faces and 20 White male faces with neutral facial expressions were selected for use from the Chicago Face Database (Ma et al., 2015). Critically, both the Black faces (mean Black classification = 98%, $SD = 2\%$) and the White faces (mean White classification = 99%, *SD* = 2%) were high on racial prototypicality. Identifiers and additional norming data for each image are available on OSF (https://osf.io/spzx9/).

Procedure and Measures

Learning Phase. As in previous experiments, the learning phase consisted of an exposure task and a subsequent categorization task.

The procedure of the learning phase was identical across the valence instruction and contingency instruction conditions; the two differed from each other only in the initial instructions provided prior to the exposure task. Specifically, in the valence instruction condition, the initial instructions referred to moral and immoral behaviors (without mentioning the race of the targets) and asked participants to guess which person performed which behavior. As such, this condition stayed relatively close to the original Dasgupta and Greenwald (2001) procedure. In the contingency instruction condition, the initial instructions referred to moral Black and immoral White individuals and, as such, directed participants' attention both to the racial group membership of the targets and the contingency between race and valence of behaviors. This manipulation was implemented to maximize the possibility of obtaining an effect, should one exist in the population.

Following these initial instructions, participants in both conditions completed the same exposure task, which was modeled after the exposure tasks in Experiments 1–3. However, in this version of the exposure task, (a) the targets were novel, rather than well-known, and (b) the statements contained positive and negative behaviors rather than biographical details of famous and infamous individuals. Ten Black targets and 10 White targets were randomly selected for inclusion from the set of 40 facial images described above. Each Black target was randomly assigned to appear with one correct and one incorrect positive behavior and each White target was randomly assigned to appear with one correct and one incorrect negative behavior, selected from the set of 100 behaviors referred to above. Given that the targets and the behaviors were unknown to participants, initially participants had to guess which description was accurate and learned the correct description for each target over time. Therefore, the number of trials on the exposure tsk was doubled from 40 in Experiments 1–3 to 80 in the present experiment.

The categorization task was identical to the categorization task completed in Experiments 1–3, with the exception that targets' faces, rather than names, were displayed on each trial because the targets had not been referred to using names during the exposure task. As in the previous experiments, the control condition was procedurally matched to the experimental conditions but used flowers and insects instead of humans of different races as targets.

Test Phase. The test phase was also highly similar to previous experiments and consisted of measurement of implicit racial evaluations, measurement of explicit racial evaluations, and exploratory items.

The IAT used to measure implicit racial evaluations was procedurally identical to the IATs used in Experiments 1–3. However, the category labels were *White Americans* and *Black Americans* rather than *White* and *Black*. Moreover, the category stimuli were six facial images each, selected from the same set as the faces used in the learning phase but not identical to them. As such, the IAT in the present experiment still measures generalization, but generalization from one set of novel exemplars to another set of novel exemplars, rather than generalization from famous to nonfamous exemplars.

The explicit evaluation measures were identical to the ones administered in Experiments 1–3. Similar to the previous experiments, the explicit evaluation items had high internal consistency (α = .91 for evaluations of both White Americans and Black Americans) and were therefore aggregated to form a single index. The exploratory contingency memory items were also identical to the ones used in Experiments 2–3.

In addition, as the last item of the experiment, participants were randomly assigned to one of two exploratory items asking them to generate either a list of universally admired Black individuals or a list of universally disliked White individuals. This item served to facilitate stimulus construction in subsequent experiments and is thus not discussed further. However, the responses are available for reuse in the open data (https://osf.io/ckdzt/).

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations in all three conditions, including control (*M* = 0.12, *SD* = 0.42), valence instruction (*M* = 0.12, *SD* = 0.46), and contingency instruction (*M* = 0.07, *SD* = 0.46). We note the decrease in the overall magnitude of implicit racial evaluations, which—given the similarity in samples across Experiments 1–3 and Experiment 4—is most likely due to the different stimuli used on the IAT. Specifically, the name stimuli used in Experiments 1–3 are not customarily used in contemporary implicit social cognition research given that they confound race with social class.

Most importantly, the three conditions did not significantly differ from each other, with the Bayes Factor providing strong support for the null hypothesis, $F(2, 1052) = 1.63, p = .197, \mathrm{BF}_{01} = 18.72, \eta^2 < 0.01$. As such, the present experiment, similar to Experiments 1–3, failed to produce any differences across conditions.

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations in all three conditions, including control (*M* = −14.31, *SD* = 30.98), valence instruction (*M* = −12.78, *SD* = 29.92), and contingency instruction (*M* = −15.46, *SD* = 29.53). In line with Experiments 1–3, the three conditions did not significantly differ from each other, with the Bayes Factor providing very strong support for the null hypothesis, $F(2, 1034) = 0.65, p = .521, \mathrm{BF}_{01} = 47.38, \eta^2 < 0.01$.

Moderation by Participant Race

Participant race did not moderate condition effects on implicit evaluations, $F(13, 1030) = 0.74, p = .727, \mathrm{BF}_{01} = 742.26$, partial $\eta^2 < 0.01$. Although the frequentist analysis suggested that participant race moderated condition effects on explicit evaluations, the Bayesian analysis provided strong evidence for the null hypothesis and the effect size was small, $F(13, 1012) = 1.91, p = .026, \mathrm{BF}_{01} = 23.40$, partial $\eta^2 = 0.02$. As such, we refrain from interpreting this effect.

Discussion

In Experiment 4, we tested the hypothesis that the condition difference originally obtained by Dasgupta and Greenwald (2001) may not have emerged in the present replication attempts due to subtyping on the basis of fame. That is, participants may not have applied the evaluative information encountered in the context of famous Black American and infamous White American exemplars at learning to the more general categories of Black Americans and White Americans at test. The present experiment provides strong evidence against this possibility: We did not find any shifts in implicit racial evaluations in a study that closely mirrored the critical elements of the original Dasgupta and Greenwald (2001) procedure but used different exemplars drawn from the same set of nonfamous individuals at encoding and retrieval.

As such, contrary to initial expectations, we conclude that in diverse online samples of U.S. citizens, exposure to counterattitudinal exemplars does not seem to shift implicit racial evaluations toward neutrality in the early 2020s. Critically, the lack of replication was robust across different versions of the procedure, including (a) in close replication attempts (Experiments 1–2), (b) when participants were exposed to the most strongly valenced exemplars in the learning phase (Experiment 3), and (c) when the learning and test phase relied on the same pool of novel individuals and, as such, evaluative learning did not require generalization from famous exemplars to the more general category (Experiment 4). Notably, in Experiment 4, even explicitly drawing participants' attention to the race–valence contingency was not sufficient to produce a significant shift in implicit racial evaluations. We address the issue of contingency awareness more systematically in the remaining experiments and return to this issue in the General Discussion.

Experiment 5: Shifting Implicit Racial Evaluations Using Repeated Evaluative Pairings and Evaluative Statements

The failure to replicate the shifts in implicit racial evaluations observed by Dasgupta and Greenwald (2001) is puzzling for several reasons, the primary one being that other procedures, far less potent, have been shown to create malleability in implicit evaluations. Based on the extensive research by De Houwer and colleagues (De Houwer, 2006; De Houwer & Vandorpe, 2010; Gast & De Houwer, 2013; Zanon et al., 2014), Kurdi and Banaji (2017) demonstrated a surprising result: A simple verbal statement indicating which group is good and which group is bad can produce large shifts not only in implicit evaluations of novel targets but even in some of the strongest preexisting implicit evaluations of familiar groups (specifically, antielderly and antiforeign evaluations; for a related finding, see Hütter & De Houwer, 2017). In addition, the same experiments showed that repeated evaluative pairings of specific exemplars with valenced images can also create malleability in implicit age and nationality evaluations, albeit to a lesser degree than verbal statements do.

