23/04/2024 23:13

Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time

August 16, 2016

Calvin K. Lai,[1] Allison L. Skinner,[2] Erin Cooley,[3] Sohad Murrar,[4] Markus Brauer,[4] Thierry Devos,[5] Jimmy Calanchini,[6] Y. Jenny Xiao,[7] Christina Pedram,[8] Christopher K. Marshburn,[8] Stefanie Simon,[9] John C. Blanchar,[10] Jennifer A. Joy-Gaba,[11] John Conway,[12] Liz Redford,[12] Rick A. Klein,[12] Gina Roussos,[13] Fabian M. H. Schellhaas,[13] Mason Burns,[14] Xiaoqing Hu,[15] Meghan C. McLean,[16] Jordan R. Axt,[17] Shaki Asgari,[18] Kathleen Schmidt,[19] Rachel Rubinstein,[16] Maddalena Marini,[1] Sandro Rubichi,[20] Jiyun-Elizabeth L. Shin,[21] & Brian A. Nosek[17, 22]

[1]Harvard University
[2]University of Washington
[3]Colgate University
[4]University of Wisconsin - Madison
[5]San Diego State University
[6]University of California, Davis
[7]New York University
[8]University of California, Irvine
[9]Carleton College
[10]University of Arkansas
[11]Virginia Commonwealth University
[12]University of Florida
[13]Yale University
[14]Purdue University
[15]University of Texas at Austin
[16]Rutgers University - New Brunswick
[17]University of Virginia
[18]Measurement Incorporated
[19]Wesleyan University
[20]University of Modena and Reggio Emilia
[21]Stony Brook University
[22]Center for Open Science

Correspondence should be addressed to: Calvin Lai, Department of Psychology, William James Hall, 33 Kirkland Street, Harvard University, Cambridge, MA, 02138. Email: cklai4@gmail.com.

Word count: 10,314 (main text + abstract).

Authors' note: This project was supported by a National Science Foundation Graduate Research Fellowship and Project Implicit, Inc. Lai and Axt are consultants and Nosek is an officer of Project Implicit, Inc., a non-profit organization that includes in its mission "To develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors." We thank the authors from RIRP:I for their contributions to the development of the interventions employed in this article and John Mitterer for his assistance with data collection for Study 1. Contributions: Conceived research: Lai & Nosek; Designed research: Lai & Nosek; Performed research: Lai and all authors: Analyzed data: Lai; Wrote paper: Lai & Nosek.; Revised paper: all authors. All study materials, data, and analyses are available at https://osf.io/um4ye/.

Abstract

Implicit preferences are malleable, but does that change last? We tested nine interventions (eight real and one sham) to reduce implicit racial preferences over time. In two studies with a total of 6,321 participants, all nine interventions immediately reduced implicit preferences. However, none were effective after a delay of several hours to several days. We also found that these interventions did not change explicit racial preferences and were not reliably moderated by motivations to respond without prejudice. Short-term malleability in implicit preferences does not necessarily lead to long-term change, raising new questions about the flexibility and stability of implicit preferences.

Word Count: 100

Keywords: attitudes, racial prejudice, implicit social cognition, malleability, Implicit Association Test

Full Citation

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., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein., R, Marini, M., Rubichi, S., Shin,. J. L., & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001-1016.

Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time

Early theories of implicit social cognition suggested that implicit associations were largely stable. These claims were supported by evidence that changes in conscious belief did not lead to corresponding changes in implicit associations (e.g., Devine, 1989; Shiffrin & Schneider, 1977; Wilson, Lindsey, & Schooler, 2000). The psychologist John Bargh referred to the stability of implicit cognitions as the "cognitive monster": "Once a stereotype is so entrenched that it becomes activated automatically, there is really little that can be done to control its influence" (p. 378, Bargh, 1999). This dominant view has changed over the past fifteen years to one of implicit malleability, with many studies finding that implicit associations are sensitive to lab-based interventions (for reviews, see Blair, 2002; Gawronski & Bodenhausen 2006; Lai, Hoffman, & Nosek, 2013). These interventions vary greatly in approach. In one, for example, participants are exposed to images of people who defy stereotypes (e.g., admired Black people / hated White people; Dasgupta & Greenwald, 2001; Joy-Gaba & Nosek, 2010). In another, participants are given goals to override implicit biases (e.g., Mendoza, Gollwitzer, & Amodio, 2010; Stewart & Payne, 2008).

In most of the research on implicit association change, the short-term malleability of associations is tested by administering an implicit measure immediately after the intervention. Studies examining long-term change in implicit associations are rare. In a meta-analysis on experiments to change implicit associations (Forscher, Lai, et al., 2016), only 22 (3.7%) of 585 studies examined whether change in implicit associations persisted beyond a single session. The studies do not provide a firm basis for knowing when lasting change will or will not happen. Of the 22 experiments, 9 (40.1%) studies found significant evidence of lasting change (e.g., Vezzalli, Capozza, Giovannini, & Stathi, 2011; Olson & Fazio, 2006), 7 (31.8%) studies did not find significant evidence (e.g., Jang & Kim, 2011; Thomas, Judge, Brownell, & Vartanian, 2006), and 6 (27.2%) studies found mixed evidence (e.g., O'Brien, Puhl, Latner, Mir, & Hunter, 2010; Sportel, de Hullu, de Jong, & Nauta, 2013). As such, cumulative knowledge about the mechanisms and conditions necessary for changing implicit associations is only beginning to develop. The central interest of the present article is to systematically examine when short-term malleability in implicit associations translates into persisting change.

Comparative Approaches to Intervention Research

A standard model of intervention research is to isolate mechanisms in order to study how those mechanisms work. However, an exclusive focus on isolating mechanisms within interventions can impede progress. A complementary strategy takes a comparative approach by examining many interventions simultaneously. This strategy can reveal differences in effectiveness that would otherwise be difficult to uncover when testing interventions in isolation. Once revealed, mechanism-focused research can unpack the causes underlying effective interventions.

Driven by a lack of comparative work on implicit bias reduction approaches, Lai and colleagues (2014) experimentally compared the effects of 17 interventions and one sham intervention on implicit racial preferences in Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 interventions (RIRP:I). Relative to a control condition, nine of

the 18 interventions were effective at reducing implicit biases when assessed immediately following administration of the intervention. The effective interventions varied widely in design and hypothesized mechanism. Effective interventions in RIRP:I tended to be highly self-relevant, emotionally evocative, and either gave experiences with positive Black exemplars and negative White exemplars or concrete strategies to override bias. Interventions that were ineffective tended to induce reflection on egalitarian values or encourage taking the perspective of Black individuals.

Overview

We conducted two large-scale confirmatory experiments to examine the durability of implicit bias reduction effects from all nine effective interventions in RIRP:I. Five interventions gave participants experiences with counterstereotypical exemplars, one intervention primed multicultural ideology, two interventions employed evaluative conditioning, and two interventions gave intentional strategies to overcome bias. In Study 1, we investigated intervention effectiveness on implicit and explicit racial preferences immediately and after a delay of several hours to several days in a sample of 1,021 North American students from two universities. We also assessed students' support for affirmative action policies to examine whether changes in racial preferences transferred to changes in racially-relevant political preferences. In Study 2, we tested the nine interventions again with a shorter delay between sessions and a larger sample of 5,295 participants from 17 American universities. These findings provide new insight into the durability of implicit bias change, establishing a new frontier for understanding the conditions under which shifts in implicit preferences reflect short-term malleability or longer-term change.

Study 1

Method

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in this article. All materials and supplemental analyses are available here: https://osf.io/um4ye/. This study's analysis plan was pre-registered before data collection at https://osf.io/zeupk/. A version of this study's design and analysis plan was peer reviewed by the editor and ad hoc reviewers at this journal (https://osf.io/kztme/).

Participants

Participants were non-Black undergraduates (83.3% White, 73.7% female, median age of 18) from Brock University and the University of Virginia. Our plan for determining sample size was to collect as many participants as we could in the Fall 2013 semester. 1192 participants from Brock University and 159 participants from the University of Virginia began the study at Time 1 (T1). Of those 1391 participants, 261 (18.7%) were excluded because they did not finish T1 or took T1 multiple times, 53 (3.8%) because they identified as Black or White/Black multi-racial, 13 (.9%) because they chose not to report their racial identity, 29 (2.0%) because they responded too quickly or made too many errors on the implicit measure (see Dependent Measures section for more detail), and 14 (1.0%) because they accessed the second session before the first session. This left a final sample of 1021 participants who completed T1, of which 872 (85.4%) also

completed Time 2 sessions (T2). In terms of statistical power to detect an effect size of Cohen's $d = .32$ (the average effect size of the effective interventions from RIRP:I) for each individual effect, we had 38% power to detect a reduction against control at T1 at $p < .01$ and 31% power to detect a reduction against control at T2 at $p < .01$.

Procedure

Participants were shown a link to the study (delivered via email at Brock University and via the participant pool website at the University of Virginia) and instructed to complete it online. Two-thirds of participants were randomly assigned to begin the study by taking a pretest Race Implicit Association Test (IAT) and one-third were assigned to take nothing at all. This was done to allow for analysis of within-subjects change and analysis of unique effects from taking a pretest (Solomon, 1949). Participants were then randomly assigned to one of nine intervention conditions or a control condition with no intervention. As the final part of the first session, they took a posttest Race IAT and a measure of explicit racial prejudice. Participants at Brock University also completed a demographics questionnaire (University of Virginia participants' demographic data came from a research pool prescreen questionnaire). Procedurally this session was similar to Study 4 in RIRP:I.

Between two and four days after T1 (and with reminders after 2 or 3 days), participants were emailed a link for T2.[1] On average, participants returned for the second session after 3.28 days ($SD = 1.97$ days). In that session, they completed the Race IAT, two items assessing support for pro-Black affirmative action, a measure of explicit racial prejudice, and an item assessing their effort in the study. See Figure 1 for a schematic of the procedure.

[1] Due to an error, participants at Brock University were emailed a link to the second session at the same time as the first one. When examining the overall sample, 14 participants took the second session before the first, 91 participants took the second session in under an hour after the first session, 26 participants took it between 1 - 24 hours, 85 took it between 1 and 2 days, 359 took it between 2 and 4 days, and 312 took it 4 days after or later. The 14 participants who took the second session before the first were excluded from all analyses and the rest of the participants were included in all analyses.

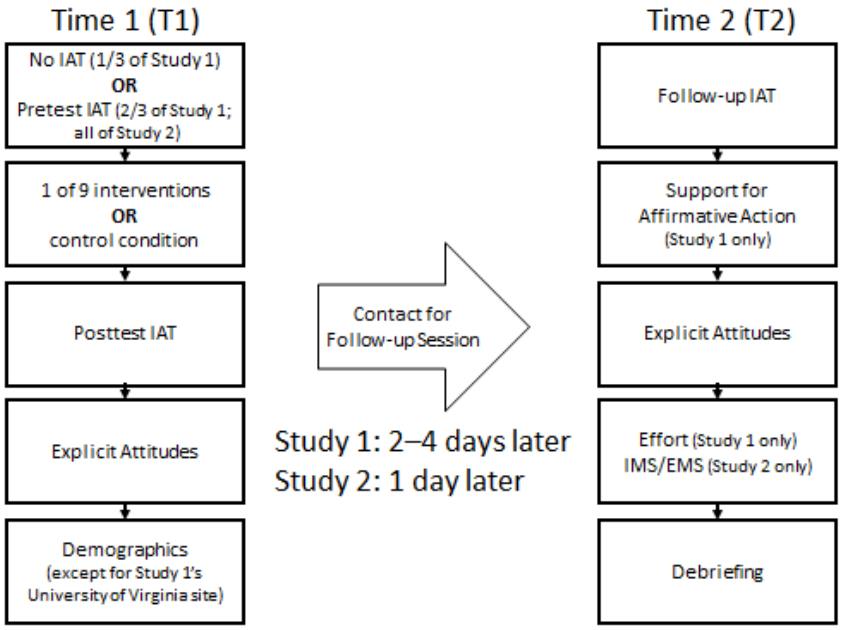

Figure 1. Procedure for Studies 1 and 2.

Dependent Measures

Implicit Association Test (IAT). The IAT assesses the relative strength of associations between two concepts (i.e., White people, Black people) and two attributes (i.e., Good, Bad; see Nosek, Greenwald, & Banaji, 2007, for a review). It does so by comparing how quickly participants respond when one set of concepts/attributes are paired together (e.g., White people + Good / Black people + Bad) with how quickly they respond when another set of concepts/attributes are paired together (e.g., White people + Bad / Black people + Good). Table 1 describes the block structure and method-related randomization (i.e., for order and practice effects) of the IAT. The procedure followed the recommendations of Nosek, Greenwald, and Banaji (2005) but with five blocks instead of seven and fewer trials for each block (16, 24, or 32 trials instead of 20 or 40 trials) to reduce the total time required. Participants were instructed to categorize words and images as quickly and accurately as possible. The IAT was scored with the D2 algorithm recommended by Greenwald, Banaji, and Nosek (2003). A positive *D* score indicates faster responding when White faces were paired with good words and Black faces were paired with bad words compared to the reverse. Positive scores are interpreted as an implicit preference for White people relative to Black people.[2] Participants were excluded from all analyses if (a) more than 10% of critical trials were faster than 300 ms across all IATs they completed, (b) if the error rate was higher than 30% across all IATs, (c) if more than 25% of trials were faster than 300 ms in any critical block in any IAT, or (d) if the error rate was higher than 50% in any critical block in any IAT. We excluded 29 (2.0%) participants in Study 1 and 101 (1.6%) participants in Study

[2] By 'implicit preferences' we mean indirectly assessed preferences, in contrast to explicit or directly assessed preferences. Evidence for malleability or change in IAT scores do not guarantee a change in associations because measures are influenced by additional influences (Calanchini & Sherman, 2013; Nosek et al., 2007)

2 for fulfilling these exclusion criteria. Participant exclusion rates did not differ by condition in Study 1, $\chi^2(18, N = 1050) = 21.86$, $p = .24$, or in Study 2, $\chi^2(18, N = 5396) = 16.78$, $p = .54$.

Table 1
Block Sequence in the Race Implicit Association Test (IAT)

Block	No. of trials	Function	Categories assigned to left-key response	Categories assigned to right-key response
1	16	Practice	Black people	White people
2	16	Practice	Bad	Good
3	32	Critical	Black people + Bad	White people + Good
4	24	Practice	White people	Black people
5	32	Critical	White people + Bad	Black people + Good

Note. The combined trial blocks (Blocks 3 and 5) alternated between trials that present good/bad stimuli and trials that present White people/Black people stimuli. The categories in Block 1/3 were counterbalanced with Block 4/5 to address the possibility of order effects (Greenwald et al., 1998). We also sought to reduce practice effects by randomizing features that are not important to the measure. Both studies contained two variations of the IAT: one variant had a black background and Good/Bad as evaluative categories; the other variant had a white background, Pleasant/Unpleasant as evaluative categories, and a different set of image/word stimuli. These IAT variants were counterbalanced at Time 1 and fully randomized at Time 2.

Explicit racial preferences. Participants completed three self-report items measuring racial prejudice. One assessed relative preference for White people compared to Black people on a 7-point Likert scale ranging from “I strongly prefer Black people to White people” to “I strongly prefer White people to Black people.” The other two items were feeling thermometers rating warmth for White people and Black people on a 7-point scale ranging from “Very cold” to “Very warm.” For analyses, a difference score was computed between the two feeling thermometers and averaged with the racial preference measure after standardizing each ($SD = 1$) while retaining their rational zero points of no preference between White people and Black people. More positive scores indicated a greater explicit preference for White people over Black people.

Support for affirmative action. In the second session, participants completed two self-report items measuring support for affirmative action by responding no (0) or yes (1).[3] One item assessed support for affirmative action in corporate settings: "A corporate personnel officer is evaluating a Black job applicant and a White job applicant who are identically qualified except the White applicant has more prior experience in related work. Is there a reasonable justification for this personnel officer hiring the Black applicant rather than the White applicant?" The other

[3] As a reviewer noted, these items did not sufficiently capture the dynamics of affirmative action in North America (for examples of more thorough assessments, see Federico & Sidanius, 2002; Haley & Sidanius, 2006). We include analyses of these items in this manuscript but caution against overgeneralization of the findings due to their limited scope.

assessed support for affirmative action in higher education: "A college admissions officer considers applications from Black applicants and White applicants with similar credentials and cannot accept all of them. Should the admissions officer more often accept Black applicants than White applicants?"

Effort on study. Participants completed two 5-point self-report items assessing their effort and motivations for taking the study. The items were "Did you care about your performance in the study?" and "What level of effort did you put forth in the study?" The first item had the response options "Not at all", "Slightly", "Somewhat", "Very much", and "A great deal". The second item had the response options "No effort", "Slight effort", "Moderate effort", "Strong effort", and "Extreme effort". On average, participants reported between moderate and strong effort (M = 3.66, SD = .72).

Interventions

RIRP:I identified nine interventions that shifted IAT scores immediately following the intervention. For Study 1, we included all nine with no or minor revisions plus a no-intervention control condition (for more information about the development of these interventions, see Lai et al., 2014). Next, we describe each of these nine interventions in four categories: Exposure to Counterstereotypical Exemplars, Appeals to Egalitarian Values, Evaluative Conditioning, and Intentional Strategies to Overcome Biases. Details of the intervention procedures and links to self-administer the procedures are available at: https://osf.io/vk24l/.

Exposure to Counterstereotypical Exemplars

Four interventions were designed to reduce IAT scores through experiences with positive Black exemplars and negative White exemplars: Vivid Counterstereotypic Scenario, Practicing an IAT with Counterstereotypical Exemplars, Shifting Group Boundaries Through Competition, and Shifting Group Affiliations Under Threat.

Vivid Counterstereotypic Scenario. Participants in this intervention read a vivid second-person story in which they are the protagonist. The participant imagines walking down a street late at night after drinking at a bar. Suddenly, a White man in his forties assaults the participant, throws him/her into the trunk of his car, and drives away. After some time, the White man opens the trunk and assaults the participant again. A young Black man notices the second assault and knocks out the White assailant, saving the day. After reading the story, participants are told the next task (i.e., the race IAT) was supposed to affirm the associations: White = Bad, Black = Good. Participants were instructed to keep the story in mind during the IAT. This intervention employs extreme counterstereotypes - a White villain and a Black hero (Dasgupta & Greenwald, 2001). It is also self-relevant: participants imagine themselves in the situation. A study that compared second and third-person perspectives with a variant of this story found self-relevance to be essential for effectiveness (Marini, Rubichi, & Sartori, 2012). Lastly, the content and style was emotionally involving (Rudman, 2004). In RIRP: I, this intervention was the most effective out of 17 tested, d = .49, 95% CI [.41, .58].