However, implicit racial evaluations may be unique, and we do not yet know whether the Kurdi and Banaji (2017) procedure would have demonstrated a change in implicit racial evaluations, as it did in the context of age and nationality. As such, both procedures (repeated evaluative pairings and evaluative statements) are implemented in Experiment 5 to conduct a first test of whether either or both of these interventions can shift implicit racial evaluations. If they do not, then implicit racial evaluations may be uniquely resistant to change (Kurdi, Krosch, & Ferguson, 2023); if they do, then we can conclude that implicit racial evaluations can shift although the Dasgupta and Greenwald (2001) paradigm is not sufficient to do so for theoretical reasons that remain to be identified.

More generally, single failures to replicate psychological findings are not always persuasive given that, in many cases, multiple studies may be needed to adequately replicate the original work. Moreover, such replication attempts leave the community of scientists and the interested public without a resolution as to when and why the effect may emerge and disappear. As such, the initial failures to replicate the Dasgupta and Greenwald (2001) result provided an opportunity to go further by attempting to isolate the factor(s) necessary to produce shifts in implicit racial evaluations, consistent with the original finding.

Method

Participants and Design

Participants were 1,007 White American adult volunteers recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT ($n = 19$) or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 5$). These exclusions left 983 participants in the final sample. In the final sample, 648 participants were female, 307 participants male, and 22 participants of other genders. Mean participant age was 37 years ($SD = 16$ years).

The experiment consisted of a learning phase and a test phase. In the learning phase, participants were assigned to one of four conditions: (a) control ($n = 257$), (b) repeated evaluative pairings (REP; $n = 241$), (c) evaluative statements (ES; $n = 246$), or (d) combined ES + REP ($n = 239$). The latter three conditions were designed to shift implicit racial evaluations toward neutrality using different procedures; the control condition was structurally matched with the ES + REP condition but did not involve exposure to valenced information in conjunction with racial exemplars. Similar to the previous experiments, the test phase consisted of implicit and explicit racial evaluation measures, followed by a set of exploratory contingency memory items.

Materials

Valenced Line Drawings. Five positive and five negative line drawings were borrowed for use from Kurdi and Banaji (2017).

Images of Black and White Exemplars. The same images used in Experiment 4, selected from the Chicago Face Database (Ma et al., 2015), were retained for use.

Procedure and Measures

The procedure and measures were modeled closely after Kurdi and Banaji (2017). As in the previous experiments, participants completed a learning phase whose content differed depending on condition assignment. The test phase involved measurement of implicit and explicit racial evaluations and the administration of some exploratory items.

Learning Phase. For the purposes of the learning phase, participants were assigned either to one of three procedures designed to shift implicit racial evaluations toward neutrality (REP, ES, and

ES + REP) or a procedurally matched control. The four conditions were designed to have the same approximate duration of 2.5 min. Further procedural details not mentioned here are described in detail in Kurdi and Banaji (2017) and are available in additional online materials (https://osf.io/spzx9/).

REP Condition. In this condition, participants were exposed to pairings of Black exemplars with positively valenced line drawings and White exemplars with negatively valenced line drawings. Five Black and five White exemplars were randomly selected from the set described above for use in the REP procedure.

Initial instructions informed participants that they would see two types of faces and two types of drawings and that they should learn the association between a certain type of face and a certain type of drawing (rather than associations between individual images). Race and valence—the dimensions along which the images differed—were not mentioned in the text of the instructions. Participants were then exposed to the full set of faces and the full set of line drawings and instructed again to learn the relationship between the images that they would see.

The learning procedure consisted of 36 stimulus pairings (18 Black–positive and 18 White–negative), presented in randomized order. Each trial consisted of (a) a fixation cross (presented for 500 ms), (b) the simultaneous presentation of a face and a line drawing (2,500 ms), and (c) an intertrial interval (500 ms). Line drawings and facial stimuli were sampled randomly, without replacement. When the entire set of images was exhausted, random sampling began anew.

ES Condition. In this condition, initial instructions informed participants that, over the course of the learning task, Black faces would always be paired with pleasant images and White faces would always be paired with unpleasant images. These instructions were presented multiple times to make the duration of this condition equal to that of the REP condition. Critically, participants were not exposed to actual stimulus pairings before completing the dependent measures; rather, instructions about upcoming stimulus pairings served as the sole source of valenced information.

ES + REP Condition. This condition consisted of a combination of the ES and REP conditions described above. Specifically, participants were first verbally informed of the Black–good and White–bad contingencies (like in the ES condition) and then exposed to actual stimulus pairings (like in the REP condition). To keep the duration of the three experimental conditions comparable to each other, in this condition, participants were exposed to 20 (rather than 36) stimulus pairings.

Control Condition. The control condition was procedurally matched to the ES + REP condition. However, participants were exposed to either (a) pairings of positive line drawings with other positive line drawings, negative line drawings with other negative line drawings, Black faces with other Black faces, and White faces with other White faces or (b) pairings of positive line drawings with negative line drawings and of Black faces with White faces. In this way, the control condition exposed participants to the same materials the same number of times as the ES + REP condition did but did not create any contingency between racial groups and valences.

Test Phase. Similar to the previous experiments, the test phase consisted of measurement of implicit racial evaluations using the IAT, measurement of explicit racial evaluations using self-report items, and exploratory measures.

Implicit Racial Evaluations. The procedure of the IAT was the same as in previous experiments but used different stimuli, as described below. Moreover, unlike in the previous experiments and in keeping with the procedure of Kurdi and Banaji (2017), the order of critical blocks was not randomized. Rather, given the focus on the relative effectiveness of the three interventions relative to baseline (as opposed to absolute deviation from zero), all participants completed the Black people/good–White people/bad block first. This order created alignment between the information to which participants had been exposed in the learning phase and the first critical block of the IAT.

Category labels on the IAT were *White people* and *Black people*. In keeping with the procedure of Kurdi and Banaji (2017), the same Black and White exemplars were used during the learning phase and on the IAT. As such, the present experiment, unlike Experiment 1 of Dasgupta and Greenwald (2001) and the present Experiments 1–4, does not constitute a test of generalization. We return to this issue in Experiments 6–7 below. Attribute labels on the IAT were *Pleasant* and *Unpleasant*. Attribute stimuli for the former attribute included *love*, *peace*, *happy*, *sweet*, *glory*, and *success*, and for the latter attribute included *hate*, *war*, *devil*, *bomb*, *bitter*, *agony*, and *failure*.

Explicit Racial Evaluations. Explicit evaluations of each racial group were measured using (a) a feeling thermometer item, (b) a good–bad semantic differential item, and (c) an honest–dishonest semantic differential item. Participants used 101-point sliding scales to enter their responses. All six items were presented in individually randomized order. Given acceptable internal consistency among the items measuring explicit evaluations of White Americans (Cronbach's $\alpha = .80$) and Black Americans ($\alpha = .78$), the items were aggregated, and a relative index was created, as in the previous experiments.

Exploratory Measure. Contingency memory was measured using an item that asked participants to indicate what they had learned in the learning phase of the experiment. The options included (a) Black people are good and White people are bad (correct response in REP, ES, and ES + REP), (b) White people are good and Black people are bad (foil), and (c) nothing that would have indicated whether the groups are good or bad (correct response in control).

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations in all four conditions, including control ($M = 0.38$, $SD = 0.44$), REP ($M = 0.23$, $SD = 0.45$), ES ($M = 0.12$, $SD = 0.49$), and ES + REP ($M = 0.07$, $SD = 0.50$). Most importantly, unlike in Experiments 1–4, the four conditions clearly and significantly differed from each other, with the Bayes Factor providing extreme support for the alternative hypothesis, $F(3, 979) = 20.75$, $p < .001$, $\mathrm{BF}_{10} = 1.15 \times 10^{10}$, $\eta^2 = 0.06$.

In following up on the significant omnibus test, we found that implicit racial evaluations shifted toward neutrality in all three conditions relative to control (all $p < .001$). The ES and ES + REP conditions both produced significantly larger shifts than the REP condition did ($p \leq .007$), whereas the ES and ES + REP conditions did not differ from each other ($p = .281$). This pattern of results perfectly mirrors the findings reported by Kurdi and Banaji (2017) with respect to novel social targets and preexisting categories such as age and nationality.

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations in all four conditions, including control ($M = -2.41$, $SD = 9.05$), REP ($M = -3.38$, $SD = 9.45$), ES ($M = -5.34$, $SD = 13.27$), and ES + REP ($M = -3.61$, $SD = 13.48$). Also in a deviation from implicit evaluations, although the four conditions significantly differed from each other in the frequentist analysis, the Bayesian analysis provided moderate evidence for the null hypothesis and the effect size was exceedingly small, $F(3, 958) = 2.78$, $p = .040$, $BF_{01} = 6.28$, $\eta^2 < 0.01$. As such, we refrain from interpreting this effect.

Discussion

Experiment 5 provides solid evidence supporting the idea that implicit racial evaluations can shift. Moreover, it demonstrates that such shifts are not only possible, but that conditions that were present in the Dasgupta and Greenwald (2001) studies, including (a) the use of college samples and (b) physical presence of an experimenter, are not required to produce such effects. Indeed, in the present experiment, implicit evaluations shifted significantly and considerably toward neutrality in response to three manipulations, the first one involving exposure to pairings of category members with valenced images, the second one involving verbal descriptions describing such stimulus pairings, and the third one relying on a combination of both manipulations. In fact, the pattern of findings involving implicit racial evaluations in the present experiment mirrored perfectly the results obtained by Kurdi and Banaji (2017) on tests of age and nationality. Implicit racial evaluations are malleable.

The results of Experiment 5 suggest that the puzzle posed by the lack of replications observed in Experiments 1–4 is explained by some specific aspect(s) of the Dasgupta and Greenwald (2001) paradigm, potentially in conjunction with broader societal changes that have unfolded in the United States since the 1990s. With the knowledge that implicit racial evaluations are malleable and that shifts in implicit evaluations can be achieved in tasks administered virtually rather than in person, two final experiments were conducted to identify the psychological conditions required to produce the effect. Specifically, inspired by recent propositional accounts of implicit evaluation (De Houwer, 2014; Ferguson et al., 2019; Kurdi & Dunham, 2020; Mandelbaum, 2016), we sought to use the Dasgupta and Greenwald (2001) procedure but modified it in such a way as to direct participants' attention to (*a*) the racial category membership of the exemplars and the valence of descriptions and (b) the contingency between the race and valence.

Experiment 6: Directing Attention to the Race–Valence Contingency I

When the Dasgupta and Greenwald (2001) experiments were conducted in the late 1990s, the overwhelming theoretical consensus about the nature of implicit evaluations (and the underlying attitudinal representations) was that (a) they reflect co-occurrences of categories (such as Black Americans and White Americans) with attributes (such as positive and negative) in the environment and (b) that these co-occurrences are registered in an automatic and stimulus-driven manner (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). That is, passive exposure to positively valenced Black exemplars and negatively valenced White exemplars should be sufficient to shift implicit racial evaluations toward neutrality; intentional processing of the stimuli was thought to be unnecessary. Specifically, on this view, whether participants attend to the racial category membership of exemplars, the valence of descriptions, or the contingency between the two should not modulate responding on the IAT and other indirect measures of social cognition.

However, theoretical understanding of how the attitudinal representations reflected by implicit evaluations are acquired and how implicit evaluations can shift in the face of counterattitudinal information has changed considerably over the past two decades. Indeed, several contemporary accounts recognize the possibility that the way in which participants engage with information to which they are exposed can modulate the updating of implicit evaluations because implicit evaluations reflect not merely the number of co-occurrences experienced in the environment but also the inferences that participants make from those co-occurrences. This view was foreshadowed by the Ohio State attitude theorists who in the 1960s suggested that it was not the message but rather the cognitive response to it that determined persuasion (Greenwald et al., 1968). Specifically, it has now been demonstrated that (a) a major predictor of the extent of evaluative learning in general (e.g., Pleyers et al., 2007; Stahl & Unkelbach, 2009) and implicit evaluation updating in particular (e.g., Hofmann et al., 2010; Stahl et al., 2009) is awareness of the contingency between targets and positive and negative valence, and (b) verbal information about the nature of stimulus pairings can influence the way in which stimulus pairings produce shifts in implicit evaluations (Kurdi, Morehouse, & Dunham, 2023).

If awareness of (a) racial categories, (b) valence, and (c) the relationship between the two is a precondition for (or at least facilitator of) shifts in implicit evaluations, then the exposure task constituting the critical manipulation of the original Dasgupta and Greenwald (2001) paradigm can be said to create suboptimal conditions for learning because participants' attention is (a) not directed toward racial categories (in fact, race remains entirely unmentioned before the racial categorization task), (b) not directed toward valence categories (in fact, because the accurate and inaccurate descriptions are matched on valence, participants' attention is directed to specific semantic content rather than valence), and (c) not directed toward the contingency between racial categories and valence (in fact, the "knowledge test" framing creates incidental learning conditions when it comes to race, valence, and the relationship between the two).

As such, the goal of the present experiment was to increase participants' awareness of exemplars' racial category membership, description valence, and the contingency between the two. Specifically, (a) both the initial instructions and each accurate and inaccurate description during the exposure task mentioned exemplars' racial group membership, (b) the initial instructions mentioned valence ("admired Black people" vs. "disliked White people"), and (c) participants were asked to focus on the relationship between racial categories and valence as they were completing the exposure task. According to the propositional perspectives (De Houwer, 2014; Ferguson et al., 2019; Kurdi & Dunham, 2020; Mandelbaum, 2016) and related accounts focusing on the role of contingency awareness in evaluative learning (e.g., Corneille & Stahl, 2019), implicit racial evaluations should be considerably more likely to exhibit malleability under these conditions.

Method

Participants and Design

Participants were 1,633 adult volunteers from the United States recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT ($n = 31$) or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention ($n = 17$). These exclusions left 1,585 participants in the final sample.

In the final sample, 1,107 participants were female, 407 participants male, and 49 participants of other genders. Mean participant age was 37 years ($SD = 15$ years). One thousand seven participants identified as White, 191 participants as Black, 187 participants as Hispanic, 98 participants as multiracial, 71 participants as Asian, 12 participants as Middle Eastern, 12 participants as Native American, and five participants as Pacific Islander.

Similar to all previous experiments, Experiment 6 consisted of a learning phase and a test phase. In the learning phase, participants were assigned to an experimental condition ($n = 795$) designed to shift implicit racial evaluations or a procedurally matched control condition ($n = 790$). In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory contingency memory items.

Procedure and Measures

Learning Phase. The learning phase was identical to the learning phase of Experiment 2, which followed closely the procedure by Dasgupta and Greenwald (2001), with two major deviations.

First, the initial instructions preceding the exposure task explicitly referred to the contingency between racial groups and valences ("admired Black people such as artists, athletes, and leaders" vs. "disliked White people such as mass murderers, terrorists, and criminals"). In addition, participants were specifically asked to focus on the broader racial group membership of the exemplars as they were completing the task. This feature of the paradigm represents a critical departure from Dasgupta and Greenwald (2001) in which the "knowledge test" framing, which did not contain explicit mention of race or stimulus valence, created incidental learning conditions with respect to both and with respect to the contingency between them.

Second, to consistently remind participants of exemplar race, the correct and incorrect descriptions used in the exposure task were modified to include explicit mention of the person's racial group membership. For example, the correct description for Colin Powell was edited to be "former Black Chairman, Joint Chiefs of Staff for the U.S. Department of Defense" and the incorrect description was edited to be "Black U.S. Ambassador to the United Nations."