Practicing an IAT with Counterstereotypical Exemplars. Exposure to counterstereotypical Black and White exemplars can shift implicit racial preferences (Dasgupta & Greenwald, 2001;

Joy-Gaba & Nosek, 2010). We employed a variation of the IAT procedure to reinforce positive associations with Blacks and negative associations with Whites. Participants completed 20 practice trials, followed by the combined blocks of the race IAT that paired Black with Good and White with Bad (32 trials). The stimulus items representing Blacks and Whites were the same as those used in the race IAT, plus six famous positive Black exemplars (e.g., Oprah Winfrey) and six infamous negative White exemplars (e.g., Adolf Hitler). Before the IAT practice, participants were shown pictures of each of these exemplars along with brief one-line descriptions of what they are known for. Study 1 included some negative White exemplars that participants may not know/remember (i.e., John Gotti, Timothy McVeigh, Charles Manson, Ted Bundy), but participants were reminded of their notorious behavior. In Study 2, we replaced those negative exemplars with more recent exemplars (i.e., Bernie Madoff, Anders Breivik, Jared Loughner, Jerry Sandusky) and similar reminders of their notorious behavior. This intervention was the third-most effective in RIRP: I, d = .40, 95% CI [.30, .49].

Shifting Group Boundaries Through Competition. Participants played in a simulated dodgeball game in which their teammates were Black and their opponents were White. The Black teammates saved the participants from being knocked out and were good sports, whereas the opposing all-White team engaged in unfair play and were bad sports. At the end of the intervention, participants were instructed to make intentions to think "Black = Good" and "White = Bad" and to remember how their Black teammates helped them and their White enemies hurt them during the IAT. This intervention was motivated by evidence that intense competition and strong outgroup threats lead to negative outgroup attitudes (Riek, Mania, & Gaertner, 2006). We expected that flipping the script, (i.e., by cooperating with Black outgroup members to compete against White ingroup members) would produce the opposite effect: more positive outgroup attitudes and reduced ingroup favoritism. This intervention was the second-most effective out of the nine effective interventions in RIRP: I, d = .45, 95% CI [.36, .55].

Shifting Group Affiliations Under Threat. Participants read a vivid and threatening post-nuclear war scenario. They were then shown profiles of people described as "close friends" in their camp, all of whom were Black and had helpful survival skills (e.g. a doctor who worked with Doctors Without Borders). They also viewed profiles of "terrible enemies" that were all villainous White people who plotted to destroy their camp. After reading those profiles, participants were told to "Please imagine and think about the friends and enemies you just read about while you complete these tasks." The rationale for this intervention was similar to Shifting Group Boundaries Through Competition. Outgroup threats lead to more negative outgroup attitudes (Riek et al., 2006), so flipping the group memberships may have the opposite effect. This intervention was the eighth-most effective in RIRP: I, d = .28, 95% CI [.18, .37].

Appeals to Egalitarian Values

In RIRP: I, five interventions attempted to reduce implicit bias by appealing to deeply-held egalitarian values. Of these, one was successful: Priming Multiculturalism.

Priming Multiculturalism. To improve intergroup relations, some endorse multiculturalism - the idea that racial/ethnic differences should be appreciated and celebrated. Experimental evidence suggests that considering multiculturalism reduces implicit racial preferences relative to

considering colorblindness - the idea that racial/ethnic differences should be ignored (Richeson & Nussbaum, 2004; Wolsko, Park, Judd, & Wittenbrink, 2000). This intervention examined the effect of multiculturalism on racial preferences by encouraging participants to adopt a multicultural perspective. Following Richeson and Nussbaum (2004), participants read a prompt advocating multiculturalism, summarized the prompt in their own words, and then listed one reason why multiculturalism "is a positive approach for improving relationships between groups." Finally, participants were given instructions to think "Black = Good" as they took the IAT. Priming Multiculturalism was the seventh-most effective intervention in RIRP: I, d = .29, 95% CI [.17, .40].

Evaluative Conditioning

Repeatedly pairing attitude objects (e.g., Black/White faces) with other valenced stimuli (e.g., positive/negative words) can alter implicit associations (De Houwer, Thomas, & Baeyens, 2001; Olson & Fazio, 2001, 2002, 2006). Two variations of evaluative conditioning were included: Evaluative Conditioning and Evaluative Conditioning with the GNAT.

Evaluative Conditioning. Participants viewed 20 Black faces paired with positive words and 20 White faces paired with negative words. On each trial, participants saw a pairing for one second and categorized the face as "Black" or "White". They were also instructed to memorize the positive/negative word for later testing. After the task, participants recalled as many of the positive/negative words as possible. This intervention was the ninth-most effective in RIRP: I, d = .21, 95% CI [.12, .30].

Evaluative Conditioning with the GNAT. Participants completed a Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001) modified to condition new associations. Participants saw Black faces paired with good or bad words. They pressed the spacebar (i.e., 'Go') when a Black face was paired with a good word and made no response (i.e., 'No-Go) when a Black face was paired with a bad word. They were also instructed to count the number of times they saw a Black person and a good word paired together. A majority of the trials (46 out of 80) were 'Go' trials (i.e., Black faces paired with good words). Afterward, participants reported how many Black/good pairings they counted. This intervention was the sixth-most effective in RIRP: I, d = .32, 95% CI [.24, .41].

Intentional Strategies to Overcome Biases

Performance on implicit measures can be altered via strategies to override implicit bias. Two interventions gave participants strategies to alter the expression of implicit associations: Using Implementation Intentions and Faking the IAT. These interventions differ in that Using Implementation Intentions provides a strategy to alter the expression of implicit biases themselves, whereas Faking the IAT gives participants strategies to subvert the procedure, which presumably does not have an effect on actual implicit associations. The latter is a sham intervention for comparative purposes.

Using Implementation Intentions. Making desired behaviors more accessible and automatic is an effective approach for aligning intentions with behavior (Stewart & Payne, 2008). A popular method for doing so is implementation intentions: if-then plans that tie a behavioral response to a

situational cue (Gollwitzer, 1999). Participants learned about the tendency for people to exhibit implicit biases for Whites over Blacks, then were told that they could overcome that bias by committing themselves to an implementation intention by saying to themselves silently, "If I see a Black face, then I will respond by thinking 'good.'" In Study 1, participants also took an abbreviated IAT at the beginning of the intervention to familiarize themselves with the task. Implementation Intentions was the fifth-most effective intervention in RIRP: I, d = .38, 95% CI [.30, .47].

Faking the IAT. The IAT is resistant to naive fakers (Banse, Seise, & Zerbes, 2001; Kim, 2003), but is susceptible to faking when given concrete instructions or experience with the IAT (Fiedler & Bluemke, 2005; Steffens, 2004). As a comparison to the "true" interventions, participants completed an adapted version of a faking manipulation from Cvencek and colleagues (2010). Participants first learned about the tendency for people to exhibit implicit biases for Whites over Blacks. Then, they were told to alter their responses on the IAT by slowing down when "Black and Bad" are paired together and speeding up when "White and Bad" are paired together. In Study 1, participants also took an abbreviated IAT at the beginning of the intervention to familiarize themselves with it. Faking the IAT was the fourth-most effective intervention in RIRP: I, d = .39, 95% CI [.31, .47].

Results

For a complete description of our pre-registered analysis plan (and deviations from that plan), see https://osf.io/zeupk/. Most analyses in this section were conducted with and without data collection site as a covariate. The pattern of results did not change for any analysis due to the inclusion of this covariate and we report only the versions with the site covariate in this section. Analyses without site as a covariate and other supplemental analyses (e.g., analyses using listwise deletion) that are not reported in the main text are available at https://osf.io/um4ye/. Due to the number of analyses we computed for this section, we set our alpha criterion as p = .01 instead of the conventional p = .05.

Implicit Racial Preferences

Participants completed two or three IATs over the course of two sessions. Overall, participants had IAT scores preferring Whites over Blacks at pretest, posttest, and follow-up assessments (Ns = 670, 1016, 866; Ms = .64, .37, .48; SDs = .40, .48, .40; ds = 1.61, .77, 1.21). These IAT scores were positively, but not strongly, correlated ($r_{\text{pretest-posttest}}(666) = .22$, $r_{\text{pretest-follow-up}}(561) = .22$, $r_{\text{posttest-follow-up}}(859) = .30$). The relatively weaker correlations compared to prior research (Nosek et al., 2007; Bar-Anan & Nosek, 2014) could be attributable to differential sensitivity to the interventions or using a shortened version of the IAT (5-block instead of 7-block).

IAT scores were moderated by time of assessment, $F(1, 864) = 30.67$, $p < .001$, $\eta^2_p = .03$, condition, $F(9, 850) = 3.07$, $p = .001$, $\eta^2_p = .03$, data collection site, $F(1, 850) = 9.11$, $p = .003$, $\eta^2_p = .01$, and an interaction between time and condition, $F(9, 850) = 7.44$, $p < .001$, $\eta^2_p = .07$, but not by an interaction between time and site, $F(1, 850) = 3.09$, $p = .079$, $\eta^2_p = .00$. Follow-up analyses found that condition had significant effects on IAT scores (controlling for site) at

posttest, $F(9, 1005) = 8.05$, $p < .001$, $\eta^2_p = .07$, but not at pretest, $F(9, 659) = .77$, $p = .64$, $\eta^2_p = .01$, or at follow-up, $F(9, 855) = 2.28$, $p = .016$, $\eta^2_p = .02$. See Table 2 for a summary of implicit preferences by condition.