Test Phase. The test phase was also highly similar to the test phase of Experiment 2, and thus to Experiment 1 of Dasgupta and Greenwald (2001), with the exception that the IAT used *White Americans* and *Black Americans* as category labels and a set of six images from the Chicago Face Database (Ma et al., 2015) from each category as category stimuli to remove the social class confound discussed in the Method section of Experiment 4 above. As in the previous experiments, the items measuring explicit evaluations of White Americans ($\alpha = .91$) and Black Americans ($\alpha = .90$) were highly reliable and were thus used to create an index of overall explicit evaluations.

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations both in the control condition ($M = 0.17$, $SD = 0.42$) and in the experimental condition ($M = 0.09$, $SD = 0.46$). Critically, unlike in Experiments 1–4, the two conditions significantly differed from each other, with the Bayes Factor providing very strong support for the alternative hypothesis, $t(1574.37) = 3.63$, $p < .001$, $\text{BF}_{10} = 36.95$, Cohen's $d = 0.18$. That is, the manipulation implemented in this experiment led to a significant shift of implicit racial evaluations toward neutrality.

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations both in the control condition ($M = -11.83$, $SD = 29.16$) and in the experimental condition ($M = -13.61$, $SD = 30.99$). Unlike with implicit evaluations as the dependent measure, the two conditions did not significantly differ from each other, with the Bayes Factor providing moderate support for the null hypothesis, $t(1545.24) = 1.17$, $p = .243$, $\text{BF}_{01} = 8.94$, Cohen's $d = 0.06$.

Moderation by Participant Race

Participant race did not moderate condition effects on implicit evaluations, $F(7, 1567) = 1.15$, $p = .328$, $\text{BF}_{01} = 64.80$, partial $\eta^2 < 0.01$, or on explicit evaluations, $F(7, 1536) = 1.48$, $p = .170$, $\text{BF}_{01} = 108.96$, partial $\eta^2 < 0.01$.

Discussion

In Experiment 6, we modified the initial instructions and the stimuli used during the exposure task to increase awareness of (a) Black and White racial categories, (b) positive and negative valence, and (c) the fact that all Black exemplars are positive and all White exemplars are negative. With these modifications in place, we obtained a small but statistically significant difference between the control and experimental conditions. This result deviated from the findings of Experiments 1–4 reported above in which no shift in implicit evaluations was produced in paradigms that followed the original Dasgupta and Greenwald (2001) in creating incidental learning conditions with respect to exemplar race, description valence, and the contingency between the two.

As such, in line with propositional perspectives (De Houwer, 2014; Ferguson et al., 2019; Kurdi & Dunham, 2020; Mandelbaum, 2016) and related accounts of evaluative learning (e.g., Corneille & Stahl, 2019), the present findings suggest that counterattitudinal exemplars have the potential to shift implicit racial evaluations toward neutrality, provided that participants become aware of the race–valence contingency. If they do not, shifts in implicit evaluation may not occur at all, or they may occur to a considerably weaker degree that was not detectable on the previous tests. However, the role of contingency awareness notwithstanding, the present results need not imply that shifts in implicit evaluation were mediated by inferential reasoning. We return to issues of cognitive mechanism in more detail in the General Discussion below.

Experiment 7: Directing Attention to the Race–Valence Contingency II

The results of Experiment 6 suggest that exposure to counterattitudinal exemplars has the potential to shift implicit racial evaluations toward neutrality provided that participants become aware of the Black–good and White–bad contingencies during learning. In Experiment 7, we sought to examine the robustness of this finding by probing whether it can emerge in a procedure considerably different from the one implemented in Experiment 6.

Specifically, in the present experiment, participants were passively exposed to pairings of famous Black and infamous White exemplars with a single correct description rather than attempting to distinguish the correct description applicable to each individual from a valence-matched foil. This modification was implemented because we reasoned that the choice task implemented by Dasgupta and Greenwald (2001) may have inadvertently directed participants' attention to the semantic details of individuals' biographies rather than their valence, which was held constant, thus interfering with evaluative learning processes (Gast & Rothermund, 2011). Driven by the same consideration, the descriptions were also shortened, with only the gist of the information retained.

In addition, Experiment 6 raises the possibility that exposure to counterattitudinal exemplars can shift implicit racial evaluations only to the extent that the Black–good and White–bad contingencies are made explicit to participants by the experimenter. However, if this were the case, then this would (a) severely limit the usability of exposure to counterattitudinal exemplars as an intervention and (b) suggest that incidental exposure to counterattitudinal Black and White individuals in one's daily life is unlikely to produce any shifts in implicit racial evaluations (let alone changes in the underlying racial attitudes). To directly probe whether this is the case, in the present experiment we implemented (a) a REP-like condition in which participants were merely instructed to learn the relationship between two types of individuals and two types of descriptions without mentioning race or valence ahead of time and (b) an ES + REP-like condition in which the Black–good and White–bad contingencies were made fully explicit prior to the exposure task.

Method

Participants and Design

Participants were 1,566 White American volunteers recruited via Project Implicit. As preregistered, we excluded participants from subsequent analyses if they (a) did not complete the IAT (n = 30) or (b) had response latencies of 300 ms or lower on at least 10% of IAT trials, indicating inattention (n = 8). These exclusions left 1,528 participants in the final sample. In the final sample, 1,013 participants were female, 478 participants male, and 19 participants of other genders. Mean participant age was 43 years (SD = 15 years).

Similar to all previous experiments, Experiment 7 consisted of a learning phase and a test phase. In the learning phase, participants were assigned to one of three conditions: (a) REP (n = 508); (b) ES + REP (n = 535); and (c) control (n = 485). Whereas the two former conditions involved exposure to positive Black and negative White exemplars in the same procedure but with different instructions preceding such exposure, the control condition was procedurally matched to the REP condition but did not present Black or White exemplars to participants. In the test phase, implicit and explicit racial evaluations were measured, followed by a set of exploratory contingency memory items.

Procedure and Measures

Learning Phase. Similar to Experiment 5 but unlike the remaining experiments, the learning phase consisted solely of an exposure task, without a subsequent categorization task.

The procedure of the learning phase was identical across the REP and ES + REP conditions; the two differed from each other only in the initial instructions provided prior to the exposure task. Specifically, in the REP condition (just as the REP condition of Experiment 5), the initial instructions asked participants to learn the relationship between two types of faces and two types of descriptions without referring to racial group membership or stimulus valence. In the ES + REP condition (just as the ES + REP condition of Experiment 5), the initial instructions explicitly mentioned the race of the exemplars and the contingency between race and valence.

Following these initial instructions, participants in both conditions completed the same exposure task, which was procedurally modeled after the REP and REP + ES conditions of Experiment 5. That is, participants passively watched pairings between exemplars and descriptions, as opposed to Experiment 1 of Dasgupta and Greenwald (2001) in which each trial involved participants making a choice between two descriptions (one accurate and one inaccurate). The exemplars were the most strongly valenced exemplars used in Experiment 3, with the exception that Will Smith was replaced by John Lewis. The reason for this is that the experiment was conducted after Will Smith slapped Chris Rock at the Academy Awards on March 27, 2022, which may have shifted societal evaluations of him in the negative direction. Moreover, to ease the cognitive load imposed on participants, the descriptions were shortened to contain one to three words, for example, "civil rights leader" or "convicted terrorist."

The control condition was procedurally matched to the REP condition but involved exposure to exemplars of flowers and insects rather than Black and White individuals. Unlike in Experiment 5, the number of stimulus pairings in all three conditions was fixed to 40.

Test Phase. The test phase was highly similar to the test phase of Experiment 2, and thus to Experiment 1 of Dasgupta and Greenwald (2001), with the exception that the IAT used *White Americans* and *Black Americans* as category labels and a set of six grayscale facial images from the Project Implicit demonstration website (https://implicit.harvard.edu/) from each category as category stimuli to remove the social class confound discussed in the Method section of Experiment 4 above. As in the previous experiments, items measuring explicit evaluations of White Americans (α = .91) and Black Americans (α = .90) were highly reliable and were thus used to create an index of overall evaluation.