Table 2
Implicit Racial Preferences (Study 1)

	Pretest			Posttest				Follow-up			
Condition	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*N*	*M*	*SD*	*d*
Control	62	.60	.40	102	.54	.40		80	.44	.38	
Exposure to Counterstereotypical Exemplars											
Vivid Counterstereotypic Scenario	71	.61	.42	100	.25	.47	.67***	91	.53	.36	-.24
Practicing an IAT with Counterstereotypical Exemplars	75	.63	.49	111	.36	.44	.43**	94	.48	.40	-.10
Shifting Group Boundaries Through Competition	80	.61	.33	114	.43	.47	.26	96	.44	.41	.01
Shifting Group Affiliations Under Threat	73	.64	.38	104	.39	.49	.35*	91	.58	.42	-.34*
Appeals to Egalitarian Values											
Priming Multiculturalism	59	.77*	.32	88	.44	.41	.24*	73	.56	.33	-.32
Evaluative Conditioning											
Evaluative Conditioning	66	.64	.46	97	.50	.38	.11	81	.46	.40	-.06
Evaluative Conditioning with the GNAT	70	.62	.35	108	.35	.44	.44**	92	.36	.42	.20
Intentional Strategies to Overcome Biases											
Using Implementation Intentions	62	.66	.35	103	.36	.42	.43**	92	.46	.42	-.04
Faking the IAT	52	.66	.46	89	.05	.67	.90***	76	.50	.38	-.16

Note. Descriptive statistics reflect *D* scores (Greenwald et al., 2003), and positive values indicate greater preference for White people compared to Black people. *d* = Cohen's *d* effect size reflecting change in preference relative to control. Significance is from a *t*-test contrasting an experimental condition against the control condition. * $p < .05$. ** $p < .01$. *** $p < .001$.

Five of the nine interventions significantly reduced IAT scores relative to the control condition at posttest controlling for site: Vivid Counterstereotypic Scenario, $F(1, 199) = 21.38$, $p < .001$, $\eta^2_p = .10$, Practicing an IAT with Counterstereotypic Exemplars, $F(1, 210) = 8.69$, $p = .004$, $\eta^2_p = .04$, Evaluative Conditioning with the GNAT, $F(1, 207) = 9.69$, $p = .002$, $\eta^2_p = .05$, Using Implementation Intentions, $F(1, 202) = 8.47$, $p = .004$, $\eta^2_p = .04$, and Faking the IAT, $F(1, 188) = 35.93$, $p < .001$, $\eta^2_p = .16$. The four interventions that failed to significantly reduce posttest IAT scores (controlling for site) were Shifting Group Boundaries through Competition, $F(1, 213) = 3.56$, $p = .061$, $\eta^2_p = .02$, Shifting Group Affiliations Under Threat, $F(1, 203) = 5.18$, $p = .024$, $\eta^2_p = .03$, Evaluative Conditioning, $F(1, 196) = .57$, $p = .45$, $\eta^2_p = .00$, and Priming Multiculturalism, $F(1, 187) = 2.97$, $p = .086$, $\eta^2_p = .02$. However, all intervention effects were in the expected direction, with some showing weaker effect sizes than in RIRP: I. Lower power of the design may be contributing to non-significance of some effects at posttest.

Overall, IAT scores were slightly smaller at follow-up compared to pretest for intervention conditions. However, this was true of the control condition as well, which likely reflects the reduction of IAT effects as a function of experience with the measure (Greenwald et al., 2003). None of the interventions significantly reduced IAT scores relative to control at follow-up controlling for site, *p*s > .01.

Because participants varied in the amount of time between posttest and follow-up, it is possible that participants who chose to take the follow-up session earlier showed more evidence of persistent change than participants who took the follow-up session later. To examine this possibility, we investigated whether the amount of time between the intervention and the posttest assessment predicted how much IAT scores rebounded. We tested a model that predicted follow-up IAT scores from condition, time between sessions (in seconds), site, and the interaction between condition and time between sessions. We found no main effect of time between sessions, $F(1, 845) = 1.19$, $p = .28$, $\eta^2_p = .00$, and no interaction between condition and time between sessions, $F(9, 845) = .94$, $p = .49$, $\eta^2_p = .01$. Most follow-up sessions were completed 24-96 hours after intervention, suggesting that the interventions cease to have detectable effects in less than that amount of time.

Explicit Racial Preferences

In RIRP: I, we found that interventions were ineffective at changing explicit racial preferences. In the current study, we tested to see if interventions can change explicit racial prejudice again. Participants reported preferences for Whites over Blacks at both posttest and follow-up assessments (*N*s = 989, 860; *M*s = .46, .46; *SD*s = .85, .88; *d*s = .54, .52), and these preferences at posttest and follow-up were highly correlated, $r(843) = .84$. Consistent with RIRP: I, condition did not affect explicit preferences (controlling for site) at posttest, $F(9, 978) = .69$, $p = .72$, $\eta^2_p = .01$, at follow-up, $F(9, 849) = 1.47$, $p = .15$, $\eta^2_p = .02$, or on the average of posttest and follow-up scores, $F(9, 834) = 1.37$, $p = .20$, $\eta^2_p = .02$. See Table 3 for a summary of explicit preferences by condition.

Table 3
Explicit Racial Preferences (Study 1)

	Posttest				Follow-up			
Condition	*N*	*M*	*SD*	*d*	*N*	*M*	*SD*	*d*
Control	98	.48	.81		79	.47	.89	
Exposure to Counterstereotypical Exemplars								
Vivid Counterstereotypic Scenario	100	.54	.86	-.07	91	.51	.80	-.05
Practicing an IAT with Counterstereotypical Exemplars	110	.41	.84	.08	94	.37	.66	.13
Shifting Group Boundaries Through Competition	110	.45	.82	.04	96	.44	.84	.03
Shifting Group Affiliations Under Threat	99	.43	.82	.06	90	.51	.80	-.05
Appeals to Egalitarian Values								
Priming Multiculturalism	88	.51	1.02	-.03	74	.46	1.07	.01
Evaluative Conditioning								
Evaluative Conditioning	93	.56	.97	-.09	80	.71	1.17	-.23
Evaluative Conditioning with the GNAT	104	.33	.69	.20	90	.26	.85	.24
Intentional Strategies to Overcome Biases								
Using Implementation Intentions	100	.41	.81	.09	93	.40	.79	.08
Faking the IAT	87	.49	.85	-.01	76	.45	.90	.02

Note. The explicit measures are an average between two items after standardizing each measure (*SD* = 1) while retaining a rational zero point indicating no preference. More positive scores indicate greater preference for White people over Black people. *d* = Cohen's *d* effect size reflecting change in preference relative to control.

Implicit-Explicit Relations

Implicit and explicit preferences were weakly related at posttest, $r(982) = .10$, $p = .003$, and not related at follow-up, $r(851) = .06$, $p = .063$. These relations were much weaker than previously observed (Nosek, 2007; Nosek et al., 2007), perhaps due to the interventions or to the undergraduate sample. To test the effect of interventions on the strength of implicit-explicit relations, we tested models with condition, site, implicit preferences, and an interaction between condition and implicit preferences in predicting explicit preferences. We did not find an interaction between condition and implicit preferences in predicting explicit preferences at either posttest, $F(9, 963) = 1.80$, $p = .064$, $\eta^2_p = .02$, or at follow-up, $F(9, 832) = .57$, $p = .83$, $\eta^2_p = .01$. We also tested the reverse: condition, site, explicit preferences, and an interaction between condition and explicit preferences in predicting implicit preferences. We did not find an interaction either at posttest, $F(9, 963) = 1.51$, $p = .14$, $\eta^2_p = .01$, or at follow-up, $F(9, 832) = .89$, $p = .53$, $\eta^2_p = .01$.[4]

Support for Affirmative Action

Participants did not support affirmative action overall, $M = .16$, $SD = .26$. Only 223/867 (25.7%) participants supported affirmative action in corporate settings and only 56/850 (6.6%) supported affirmative action in higher education. Overall support for affirmative action was not significantly related to follow-up implicit preferences at $p < .01$, $r(861) = -.08$, $p = .022$, or follow-up explicit prejudice, $r(858) = -.05$, $p = .19$. Experimental condition did not affect support for affirmative action in corporate settings, $\chi^2(9, N = 867) = 8.99$, $p = .44$, higher education, $\chi^2(9, N = 794) = 5.39$, $p = .80$, or overall, $F(9, 860) = 1.06$, $p = .39$, $\eta^2 = .01$.

Robustness Checks and Attrition Rates

Robustness checks. We examined the robustness of the reported analyses in a number of ways. In this section, we summarize the results of these tests (See https://osf.io/um4ye/ for a supplement containing a full review). We tested alternative statistical models of implicit bias change, whether taking a pretest IAT influenced intervention effectiveness, whether demographic characteristics (i.e., university affiliation and participant race) was related to intervention effectiveness, and whether IAT variant (i.e., stimuli, category labels, and background color) and IAT order (i.e., taking the compatible vs. incompatible block first) were related to IAT scores. We did not find evidence for any of these models except for IAT order, suggesting that the effects were relatively robust across design factors. For each of the three IATs, participants were randomly assigned to take the compatible block first or the incompatible block first. Taking the compatible block first led to smaller pretest IAT scores, no difference in the posttest IAT, and larger follow-up IAT scores. IAT order did not interact with experimental condition in predicting IAT scores.

[4] Note that these analyses include a predictor that was measured after the intervention. However, there was no effect of condition on explicit preferences (see Explicit Racial Preferences section). That means it was unlikely there was a confounding influence in the tests.

Attrition rates. To examine attrition rates, we included all participants who began the study except for participants who took T1 multiple times, participants who identified as Black or White/Black multi-racial or chose not to report their racial identity, and participants who accessed the second session before the first session. This left a sample of 1124 participants who began the study, of which 1050 (93.4%) completed the first session and 929 (82.7%) completed both sessions.

We did not find evidence for differential attrition in this study. We first tested to see if experimental condition predicted attrition rates. It did not predict attrition rates within the first session, $\chi^2(9, N = 1108) = 16.86$, $p = .051$, for participants who completed the first session but did not complete the second, $\chi^2(9, N = 1050) = 7.77$, $p = .56$, or overall attrition rates, $\chi^2(9, N = 1108) = 8.77$, $p = .46$. As suggested by reviewers, we also tested models examining attrition as a function of pretest IAT scores and as a function of experimental condition. Pretest IAT scores did not predict first session attrition rates, $\chi^2(1, N = 716) = 1.55$, $p = .21$, attrition rates for participants who completed the first session, $\chi^2(1, N = 682) = 1.13$, $p = .29$, or attrition rates overall, $\chi^2(1, N = 716) = 2.67$, $p = .10$. Experimental condition did not predict attrition rates in these models either, ps = .20, .40, .20. Lastly, we tested whether there were differences in overall attrition rates as a function of an interaction between experimental condition and gender, age, and religiosity. We found no significant evidence for main or interactive effects of experimental condition and demographics in these models, ps > .10.