Results

Implicit Racial Evaluations

Participants exhibited pro-White/anti-Black implicit evaluations in all three conditions, including control (M = 0.37, SD = 0.44), REP (M = 0.26, SD = 0.47), and ES + REP (M = 0.19, SD = 0.48).

Most importantly, the three conditions substantially and significantly differed from each other, with the Bayes Factor providing extreme support for the alternative hypothesis, $F(2, 1525) = 19.32$, $p < .001$, $\text{BF}_{10} = 1.03 \times 10^6$, $\eta^2 = 0.03$. In following up on the significant omnibus test, we found that each pairwise difference between conditions was significant ($p \leq .016$). That is, both the REP and the ES + REP condition significantly shifted implicit racial evaluations toward neutrality but the latter to a larger extent than the former.

Explicit Racial Evaluations

In a deviation from implicit evaluations, participants exhibited pro-Black/anti-White explicit evaluations in all three conditions, including control ($M = -9.55$, $SD = 24.43$), REP ($M = -9.55$, $SD = 24.08$), and ES + REP ($M = -8.90$, $SD = 21.47$). Unlike with implicit evaluations as the dependent measure, the three conditions did not significantly differ from each other, with the Bayes Factor providing extreme support for the null hypothesis, $F(2, 1476) = 0.13$, $p = .877$, $\text{BF}_{01} = 110.86$, $\eta^2 < 0.01$.

Discussion

Experiment 7 provides evidence for the robustness of the finding that exposure to counterattitudinal exemplars can create shifts in implicit racial evaluations toward neutrality, provided that participants are aware of the race–valence relationship. Notably, although the ES + REP condition, which made the Black–good and White–bad contingencies explicit to participants, produced a larger effect than the REP condition, which did not, implicit evaluations also shifted significantly in the latter condition relative to control. Overall, these results suggest that although awareness of the race–valence contingency considerably modulates the extent to which implicit evaluations exhibit malleability in response to counterattitudinal exemplars, the contingency need not be made explicit by the experimenter. Rather, participants may be able to detect the contingency on their own if their attention is not distracted by some other task goal and, once they do, implicit evaluations can shift in response to this self-generated recognition.

Exploratory Analyses of Contingency Memory (Experiments 1–7)

In terms of mean levels, the present experiments suggest that awareness of the contingency between racial categories and valence is a critical moderator of whether exposure to counterattitudinal exemplars leads to shifts in implicit evaluations. Specifically, although Experiments 1–3 and 6–7 relied on the same famous and infamous individuals, implicit evaluations shifted in the former set of experiments and not in the latter. We reasoned that the critical difference between the two sets of experiments was whether participants' attention was directed (more or less explicitly) toward the Black–good and White–bad relationships present in the stimuli used during the exposure task.

If this is the case, and contingency awareness is a major predictor of malleability in implicit evaluations, then we should expect the same relationship to emerge not only at the level of experiments but also at the level of individual participants. To this end, each experiment reported above included the collection of a set of memory measures at the end of the procedure to determine whether participants were able to accurately report the contingency between racial categories and valences. Such post hoc measures of contingency memory are not without methodological limitations (Gawronski & Walther, 2012; Kurdi et al., 2022). Nevertheless, they can be helpful in determining whether accurate declarative memory of the Black–good and White–bad contingencies predicts the extent of malleability observed on the implicit and explicit evaluation measures at the level of individual participants.

Participants were deemed to be contingency aware if they accurately reported either (a) the Black–good contingency or (b) the White–bad contingency and (c) indicated that they had become aware of this contingency during the learning task, as opposed to at the end of the experiment while answering the contingency memory questions (only Experiments 2–4 and 6–7). Choosing the appropriate threshold for contingency awareness involves an inherent tradeoff between sensitivity and specificity, and relatively conservative or liberal criteria may lead to different conclusions (Kurdi et al., 2022; Moran et al., 2021). As such, we encourage readers to explore the present data with alternative criteria for contingency awareness if they wish.

In a mixed-effects model using data from all seven experiments, with implicit evaluations as the dependent variable, contingency memory as the sole fixed effect, and random intercepts for experiments, we obtained a significant effect of contingency memory, $\chi^2(2) = 66.19$, $p < .001$. Specifically, in the control condition, participants exhibited significant pro-White/anti-Black implicit evaluations, $\beta_0 = 0.54$, $t(6.71) = 9.06$, $p < .001$. We observed a reduction in implicit racial evaluations both among participants with inaccurate contingency memory, $\beta = -0.10$, $t(6713.58) = -3.39$, $p < .001$, and among participants with accurate contingency memory, $\beta = -0.20$, $t(6831.46) = -8.15$, $p < .001$. Critically, implicit racial evaluations shifted significantly more strongly in the latter than in the former group, $\beta = 0.10$, $t(6604) = 3.18$, $p = .002$.

Contingency memory also significantly predicted the magnitude of shifts in explicit evaluations across all seven experiments, $\chi^2(2) = 14.41$, $p < .001$. In the control condition, participants exhibited significant pro-Black/anti-White explicit evaluations, $\beta_0 = -0.37$, $t(13.74) = -17.11$, $p < .001$. Explicit evaluations did not shift among participants with inaccurate contingency memory, $\beta = 0.03$, $t(1220.49) = 1.27$, $p = .206$, but they shifted toward a stronger pro-Black/anti-White stance among participants with accurate contingency memory, $\beta = -0.07$, $t(5027.44) = -2.80$, $p = .005$. The two groups of participants significantly differed from each other, $\beta = 0.11$, $t(680) = 3.42$, $p < .001$.

General Discussion

In three high-powered (total $N > 1,800$) and close-to-exact replications, we failed to obtain the effect originally reported by Dasgupta and Greenwald (2001). That is, we found no reduction in pro-White/anti-Black implicit evaluations after exposure to positive Black and negative White exemplars (Experiments 1–3). Given the substantial amount of time that has elapsed since the original results were published, we can only make informed guesses about the reasons for the lack of replication, without the certainty afforded by direct experimental tests.

At a first approximation, it is conceivable that the original result was a false positive, in which case one should expect replication attempts to yield null results. Contrary to this possibility, some of the experiments conducted as part of the only known previous independent replication attempt by Joy-Gaba and Nosek (2010) produced statistically significant results. As such, we believe that it is more likely that the effect originally obtained in the late 1990s decreased in size over time, both between 2001 and 2010 and between 2010 and 2023. Alternatively, or in addition, implicit racial evaluations may now be more difficult to shift than they were 25 years ago given that baseline levels of pro-White/anti-Black evaluations have decreased considerably (Charlesworth & Banaji, 2019). However, this perspective does not explain why sizable shifts in implicit evaluations were obtained in the present Experiments 5–7.

Notably, the original Dasgupta and Greenwald (2001) experiments were conducted in samples of University of Washington undergraduates who completed the entire procedure in person and in a research lab where they interacted with an experimenter. In contrast, in the present experiments, participants were adult U.S. volunteers who completed the intervention and all dependent measures online and in anonymity. It may be the case that these contextual differences in experimental setting (in-person vs. online, identifiable vs. anonymous, student sample vs. more heterogeneous adult sample) are, collectively or individually, responsible for the lack of replication (see Table 1). Future research should consider replicating the original research in an in-person lab context where a live experimenter interacts with a college sample. These features of the original experiment may help produce the effect if they increase participants' attention and accountability, thereby facilitating their awareness of the contingency between race and valence, which we found to be a major predictor of intervention effectiveness.

At the same time, two different procedures adapted from Kurdi and Banaji (2017), one relying on repeated evaluative pairings and the other on evaluative statements, led to significant and sizable shifts in implicit racial evaluations toward neutrality in the present work (see Experiment 5). As such, it seems that features such as a college sample, administering the experiment in a campus building, and physical presence of an experimenter are not necessary to produce reliable shifts in pro-White/anti-Black implicit evaluations toward neutrality. In line with these considerations, the replication attempt by Joy-Gaba and Nosek (2010) found no moderating effect of either study setting (in-person vs. online) or study sample (college student vs. adult volunteers).