Discussion

In Study 1, we examined whether the nine interventions from RIRP:I that were immediately effective at reducing implicit preferences continued to be effective after a delay. We found that they did not. Only five of the nine interventions replicated the immediate reduction effect at $p < .01$, and none had an effect after a delay. Implicit preferences rebounded quickly, possibly within several hours. We also found that the interventions did not have an effect on explicit preferences, implicit-explicit relations, or support for affirmative action. One interpretation for these effects is that these interventions induce short-term malleability in implicit preferences but lack the ability to induce a long-term change. Another interpretation is that this study lacked the statistical power to detect effects reliably. Supporting this interpretation, the study only had 36% power to detect intervention effects at $p < .01$ that are similar in size to the average effect in RIRP:I's nine effective interventions. In Study 2, we made an effort to reduce the plausibility of low statistical power as an explanation by conducting a large-scale study with over five times as many participants as Study 1.

Study 2

Method

Participants

Participants were non-Black undergraduates (60.3% White, 69.4% female, median age of 19) from 17 American universities. Our plan for determining sample size was to collect data from as many participants we could in the Fall 2014 semester. 6239 participants began the study at T1, of which 575 (9.2%) were excluded because they did not finish T1 or took T1 multiple

times, 217 (3.5%) because they identified as Black or White/Black multi-racial, 45 (.7%) because they chose not to report their racial identity, 101 (1.6%) because they misbehaved on the IAT (see Dependent Measures section for more detail), 2 (.03%) because their IAT data was missing due to technical issues, and 4 (.06%) because they accessed the second session before the first session. This left a final sample of 5295 participants who completed T1, of which 4888 (92.3%) also completed valid T2 sessions.

This sample was highly powered to detect very small effects. In terms of statistical power to detect an effect size of Cohen's $d = .32$ (the average effect size of the effective interventions from RIRP:I), we had 99.6% power to detect a reduction against control at T1 at $p < .01$ and 99.2% power to detect a reduction against control at T2 at $p < .01$. The T1 data had 80% power to detect $d = .21$ at $p < .01$, and the T2 data had 80% power to detect $d = .22$ at $p < .01$.

Procedure

The procedure for Study 2 was similar to Study 1 with several exceptions. First, the data was collected at 17 sites instead of 2. Most participants from these sites came from psychology participant pools and some were collected through psychology classes. Thirteen out of seventeen sites ($N = 3468$) collected data for the study online like in Study 1, whereas four sites ($N = 1827$) collected data for the first session in-lab and data for the second session online. Procedurally, all participants completed a pretest IAT instead of being randomly assigned to a pretest IAT or not. There was also one day (24 hours) between the initial session and the contact email for the follow-up session. On average, participants returned for the second session after 1.90 days ($SD = 2.97$ days). The interventions remained unchanged with the exception of the Counterstereotypic Training with the IAT, Faking the IAT, and Using Implementation Intentions interventions, which were slightly modified to accommodate changes in procedure and setting (see Interventions section for more information).

Lastly, we removed the questionnaires about affirmative action and effort from the second session and replaced them with the Internal and External Motivations to Respond without Prejudice scales (IMS / EMS; Plant & Devine, 1998). We were interested in whether IMS and EMS were affected by the interventions or moderated intervention effects. Participants rated their agreement with items from each scale on a 7-point Likert scale ranging from (*1*) "Strongly disagree" to (*7*) "Strongly agree". The IMS focuses on personal motivation to respond without prejudice and includes items such as "I attempt to act in nonprejudiced ways toward Black people because it is personally important to me" and "Because of my personal values, I believe that using stereotypes about Black people is wrong." In contrast, the EMS focuses on social pressure to respond without prejudice and includes items such as "I try to act nonprejudiced toward Black people because of pressure from others" and "I try to hide any negative thoughts about Black people in order to avoid negative reactions from others."

Results

As in Study 1, most analyses were conducted with and without data collection site as a covariate. The interpretation of the results for all analyses did not change due to the exclusion of the covariate, so we report only the analyses with site as a covariate in this section. Following

Study 1, we also set our alpha criterion at .01 due to the number of analyses that were conducted. A complete description of the pre-registered analysis plan (and deviations from that plan) is available at https://osf.io/7aqm9 and supplemental analyses such as the analyses without site as a covariate are available at https://osf.io/um4ye/.

Implicit Racial Preferences

Participants completed three IATs over the course of two sessions. Overall, participants held implicit preferences for Whites over Blacks at pretest, posttest, and follow-up assessments (*N*s = 5270, 5271, 4861; *M*s = .57, .30, 45; *SD*s = .41, .49, .42; *d*s = 1.39, .61, 1.08). These IAT scores were positively correlated at similar magnitudes as Study 1, $r_{\text{pretest-posttest}}(5256) = .25$, $r_{\text{pretest-follow-up}}(4836) = .25$, $r_{\text{posttest-follow-up}}(4837) = .27$. IAT scores were moderated by time of assessment, $F(1, 4810) = 257.07$, $p < .001$, $\eta^2_p = .05$, condition, $F(9, 4810) = 33.38$, $p < .001$, $\eta^2_p = .06$, and an interaction between time and condition, $F(9, 4810) = 54.34$, $p < .001$, $\eta^2_p = .09$, but not by data collection site, $F(19, 4810) = 1.78$, $p = .019$, $\eta^2_p = .01$, or an interaction between time and site, $F(19, 4810) = 1.48$, $p = .081$, $\eta^2_p = .01$. Condition had significant effects on IAT scores (controlling for site) at posttest, $F(9, 5242) = 72.39$, $p < .001$, $\eta^2_p = .11$, but not at pretest, $F(9, 5241) = .83$, $p = .59$, $\eta^2_p = .00$, or at follow-up, $F(9, 4832) = 1.19$, $p = .29$, $\eta^2_p = 00$. See Table 4 for a summary of IAT scores by condition.

Table 4
Implicit Racial Preferences (Study 2)

	Pretest			Posttest				Follow-up			
Condition	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*N*	*M*	*SD*	*d*
Control	507	.58	.40	506	.48	.40		465	.48	.42	
Exposure to Counterstereotypical Exemplars											
Vivid Counterstereotypic Scenario	514	.58	.40	516	.27	.45	.49***	483	.43	.44	.12
Practicing an IAT with Counterstereotypical Exemplars	537	.59	.40	534	.32	.40	.39***	483	.43	.44	.06
Shifting Group Boundaries Through Competition	537	.56	.41	535	.30	.43	.42***	496	.43	.31	.11
Shifting Group Affiliations Under Threat	525	.57	.40	528	.34	.44	.32***	474	.46	.39	.03
Appeals to Egalitarian Values											
Priming Multiculturalism	514	.54	.44	515	.35	.44	.30***	478	.43	.42	.11
Evaluative Conditioning											
Evaluative Conditioning	524	.55	.40	523	.41	.40	.16**	488	.47	.40	.02
Evaluative Conditioning with the GNAT	519	.58	.40	519	.39	.40	.22***	480	.45	.41	.07
Intentional Strategies to Overcome Biases											
Using Implementation Intentions	560	.57	.40	560	.29	.43	.44***	512	.42*	.41	.15*
Faking the IAT	533	.55	.42	535	-.16	.74	1.06***	500	.46	.42	.06

Note. Descriptive statistics reflect *D* scores (Greenwald et al., 2003), and positive values indicate greater preference for White people compared to Black people. *d* = Cohen's *d* effect size reflecting change in implicit preference relative to control. Significance is from a *t*-test contrasting an experimental condition against the control condition. * $p < .05$. ** $p < .01$. *** $p < .001$.

Eight of the nine interventions significantly reduced implicit preferences relative to the control condition at posttest controlling for site: Vivid Counterstereotypic Scenario, $F(1, 1001) =$

63.21, $p < .001$, $\eta^2_p = .06$, Practicing an IAT with Counterstereotypic Exemplars, $F(1, 1019) = 40.27$, $p < .001$, $\eta^2_p = .04$, Shifting Group Boundaries through Competition, $F(1, 1020) = 46.32$, $p < .001$, $\eta^2_p = .04$, Shifting Group Affiliations Under Threat, $F(1, 1013) = 25.31$, $p < .001$, $\eta^2_p = .02$, Priming Multiculturalism, $F(1, 1000) = 24.38$, $p < .001$, $\eta^2_p = .02$, Evaluative Conditioning with the GNAT, $F(1, 1004) = 15.12$, $p < .001$, $\eta^2_p = .02$, Using Implementation Intentions, $F(1, 1045) = 52.51$, $p < .001$, $\eta^2_p = .05$, and Faking the IAT, $F(1, 1020) = 286.73$, $p < .001$, $\eta^2_p = .22$. Evaluative Conditioning was the one intervention that failed to significantly reduce posttest IAT scores controlling for site, $F(1, 1008) = 6.16$, $p = .013$, $\eta^2_p = .01$, with a p-value not quite meeting our significance criterion in this analysis (though it did in a t-test comparison with control, as shown in Table 2).

None of the interventions significantly reduced implicit preferences relative to control at follow-up (controlling for site). However, Implementation Intentions reduced implicit racial preferences at follow-up at $p < .05$ but was not significant by our $p < .01$ criterion, $F(1, 956) = 5.00$, $p = .026$, $\eta^2_p = .01$. Given the number of tests, our default interpretation is that this is likely a false positive.