Once the initial experiments had established that the original Dasgupta and Greenwald (2001) paradigm did not produce any shifts in implicit racial evaluations, we turned to identifying conditions under which exposure to positive Black and negative White exemplars might reduce implicit pro-White/anti-Black evaluations. First, we established that the null results obtained in the initial close replication attempts were not due to two potential alternative explanations of insufficiently strong exemplar valence (Experiment 3) or subtyping on the basis of fame (Experiment 4). Then, inspired by recent propositional accounts of implicit evaluation (De Houwer, 2014; Ferguson et al., 2019; Kurdi & Dunham, 2020; Mandelbaum, 2016) and related empirical (Pleyers et al., 2007; Stahl et al., 2009; Stahl & Unkelbach, 2009) and theoretical perspectives (Corneille & Stahl, 2019), we reexamined the Dasgupta and Greenwald (2001) procedure and implemented some changes to make it more likely that a shift in implicit racial evaluations would occur.

Specifically, although at the time of conducting the Dasgupta and Greenwald (2001) experiments, the human capacity for effortful propositional thought was broadly recognized in experimental psychology, few if any theories accounted for the possibility that such processes may play a role in implicit social cognition. Rather, the formation and updating of implicit evaluations was thought to unfold via low-level, stimulus-driven processes that resulted in the incidental recording of co-occurrences present in the environment, irrespective of the way in which the participant cognitively engaged with those co-occurrences (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Reflecting this theoretical understanding, the Dasgupta and Greenwald (2001) procedure created incidental learning conditions with respect to (a) the exemplars' racial group membership, (b) the valence of the descriptions, and (c) the contingency between the two.

It is now well-established that contingency awareness is a major driver of evaluative learning even in the context of implicit evaluations (Corneille & Stahl, 2019; Hofmann et al., 2010) and that the inferences that participants make can considerably modulate shifts in implicit evaluations in response to stimulus co-occurrences (Kurdi, Morehouse, & Dunham, 2023). As such, we sought to create conditions under which participants would be more likely to become aware of the Black–good and White–bad contingencies and reasoned

Table 1
Table of Limitations

Potential limitations of the present experiments
• The present experiments were conducted online. Given that in-person experiments are less anonymous than online experiments are, accountability is generally higher. As such, it is conceivable that implicit racial evaluations may shift using the original Dasgupta and Greenwald (2001) procedure in the lab although no such effect was obtained in online samples (but see Joy-Gaba & Nosek, 2010).
• Female, liberal, young, and highly educated participants are overrepresented in the online volunteer samples on which we rely in the present experiments. It is unclear whether the results would generalize to representative samples that are more balanced in terms of gender, political orientation, age, and education level or to individuals who do not volunteer their time to participate in experiments on racial attitudes.
• In the present project, we measured intervention effectiveness immediately following the learning phase. It is unclear whether the temporary shifts in implicit evaluations observed in some of the present experiments would persist over time.
• The experiments reported in this article used short, single-session interventions to shift implicit racial evaluations. The present results (especially the results involving contingency awareness) may not generalize to settings in which interventions are administered across multiple occasions and/or over longer periods of time.
• Given their reliance on famous Black American and infamous White American exemplars, the present experiments are specific to the U.S. context. Although we have no reason to believe that the theoretical conclusion about the role of contingency awareness in the malleability of implicit evaluations is specific to the United States, given cultural differences, the specific findings involving the present stimulus materials may not generalize to other countries.

that such awareness, in turn, would be conducive to malleability in implicit evaluations. These predictions were confirmed both in a procedure that made the race–valence contingency fully explicit prior to the exemplar exposure (Experiment 6) and in a procedure that merely instructed participants to learn the relationship between two types of individuals and two types of descriptions without fully elucidating the nature of the contingency (or even referring to race or valence; Experiment 7).

We also confirmed the importance of contingency awareness at the level of individual participants. In an analysis collapsing across all experiments, we demonstrated that participants able to report the Black–good or White–bad contingency following the experiment exhibited double the effect size relative to participants who were not. Needless to say, retrospective declarative memory of contingencies is not a perfect measure of online contingency awareness during learning (Gawronski & Walther, 2012; Kurdi et al., 2022); nonetheless, in combination with the experiment-level findings discussed above, the results of the individual-level analysis clearly point toward the importance of contingency awareness as a moderator of implicit evaluation shifts in response to counterattitudinal exemplars.

The sole apparent exception from this pattern is Experiment 4 in which no effect on implicit racial evaluations emerged although participants were explicitly instructed to attend to the valence of the behaviors (valence instruction condition) or even to the fact that all positive behaviors were performed by Black individuals and all negative behaviors were performed by White individuals (contingency instruction condition). However, upon closer inspection, Experiment 4 also attests to the importance of contingency awareness in producing shifts in implicit racial evaluations, in at least two ways. First, rates of contingency awareness in Experiment 4 (41.98%) were considerably lower than in the procedurally comparable Experiments 6 and 7 in which significant malleability in implicit racial evaluations was observed (66.75% and 85.36%, respectively). Second, in line with the overall analysis presented above, implicit evaluations did shift significantly among contingency-aware participants even in Experiment 4 (although, given the low base rate of contingency awareness, no shift was observed in the entire sample).

Of course, these robust findings attesting to the central role of contingency awareness raise the question of why Dasgupta and Greenwald (2001) were able to produce shifts in implicit racial evaluations even in an incidental procedure that did not direct participants' attention to the race–valence contingency in any way and may, in fact, even have directed participants' attention away from this contingency by exposing them to two valence-matched descriptions in conjunction with each target (Gast & Rothermund, 2011). Because no test of contingency awareness was conducted among the original participants, we can only speculate here.

As mentioned in the introduction, media representation of Black Americans has become less biased (Leonard & Robbins, 2021; Shor & van de Rijt, 2023), awareness of anti-Black racism has increased (Barrie, 2020; Reny & Newman, 2021), and societal levels of anti-Black attitudes have decreased (Charlesworth & Banaji, 2019, 2022) since the Dasgupta and Greenwald (2001) experiments were conducted. As such, an experimentally created microcosm in which all Black individuals were positive and all White individuals were negative may have been more unexpected to participants in the social environment of the late 1990s than in the current social environment. Given that expectancy violation is a major driver of learning (Rescorla & Wagner, 1972), it may have been easier for participants to spontaneously become aware of the Black–bad and White–good contingencies then than it is today (see Dasgupta & Asgari, 2004; Dasgupta & Rivera, 2008). Moreover, participants' awareness of the Black–good and White–bad contingencies during exemplar exposure may have been facilitated by the in-person lab context because it encouraged accountability and facilitated attention to the experimental task.

Furthermore, we note that the present results have been obtained in the context of the immediate measurement of implicit evaluations following a short, 5-min intervention involving exemplars that, based on the relevant pretests, were likely highly familiar to most participants. As such, it remains to be seen whether the present (successful) manipulations produced their effects via genuine attitude change (i.e., the alteration of preexisting evaluative representations in long-term memory) or temporary modulations of IAT performance (Ferguson et al., in press; Kurdi & Charlesworth, 2023). Such temporary modulations may have been mediated, for example, by the selective retrieval of information from long-term memory: Exposure to positive Black and negative White exemplars may have facilitated the selective retrieval of positive information about Black individuals and negative information about White individuals. This possibility is all the more likely given that highly familiar (and consequential) social targets often have both positively and negatively valenced information attached to them in long-term memory (e.g., Zayas et al., 2017). Should this be the case, then it is unlikely that the temporary modulations in implicit racial evaluations observed in Experiments 5–7 would persist following a delay. More generally, when and under what conditions short-term malleability in implicit evaluations can translate into long-term attitude change is a major unresolved theoretical and practical question (e.g., Kurdi & Charlesworth, 2023; Lai et al., 2016).