To examine how quickly intervention effects faded, we tested a model predicting follow-up IAT scores from condition, time between sessions (in seconds), site, and the interaction between condition and time between sessions. We found no main effect of time between sessions, $F(1, 4517) = .69$, $p = .41$, $\eta^2_p = .00$, and no interaction between condition and time between sessions, $F(9, 4517) = 1.75$, $p = .072$, $\eta^2_p = .00$. This suggests that intervention effects dissipate more quickly than the variation in follow-up timing in our study could detect.

Explicit Racial Preferences

Participants reported overall preferences for Whites over Blacks at both posttest and follow-up assessments (*N*s = 5149, 4782; *M*s = .41, .41, *SD*s = .87, .88, *d*s = .46, .47). Explicit preferences at posttest and follow-up were highly correlated, $r(4711) = .86$. Condition did not affect explicit preferences (controlling for site) at posttest, $F(9, 5120) = 1.56$, $p = .12$, $\eta^2_p = .00$, at follow-up, $F(9, 4753) = 2.21$, $p = .019$, $\eta^2_p = 00$, or on the average of posttest and follow-up scores, $F(9, 4684) = 1.70$, $p = .083$, $\eta^2_p = .02$. See Table 5 for a summary of explicit preferences by condition.

Table 5
Explicit Racial Preferences (Study 2)

	Posttest				Follow-up			
Condition	*N*	*M*	*SD*	*d*	*N*	*M*	*SD*	*d*
Control	492	.44	.86	N/A	452	.44	.83	N/A
Exposure to Counterstereotypical Exemplars								
Vivid Counterstereotypic Scenario	503	.38	.84	.07	478	.36	.84	.10
Practicing an IAT with Counterstereotypical Exemplars	524	.32	.77	.15*	479	.32	.82	.15*
Shifting Group Boundaries Through Competition	518	.38	.77	.07	484	.34	.90	.12
Shifting Group Affiliations Under Threat	512	.39	.89	.06	466	.43	.92	.01
Appeals to Egalitarian Values								
Priming Multiculturalism	503	.37	.85	.08	473	.43	.92	.01
Evaluative Conditioning								
Evaluative Conditioning	513	.45	.87	-.01	480	.47	.88	-.04
Evaluative Conditioning with the GNAT	506	.44	.97	.00	471	.39	.93	.06
Intentional Strategies to Overcome Biases								
Using Implementation Intentions	553	.47	.92	-.03	508	.49	.97	-.06
Faking the IAT	525	.45	.84	-.01	491	.47	.82	-.04

Note. The explicit measures are an average between two items after standardizing each measure (*SD* = 1) while retaining a rational zero point indicating no preference. More positive scores indicate greater preference for White people over Black people. *d* = Cohen's *d* effect size reflecting change in preference relative to control. * $p < .05$.

Implicit-Explicit Relations

Implicit and explicit preferences were weakly related at posttest, $r(5125) = .16$, $p < .001$, and at follow-up, $r(4727) = .17$, $p < .001$. To test the effect of interventions on the strength of implicit-explicit relations, we tested models that predicted explicit preferences from condition, site, implicit preferences, and an interaction between condition and implicit preferences. We did not find evidence for an interaction between condition and implicit preferences at either posttest, $F(9, 5088) = .91$, $p = .52$, $\eta^2_p = .00$, or at follow-up, $F(9, 4720) = .76$, $p = .66$, $\eta^2_p = .00$. Testing the reverse models (predicting implicit preferences from condition, site, explicit preferences, and an interaction between condition and explicit preferences) did not lead to an an interaction either at posttest, $F(9, 5088) = 1.25$, $p = .26$, $\eta^2_p = .00$, or at follow-up, $F(9, 4720) = .77$, $p = .64$, $\eta^2_p = .00$.

Internal and External Motivation to Respond Without Prejudice

On average, participants were motivated to respond without prejudice internally, $M = 5.65$, $SD = 1.09$, $d = 1.51$, and externally, $M = 4.52$, $SD = 1.29$, $d = .40$. Internal and external motivations were very weakly correlated, $r(4855) = .04$, $p = .010$. Internal motivation was negatively related to explicit preferences, $r_{posttest} = -.30$, $p < .001$, $r_{follow\text{-}up} = -.31$, $p < .001$ and implicit preferences, $r_{pretest} = -.09$, $p < .001$, $r_{posttest} = -.10$, $p < .001$, $r_{follow\text{-}up} = -.05$, $p = .001$, such that higher internal motivation predicted lower pro-White/anti-Black preference. xternal motivation was positively related to explicit preferences, $r_{posttest} = .29$, $p < .001$, $r_{follow\text{-}up} = .29$, p

<.001 and implicit preferences, $r_{pretest}$ = .12, p < .001, $r_{posttest}$ = .07, p < .001, $r_{follow\text{-}up}$ = .11, p < .001, such that higher external motivation predicted greater pro-White/anti-Black preference.

Condition (controlling for site) did not affect internal motivation, $F(9, 4838) = .49$, $p = .88$, $\eta^2_p = .00$, or external motivation, $F(9, 4835) = .80$, $p = .62$, $\eta^2_p = .00$, to respond without prejudice. Further, internal motivation did not interact with condition (controlling for site) in predicting posttest IAT scores, $F(9, 4806) = 1.93$, $p = .044$, $\eta^2_p = .00$, or follow-up IAT scores, $F(9, 4802) = 1.50$, $p = .14$, $\eta^2_p = .00$. Similarly, external motivation did not interact with condition (controlling) for site in predicting posttest IAT scores, $F(9, 4803) = .97$, $p = .47$, $\eta^2_p = .00$, or follow-up IAT scores, $F(9, 4800) = 1.60$, $p = .11$, $\eta^2_p = .00$. Prior research had found that internal and external motivation interact in predicting IAT scores (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002), such that people with high internal and low external motivation show lower implicit bias than other people. We did not find evidence for this interaction (controlling for site) at any time point, ps = .62, .88, .47, η^2_ps = .00, .00, .00, nor did we find significant evidence for a three-way interaction between internal motivation, external motivation, and condition (controlling for site) at posttest, $F(9, 4776) = .55$, $p = .84$, $\eta^2_p = .00$, or follow-up, $F(9, 4773) = 1.02$, $p = .42$, $\eta^2_p = .00$. This suggests that internal and external motivations were not reliably related to how participants engaged with the interventions.

Robustness Checks and Attrition Rates

Robustness checks. In this section, we summarize our efforts to test the robustness of our results (See https://osf.io/um4ye/ for a supplement with a complete review of these tests). We tested alternative statistical models of IAT score change, whether demographic characteristics (i.e., university affiliation, participant race, characteristics of the university town population) were related to intervention effectiveness, and whether IAT variant (i.e., stimuli, category labels, and background color) and IAT order (i.e., taking the compatible vs. incompatible block first) were related to IAT scores.

As with Study 1, we did not find positive evidence for any of these robustness checks except for IAT order. Taking the compatible block first instead of the incompatible block first led to smaller pretest IAT scores, no difference in the posttest IAT, and higher follow-up IAT scores. IAT order did not interact with experimental condition in predicting IAT scores. Interestingly, we found the proportion of White students and the proportion of Black students in a university were both weakly positively related to greater implicit and explicit preferences for White people over Black people (rs ranging from .04 to .12). Neither variable interacted with experimental condition in predicting IAT scores. These results are consistent with recent findings showing that people in states with a larger proportion of Black residents tended to have higher IAT scores (Rae, Newheiser, & Olson, 2015), although the many analyses conducted suggest that one should be cautious in overinterpreting these effects.

Attrition rates. As with Study 1, we analyzed attrition rates with all participants who began the study except for participants who took T1 multiple times, participants who identified as Black or White/Black multi-racial or chose not to report their racial identity, and participants who accessed the second session before the first session. This left a sample of 5560 participants

who began the study, of which 5398 (97.1%) completed the first session and 5042 (90.7%) completed both sessions.

We did not find differential attrition by experimental condition for attrition rates in the first session, $\chi^2(9, N = 5455) = 8.59$, $p = .48$, attrition rates for participants who completed the first session but not the second, $\chi^2(9, N = 5398) = 9.90$, $p = .36$, or overall attrition rates, $\chi^2(9, N = 5455) = 8.90$, $p = .45$. As suggested by reviewers, we also tested models examining attrition as a function of both pretest IAT scores and experimental condition. Pretest IAT scores did not significantly predict at $p < .01$ attrition rates within the first session, $\chi^2(1, N = 5415) = 4.56$, $p = .033$, attrition rates for participants who completed the first session but not the second, $\chi^2(1, N = 5363) = 7.62$, $p = .006$, and attrition rates overall, $\chi^2(9, N = 5415) = 9.30$, $p = .002$. Corresponding zero-order correlations revealed a very weak positive relationship (rs = .03, .04, .04), indicating that participants with higher pretest IAT scores were more likely to complete the various phases of the study. Experimental condition did not predict attrition rates in any of these models, ps = .50, .30, .40. Lastly, we tested whether there was differential attrition for overall attrition rates as a function of an interaction between experimental condition and gender, age, and religiosity. We found no significant evidence for main or interactive effects of experimental condition and demographics in these models, ps > .10.

General Discussion

In two studies with 6,321 total participants, we compared the effectiveness of nine interventions on reducing implicit preferences immediately or after a delay. All nine interventions were effective at reducing implicit preferences immediately, and none were effective after a delay.

Interventions were effective at inducing short-term malleability in implicit preferences

Interventions varied greatly in their effectiveness at shifting implicit preferences immediately (Figure 2). The sham intervention, Faking the IAT, was most effective when meta-analytically aggregating across both studies, $d = 1.03$, 95% CI [.92, 1.15]. The most effective non-sham intervention was Vivid Counterstereotypic Scenario, $d = .52$, 95% CI [.40, .63], followed by Using Implementation Intentions, $d = .44$, 95% CI [.33, .55], Practicing an IAT with Counterstereotypical Exemplars, $d = .40$, 95% CI [.28, .51], Shifting Group Boundaries Through Competition $d = .39$, 95% CI [.28, .50], Shifting Group Affiliations Under Threat $d = .32$, 95% CI [.21, .44], Priming Multiculturalism $d = .29$, 95% CI [.18, .40], Evaluative Conditioning with the GNAT $d = .26$, 95% CI [.14, .37], and Evaluative Conditioning $d = .15$, 95% CI [.04, .26].

The current (RIRP:II) meta-analytic effect sizes were remarkably consistent with RIRP:I. Differences in effect size between RIRP:I and RIRP:II ranged from $d = .00$ to $d = .64$ with a median difference of $d = .06$. There was one outlier, however: Faking the IAT. This sham intervention had an effect in RIRP:II that was more than double the size ($d = 1.03$) of the effect in RIRP:I ($d = .39$). Without Faking the IAT, the correlation between effect sizes was $r = .91$ and the average effect sizes for RIRP:I and RIRP:II were both $d = .35$. Including Faking the IAT in the analyses reduced the correlation to $r = .51$ and increased the average effect size of RIRP:II to $d = .42$. Why was Faking the IAT so much more effective in the current research? One potential

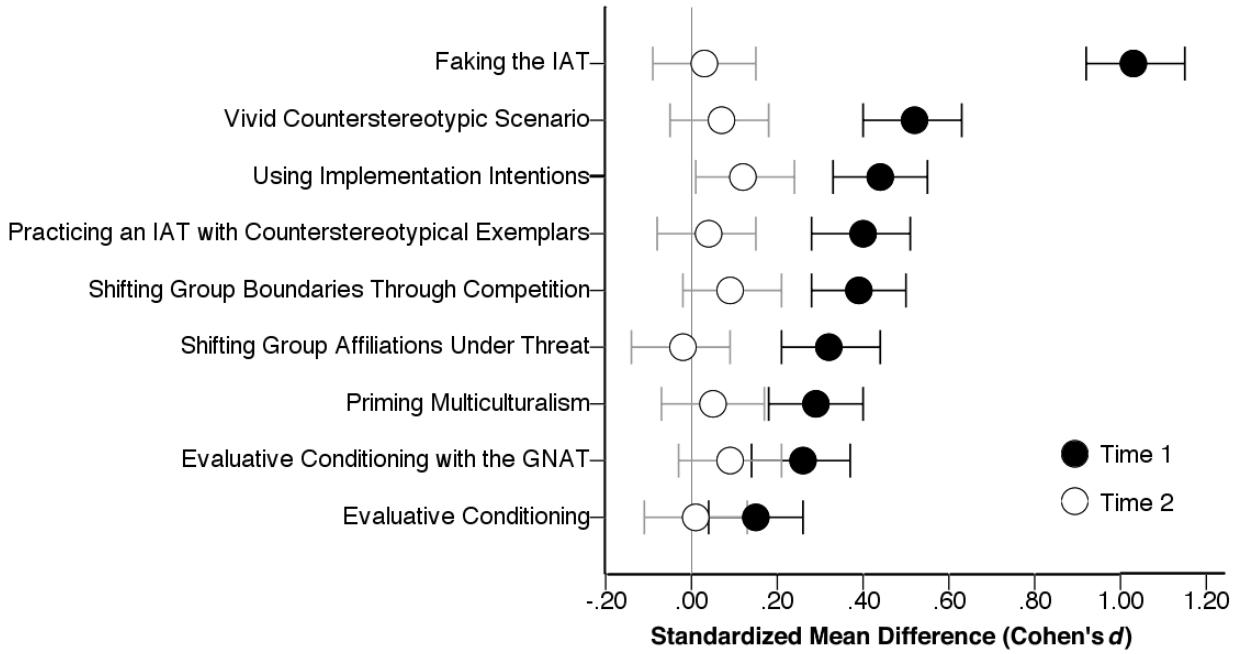

Figure 2. Meta-analytic effectiveness of interventions on implicit racial preferences, organized from most to least effective at T1. Cohen's *d* = reduction in implicit preferences compared to control; Black circles = Effect sizes at T1 (posttest); White circles = Effect Sizes at T2 (follow-up); Lines = 95% confidence interval. IAT = Implicit Association Test; GNAT = Go/No-Go Association Task.

explanation could be differences in participant motivation. RIRP:I used on-line volunteers and RIRP:II used undergraduate participants receiving credit. The latter could be more compliant with faking instructions.

As with RIRP:I, there was considerable variation in immediate effectiveness. Effect sizes ranged from d = .15 to d = 1.03, with an average d of .42 and a standard deviation of .25. Combining interventions across their four descriptive categories suggests that the most effective category was Intentional Strategies to Overcome Bias, d = .72, 95% CI [.64, .80], followed by Exposure to Counterstereotypical Exemplars, d = .41, 95% CI [.35, .46], Appealing to Egalitarian Values, d = .29, 95% CI [.18, .40], and Evaluative Conditioning, d = .20, 95% CI [.12, .28]. Interventions that were more self-relevant, emotional, and vivid tended to be more effective than those which were less involving. These category rankings are similar to RIRP:I with the exception of Intentional Strategies to Overcome Bias, which yielded an aggregate effect size that was more than double the size of the original studies. These effects are driven more by increases in Faking the IAT's effectiveness than increases in Using Implementation Intentions effectiveness.

Interventions were not effective at changing implicit preferences after a delay

In contrast to the interventions' immediate effectiveness, all nine interventions failed to create sustained change in implicit preferences after a delay of up to several days despite well-powered samples (Figure 2). Using Implementation Intentions was the closest to producing a

robust effect, d = .12, 95% CI [.01 .24]. The rest were robustly ineffective: Shifting Group Boundaries Through Competition, d = .09, 95% CI [-.02, .21], Evaluative Conditioning with the GNAT, d = .09, 95% CI [-.03, .21], Vivid Counterstereotypic Scenario, d = .07, 95% CI [-.05, .18], Priming Multiculturalism, d = .05, 95% CI [-.07, .17], Practicing an IAT with Counterstereotypical Exemplars, d = .04, 95% CI [-.08, .15], Faking the IAT, d = .03, 95% CI [-.09, .15], Evaluative Conditioning, d = .01, 95% CI [-.11, .13], and Shifting Group Affiliations Under Threat, d = -.02, 95% CI [-.14, .09].

Some participants came back to take the study within one day, whereas others came back after several days. We examined whether this was related to differences in intervention effectiveness. It's possible that some intervention effects declined rapidly, whereas others declined more slowly. We did not find evidence for this, as time between sessions did not explain variability in intervention effectiveness. This suggests that the fading away of intervention effects occurred within a day or so rather than over the course of several days.

Implicit preference malleability does not necessarily indicate implicit preference change

The most dramatic result of the current research is simultaneous strong evidence for short-term malleability in implicit preference and little evidence for long-term implicit preference change just a couple of days later. One interpretation is that implicit preferences are stable over time and are not susceptible to long-term change. Recent advances in developmental psychology appear to support this claim. Implicit preferences for social groups are observable within the first year of an infant's life (Baron, 2013) and White children as young as 3 years old exhibit implicit pro-White racial attitudes which remain stable throughout development (e.g., Dunham, Chen, & Banaji, 2013). These implicit preferences for one's own group are also present from an early age for other social categories including gender, religion, and caste (Cvencek, Greenwald, & Meltzoff, 2011; Dunham, Srinivasan, Dotsch, & Barner, 2014; Heiphetz, Spelke, & Banaji, 2013). However, the absence of developmental change in implicit preferences may reflect the stability of exposure to cultural messages rather than a deep level of stability in the implicit preferences themselves (Baron, 2015). From this view, it is possible to affect long-term change in implicit preferences but the interventions tested in the current studies were inadequate for doing so. In this section, we review three alternative explanations for how interventions could change implicit preferences in the long-term, despite the evidence shown here.

Effective mechanisms have not yet been tested. It could be that effective mechanisms for long-term implicit preference change have simply not yet been tested. While it is certainly true that researchers have not tested all mechanisms for implicit preference change, this is not a compelling dismissal of the current results. The nine interventions tested in the current research were culled from a larger pool of 18 interventions from RIRP:I that social psychological researchers thought would be maximally effective for changing implicit preferences. Those 18 interventions reflected state-of-the-art knowledge about implicit attitude change at the time. These nine interventions were also distinct because they were immediately effective at reducing implicit preferences, which is an almost-necessary condition for long-term attitude change (cf. the sleeper effect and interventions which elicit downstream exposure to external sources of attitude change; Frey & Rogers, 2014; Hovland, Lumsdaine, & Sheffield, 1949; Pratkanis, Greenwald, Leippe, & Baumgardner, 1988). Also, these interventions represent a wide range of

the published literature on bias-reduction techniques (see reviews by Blair, 2002; Dasgupta, 2013; Lai et al., 2013; Sritharan & Gawronski, 2010). However, new approaches which have been published since RIRP:I might yield stronger evidence for long-term effectiveness (e.g., Maister, Slater, Sanchez-Vives, & Tsakiris, 2015; Hu, Antony, Creery, Vargas, Bodenhausen, & Paller, 2015).