Germane to these considerations, Kurdi and Banaji (2019) found that although the evaluative statements manipulation from Kurdi and Banaji (2017) produced stronger initial shifts in implicit evaluations than repeated evaluative pairings, the effects of the latter were more stable over time than the effects of the former. Moreover, in the context of model-free reinforcement learning, working memory capacity (a crucial component of contingency awareness) predicted updating only in one-shot learning paradigms such as the present one but not when evaluative information was distributed across multiple experimental sessions (Wimmer et al., 2018; Wimmer & Poldrack, 2022). As such, it is possible that, with massive numbers of exposures spaced out over longer periods of time, implicit racial evaluations could shift (and the underlying attitude representations might change durably) in response to counterattitudinal exemplars even under incidental learning conditions, perhaps including the ones present in most Americans' daily lives. However, this possibility has yet to be empirically demonstrated and may prove challenging to demonstrate given the complex logistics involved in long-term learning experiments.

Finally, whether the present effects represent temporary malleability, genuine attitude change, or some combination of both, the precise nature of the cognitive processes mediating them remains to be further investigated. Specifically, the present results seem more challenging to reconcile with early theories of implicit social cognition assuming that implicit evaluations are updated in a purely stimulus-driven manner (e.g., Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004) than with more recent propositional approaches (e.g., De Houwer, 2014), as well as related empirical (e.g., Pleyers et al., 2007) and theoretical perspectives (e.g., Corneille & Stahl, 2019), which have emphasized the role of controlled, attention-dependent processes in evaluative learning. At the same time, it is not clear whether (a) the present results were mediated by propositional inferences and (b) if so, what exact propositional inferences may have played a role.[5] After all, the idea that learning effects can be attention-dependent is part and parcel to some classic associative learning theories outside social psychology (e.g., Mackintosh, 1975; Pearce & Hall, 1980). As such, we hope that future work will further explore these and other mechanistic questions emerging from the present experiments.

Statement of Limitations

We sought to replicate the effect of counterattitudinal exemplars on implicit racial evaluations, as originally established by Dasgupta and Greenwald (2001), and to probe the boundary conditions of this finding. No reduction in implicit pro-White/anti-Black evaluations occurred under the incidental learning conditions created by the original experiments, but we observed shifts in implicit evaluations when race, valence, and the relationship between the two were highlighted to different degrees. Notably, the present experiments were conducted in diverse samples of adults from the United States, which we consider an improvement over the more homogeneous student samples recruited for the original studies. However, the present samples were still not representative, with female, younger, more educated, and liberal participants overrepresented relative to male, older, less educated, and conservative participants. Moreover, the present results should not be expected to generalize to other countries whose macrolevel societal context (including race relations) may be fundamentally different from that of the United States. Finally, the present results were obtained following a short, single-dose intervention that was administered to participants online. The results, and especially the results regarding contingency awareness, may not generalize to in-person interventions, designs involving repeated exposures over time, or measurement of implicit evaluations following a delay.

Concluding Remarks

As scholars of implicit social cognition, we often encounter a particular type of response to our work, sometimes in the form of a friendly question and sometimes in the form of a dismissive comment. The friendly question tends to ask what individuals can do to control unwanted automatic responses to others as a function of their race, age, sexual orientation, and other social identities. The dismissive comment tends to assert that, given their automatic nature, nothing can be done to get unwanted negative implicit evaluations under control, which is claimed to absolve those exhibiting such evaluations from any responsibility for their biased behavior.

We believe that the time is ripe for us to provide two responses to these friendly questions and dismissive comments. First, the evidence now seems incontrovertible that implicit evaluations are malleable in response to a wide range of interventions. Second, in combination with a host of related findings, the present data indicate that implicit evaluations do not merely register co-occurrences present in one's environment but rather are modulated by the way in which one engages with those co-occurrences. In other words, modulation of implicit evaluations is possible—as long as one is able and willing to expend the necessary mental effort. Under what conditions such malleability can be demonstrated to facilitate enduring changes in the underlying attitude representations is a major open question, which we and others are pursuing at present.

Finally, beyond the theoretical and practical implications of the present findings for the possibility of change in racial attitudes, we believe that some implications for the interpretation of failed replication studies may also be worth highlighting. Specifically, authors of replication studies, and especially large-scale replication projects involving many findings and large teams (e.g., Open Science Collaboration, 2015), are often content to report failures of single-shot replication attempts without repeated efforts to identify the boundary conditions under which a particular result may reliably appear and other conditions under which it may disappear. We believe that there is more to be learned scientifically, and a greater contribution to be made to the field, by going beyond one-shot replication attempts and by seeking to understand when and why important findings do and do not replicate. We hope that the manner in which the present replication studies were conducted can serve as a model for future replication efforts in this regard.

[5] A reviewer of this work raised the possibility that demand effects may have, at least in part, mediated the effect in the experiments where shifts in implicit evaluations were observed (see, e.g., Corneille & Béna, 2023; Corneille & Lush, 2023). Although this possibility is not inconceivable, we believe that it is challenging to explain why such effects emerged (a) in some experiments and not in others and (b) only on the IAT and not on self-report measures, although responding on the latter is arguably easier to strategically control than responding on the former.

References

Banaji, M. R. (2004). The opposite of a great truth is also true: Homage of Koan #7. In J. T. Jost, M. R. Banaji, & D. A. Prentice (Eds.), *Perspectivism in social psychology: The yin and yang of scientific progress* (pp. 127–140). American Psychological Association. https://doi.org/10.1037/10750-010

Banaji, M. R., Fiske, S. T., & Massey, D. S. (2021). Systemic racism: Individuals and interactions, institutions and society. *Cognitive Research: Principles and Implications*, *6*(1), Article 82. https://doi.org/10.1186/s41235-021-00349-3

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology*, *32*(1), 104–128. https://doi.org/10.1006/jesp.1996.0005

Barrie, C. (2020). Searching racism after George Floyd. *Socius: Sociological Research for a Dynamic World*, *6*. https://doi.org/10.1177/2378023120971507

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. https://doi.org/10.1207/S15327957PSPR0603_8

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. https://doi.org/10.1037/0022-3514.81.5.828

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., & Banaji, M. R. (2022). Patterns of implicit and explicit attitudes: IV. Change and stability from 2007 to 2020. *Psychological Science*, *33*(9), 1347–1371. https://doi.org/10.1177/09567976221084257

Cone, J., & Calanchini, J. (2021). A process dissociation model of implicit rapid revision in response to diagnostic revelations. *Personality and Social Psychology Bulletin*, *47*(2), 201–215. https://doi.org/10.1177/0146167220919208

Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Corneille, O., & Béna, J. (2023). Instruction-based replication studies raise challenging questions for psychological science. *Collabra: Psychology*, *9*(1), Article 82234. https://doi.org/10.1525/collabra.82234

Corneille, O., & Lush, P. (2023). Sixty years after Orne's *A*merican Psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83–101. https://doi.org/10.1177/10888683221104368

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161–189. https://doi.org/10.1177/1088868318763261

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*(5), 736–760. https://doi.org/10.1521/soco.2007.25.5.736

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658. https://doi.org/10.1016/j.jesp.2004.02.003

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814. https://doi.org/10.1037/0022-3514.81.5.800

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, *26*(1), 112–123. https://doi.org/10.1521/soco.2008.26.1.112

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

De Houwer, J., & Vandorpe, S. (2010). Using the Implicit Association Test as a measure of causal learning does not eliminate effects of rule learning. *Experimental Psychology*, *57*(1), 61–67. https://doi.org/10.1027/1618-3169/a000008

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2009). Commonality and the complexity of "we": Social attitudes and social change. *Personality and Social Psychology Review*, *13*(1), 3–20. https://doi.org/10.1177/1088868308326751

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), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science*, *28*(4), 331–336. https://doi.org/10.1177/0963721419835206

Ferguson, M. J., Shen, X., Cone, J., & Mann, T. C. (in press). How do we reduce implicit bias toward outgroups? In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit bias and racism*. Cambridge University Press.

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. https://doi.org/10.1016/j.lmot.2013.03.003

Gast, A., & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion*, *25*(1), 89–110. https://doi.org/10.1080/02699931003696380

Gawronski, B., & Walther, E. (2012). What do memory data tell us about the role of contingency awareness in evaluative conditioning? *Journal of Experimental Social Psychology*, *48*(3), 617–623. https://doi.org/10.1016/j.jesp.2012.01.002

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., Brock, T. C., & Ostrom, T. M. (Eds.). (1968). *Psychological foundations of attitudes*. Academic Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behaviour. *Nature Human Behaviour*, *5*(10), 1391–1401. https://doi.org/10.1038/s41562-021-01213-6

Hewstone, M., & Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology*, *36*(2), 103–124. https://doi.org/10.1006/jesp.1999.1398

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Hütter, M., & De Houwer, J. (2017). Examining the contributions of memory-dependent and memory-independent components to evaluative conditioning via instructions. *Journal of Experimental Social Psychology*, *71*, 49–58. https://doi.org/10.1016/j.jesp.2017.02.007

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, *41*(3), 137–146. https://doi.org/10.1027/1864-9335/a000020

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*(4), 589–617. https://doi.org/10.1177/000312240607100404

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 774–788. https://doi.org/10.1037/0022-3514.81.5.774

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

Kraus, M. W., Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2019). The misperception of racial economic inequality. *Perspectives on Psychological Science*, *14*(6), 899–921. https://doi.org/10.1177/1745691619863049

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*(4), 565–579. https://doi.org/10.1037/0022-3514.68.4.565

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. https://doi.org/10.1037/xge0000239

Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, *116*(5), 681–703. https://doi.org/10.1037/pspa0000151

Kurdi, B., & Charlesworth, T. E. S. (2023). A 3D framework of implicit attitude change. *Trends in Cognitive Sciences*, *27*(8), 745–758. https://doi.org/10.1016/j.tics.2023.05.009

Kurdi, B., & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition*, *38*(Suppl.), s42–s67. https://doi.org/10.1521/soco.2020.38.supp.s42

Kurdi, B., Hussey, I., Stahl, C., Hughes, S., Unkelbach, C., Ferguson, M. J., & Corneille, O. (2022). Unaware attitude formation in the surveillance task? Revisiting the findings of Moran et al. (2021). *International Review of Social Psychology*, *35*(1), Article 6. https://doi.org/10.5334/irsp.546

Kurdi, B., Krosch, A. R., & Ferguson, M. J. (2023). Oppressed groups engender implicit positivity: Seven demonstrations using novel and familiar targets. *Psychological Science*, *34*(10), 1069–1086. https://doi.org/10.1177/09567976231194588

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2023). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology*, *124*(6), 1174–1202. https://doi.org/10.1037/pspa0000329

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Leonard, D. J., & Robbins, S. T. (Eds.). (2021). *Race in American television: Voices and visions that shaped a nation*. Greenwood.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*(5), 842–855. https://doi.org/10.1037/0022-3514.81.5.842

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. https://doi.org/10.3758/s13428-014-0532-5

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

Marini, M., Rubichi, S., & Sartori, G. (2012). The role of self-involvement in shifting IAT effects. *Experimental Psychology*, *59*(6), 348–354. https://doi.org/10.1027/1618-3169/a000163

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin*, *36*(4), 512–523. https://doi.org/10.1177/0146167210362789

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K., Balas, R., Benedict, T., Corneille, O., Douglas, S. B., Ferguson, M. J., Fritzlen, K. A., Gast, A., Gawronski, B., Giménez-Fernández, T., Hanusz, K., Heycke, T., Högden, F., … De Houwer, J. (2021). Incidental attitude formation via the surveillance task: A preregistered replication of the Olson and Fazio (2001) study. *Psychological Science*, *32*(1), 120–131. https://doi.org/10.1177/0956797620968526

Morehouse, K. N., & Banaji, M. R. (in press). The science of implicit race bias: Evidence from the Implicit Association Test. *Daedalus*.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), 421–433. https://doi.org/10.1177/0146167205284004

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 130–144. https://doi.org/10.1037/0278-7393.33.1.130

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, *19*(3), 249–254. https://doi.org/10.1111/j.1467-9280.2008.02076.x

Ratliff, K. A., Lofaro, N., Howell, J. L., Conway, M. A., Lai, C. K., O'Shea, B., Smith, C. T., Jiang, C., Redford, L., Pogge, G., Umansky, E., Vitiello, C. A., & Zitelny, H. (2020). *Documenting bias from 2007–2015: Pervasiveness and correlates of implicit attitudes and stereotypes II*. PsyArXiv. https://osf.io/jeyc7

Reny, T. T., & Newman, B. J. (2021). The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests. *The American Political Science Review*, *115*(4), 1499–1507. https://doi.org/10.1017/S0003055421000460

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Shor, E., & van de Rijt, A. (2023). Racial bias in media coverage: Accounting for structural position and public interest. *European Sociological Review*. Advance online publication. https://doi.org/10.1093/esr/jcad031

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583–592. https://doi.org/10.1037/0022-3514.89.4.583

Skinner-Dorkenoo, A. L., George, M., Wages, J. E., III, Sánchez, S., & Perry, S. P. (2023). A systemic approach to the psychology of racial bias within individuals and society. *Nature Reviews Psychology*, *2*, 392–406. https://doi.org/10.1038/s44159-023-00190-z

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01

Solié, C., Girard, B., Righetti, B., Tapparel, M., & Bellone, C. (2022). VTA dopamine neuron activity encodes social interaction and promotes reinforcement learning through social prediction error. *Nature Neuroscience*, *25*(1), 86–97. https://doi.org/10.1038/s41593-021-00972-9

Staats, A. W., Staats, C. K., & Heard, W. G. (1959). Language conditioning of meaning to meaning using a semantic generalization paradigm. *Journal of Experimental Psychology*, *57*(3), 187–192. https://doi.org/10.1037/h0042274

Stahl, C., & Unkelbach, C. (2009). Evaluative learning with single versus multiple unconditioned stimuli: The role of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 286–291. https://doi.org/10.1037/a0013255

Stahl, C., Unkelbach, C., & Corneille, O. (2009). On the respective contributions of awareness of unconditioned stimulus valence and unconditioned stimulus identity in attitude formation through evaluative conditioning. *Journal of Personality and Social Psychology*, *97*(3), 404–420. https://doi.org/10.1037/a0016196

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Turner, R. N., & Crisp, R. J. (2010). Imagining intergroup contact reduces implicit prejudice. *British Journal of Social Psychology*, *49*(1), 129–142. https://doi.org/10.1348/014466609X419901

Wimmer, G. E., Li, J. K., Gorgolewski, K. J., & Poldrack, R. A. (2018). Reward learning over weeks versus minutes increases the neural representation of value in the human brain. *The Journal of Neuroscience*, *38*(35), 7649–7666. https://doi.org/10.1523/JNEUROSCI.0075-18.2018

Wimmer, G. E., & Poldrack, R. A. (2022). Reward learning and working memory: Effects of massed versus spaced training and post-learning delay period. *Memory & Cognition*, *50*(2), 312–324. https://doi.org/10.3758/s13421-021-01233-7

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), 815–827. https://doi.org/10.1037/0022-3514.81.5.815

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *67*(11), 2105–2122. https://doi.org/10.1080/17470218.2014.907324

Zayas, V., Surenkok, G., & Pandey, G. (2017). Implicit ambivalence of significant others: Significant others trigger positive and negative evaluations. *Social and Personality Psychology Compass*, *11*(11), e12360. https://doi.org/10.1111/spc3.12360

Received June 25, 2023
Revision received September 14, 2023
Accepted September 29, 2023 ■