It is also possible that interventions are not changing implicit preferences *per se*, but are instead changing non-associative factors that are related to IAT performance (Calanchini, Sherman, Klauer, & Lai, 2014; Calanchini & Sherman, 2013; Lai et al., 2013). For example, changes in IAT scores may reflect temporary changes in task performance rather than altering associations in memory. This is likely for Faking the IAT, and could occur for other manipulations. Mathematical modeling procedures such as the Quadruple Process model that attempt to decompose the processes contributing to IAT performance could be used to test this (Calanchini, Gonsalkorale, Sherman, & Klauer, 2013; Gonsalkorale, Allen, Sherman, & Klauer, 2010; Sherman, Gawronski, Gonsalkorale, Hugenberg, Allen, & Groom, 2008). Alternatively, these interventions could be administered with multiple implicit measures as dependent variables (Bar-Anan & Nosek, 2014). Other measures may not be sensitive to the same extraneous influences, so examining effects across measures could triangulate what changes are due to idiosyncratic features of implicit measures.

Interventions need to be longer or more intensive. It is possible the mechanisms employed by interventions in the current studies can be effective, but the current interventions were not long or intensive enough to create long-lasting change. Supporting this claim, Devine and colleagues (2012) found reduced implicit racial preferences (relative to a control) after 12 weeks using an hour-long intervention that included many of the methods employed in the current research. However, follow-up replications employing larger samples did not replicate evidence for effectiveness of even this substantial intervention (Forscher, Mitamura, Dix, Cox, & Devine, 2016). This suggests that simply making interventions longer and more intensive will not be sufficient for long-term change.

It is also possible that brief interventions can be effective, but only when administered repeatedly over time in a spaced learning schedule (Greene, 1990; Hintzman & Block, 1973). Moreover, giving participants reminder cues shortly before follow-up testing might activate memories of interventions so that they are influential. To our knowledge, neither of these approaches to increasing intervention effectiveness has been systematically tested with implicit measures.

One area of promising evidence for long-term change is research involving prolonged everyday experiences. These interventions are primarily conducted outside of psychology laboratories. Prolonged interventions that have been successful in changing implicit preferences and stereotypes include: taking a semester-long class on prejudice and intergroup conflict (Rudman, Ashmore, & Gary, 2001), having an college roommate who is of a different race (Shook & Fazio, 2008), participating in a cultural music education program (Neto, da Conceiçao Pinto, & Mullet, 2015), and taking a class with a female professor (Dasgupta & Asgari, 2004; Stout, Dasgupta, Hunsinger, & McManus, 2011). In many ways, these field tests may best represent how implicit associations change in real-world settings. An important caveat for most

of these studies, however, is that they assess implicit associations while the study interventions are still ongoing. Thus, they do not provide evidence of the durability of these interventions. It could be that associations remain changed after a prolonged experience is over, or that they return to pre-intervention baseline after exposure to the intervention ends.

We have the right interventions, but the wrong population. A sobering possibility is that these interventions are effective for long-term change, but that change is too difficult for adults. Implicit racial preferences may be ingrained early in development (Dunham et al., 2013) and may only be susceptible to interventions at certain points in the lifespan (Baron, 2015). Interventions on children could be more effective because implicit preferences are more sensitive to experiences at younger ages (i.e., a critical period) or because adults' associations are more stable since they have accumulated more experiences related to an association over a lifetime. We are aware of only four studies that have experimentally examined change in children's implicit preferences to test these questions (Gonzalez, Steele, & Baron, 2016; Neto, et al., 2015; Vezzali et al., 2011; Xiao et al., 2014). As one example, Neto and colleagues' (2015) tested an education program where sixth-graders learned about music and culture from Cape Verde over the course of six months. Compared to sixth-graders who did not undergo this education program, participants showed reduced implicit and explicit anti-dark-skin preferences up to two years after completion of the education program. This suggests that children's' implicit preferences *can* remain changed for years after an intervention has taken place. Understanding *how* and *when* will be important next steps.

Ineffectiveness in changing explicit prejudice, beliefs, and motivations

Not only did all of the interventions tested here fail to reduce implicit racial preferences, but they also failed to reduce explicit racial prejudice. Moreover, the interventions did not change support for affirmative action or internal/external motivations to respond without prejudice. These null findings may reflect a general lack of malleability in these constructs or a side effect of the fact that interventions were designed to focus on changing implicit preferences. Studies that specifically target explicit racial prejudice, support for affirmative action, or motivations to respond without prejudice may yield more evidence of effectiveness.

Conclusion

The current work showcases nine interventions out of an initial field of 18 that were effective at reducing implicit preferences immediately. However, the intervention effects were fleeting, lasting less than a couple days. These findings are a testament to how the mind's prejudices remain steadfast in the face of efforts to change them. Understanding when implicit preferences can be changed in the long-term will be the next frontier for research on change in implicit associations.

References

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven implicit measures of social cognition. *Behavior Research Methods, 46,* 668-688.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York, NY: Guilford Press.

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

Baron, A. S. (2013). Bridging the gap between preference and evaluation during the first few years of life. In S. A. Gelman & M. R. Banaji (Eds.), *Social cognitive development* (pp. 281–285). New York, NY: Oxford University Press.

Baron, A. S. (2015). Constraints on the development of implicit intergroup attitudes. *Child Development Perspectives, 9,* 50–54.

Gonzalez, A. M., Steele, J., & Baron, A.S.. (2016). Reducing children's implicit racial bias through exposure to positive outgroup exemplars. *Child Development.*

Bayer, U. C., Achtziger, A., Gollwitzer, P. M., & Moskowitz, G. B. (2009). Responding to subliminal cues: do if-then plans facilitate action preparation and initiation without conscious intent? *Social Cognition, 27,* 183–201.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6,* 242–261.

Calanchini, J., Gonsalkorale, K., Sherman, J. & Klauer, C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology, 43,* 321-325.

Calanchini, J. & Sherman, J. W. (2013). Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass, 7,* 654-667.

Calanchini, J., Sherman, J., Klauer, C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin, 40,* 1285-1296.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology, 32,* 302–314.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4-year-olds: The Preschool Implicit Association Test. *Journal of Experimental Child Psychology, 109,* 187–200.

Dasgupta, N., & Greenwald, A. G. (2001). Exposure to admired group members reduces automatic intergroup bias. *Journal of Personality and Social Psychology, 81,* 800–814.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology, 47,* 233–279.

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,* 642–658.

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,* 800–814.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127,* 853-869.

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5-18.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48,* 1267–1278.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82,* 835-848.

Dunham, Y., Chen, E. E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science, 24,* 860–868.

Dunham, Y., Srinivasan, M., Dotsch, R., & Barner, D. (2014). Religion insulates ingroup evaluations: The development of intergroup attitudes in India. *Developmental Science, 17,* 311–319.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27,* 307-316.

Forscher, P.S., Mitamura, C., Dix, E. L, Cox, W. T. L., & Devine, P.G. (2016). Breaking the prejudice habit: More evidence of long-term change. Unpublished manuscript.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. E., Herman, M., Devine, P. G., & Nosek, B. A. (2016). A meta-analysis of approaches to changing implicit bias. Unpublished manuscript

Federico, C. M., & Sidanius, J. (2002). Racism, ideology, and affirmative action revisited: the antecedents and consequences of "principled objections" to affirmative action. *Journal of Personality and Social Psychology, 82,* 488-502.

Frey, E., & Rogers, T. (2014). Persistence: How treatment effects persist after interventions stop. *Policy Insights from the Behavioral and Brain Sciences, 1,* 172-179.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist, 54,* 493-503.

Gonsalkorale, K., Allen, T.J., Sherman, J.W., & Klauer, K.C. (2010). Mechanisms of group membership and exemplar exposure effects on implicit attitudes. *Social Psychology, 41,* 158-168.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1004-1011.

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,* 197-216.

Haley, H., & Sidanius, J. (2006). The positive and negative framing of affirmative action: A group dominance perspective. *Personality and Social Psychology Bulletin, 32,* 656-668.

Heiphetz, L., Spelke, E. S., & Banaji, M. R. (2013). Patterns of implicit and explicit attitudes in children and adults: Tests in the domain of religion. *Journal of Experimental Psychology: General, 142,* 864–879.

Hintzman, D. L., & Block, R. A. (1973). Memory for the spacing of repetitions. *Journal of Experimental Psychology, 99*, 70-74.

Hu, X., Antony, J. W., Creery, J. D., Vargas, I. M., Bodenhausen, G. V., & Paller, K. A. (2015). Unlearning implicit social biases during sleep. *Science, 348,* 1013-1015.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication, Vol. 3.* Princeton, NJ: Princeton University Press.

Jang, D. and Kim, D.Y., 2011. Increasing implicit life satisfaction. *Social Behavior and Personality, 39*, 229-239.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137–146.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871-88.

Kim, D. (2003). Voluntary controllability of the Implicit Association Test (2003). *Social Psychology Quarterly, 66*, 83-96.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*, 315–330.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., … Nosek, B.N. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765.

Maister, L., Slater, M., Sanchez-Vives, M. V., & Tsakiris, M. (2015). Changing bodies changes minds: Owning another body affects social cognition. *Trends in Cognitive Sciences, 19*, 6-12.

Marini, M., Rubichi, S., & Sartori, G. (2012). the role of self-involvement in shifting IAT effects. *Experimental Psychology, 59*, 348–354.

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*, 512-523.

Neto, F., da Conceiçao Pinto, M., & Mullet, E. (2015). Can music reduce anti-dark-skin prejudice? A test of a cross-cultural musical education programme. *Psychology of Music*. Advance online publication.

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science, 16*, 65-69.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19*, 625–666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166–180.

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,* 36-88.

O'Brien, K. S., Puhl, R. M., Latner, J. D., Mir, A. S., & Hunter, J. A. (2010). Reducing anti-fat prejudice in preservice health students: A randomized trial. *Obesity, 18*, 2138–2144.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*, 413–417.

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition, 20,* 89–104.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32,* 421–433.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75,* 811–832.

Pratkanis, A. R., Greenwald, A. G., Leippe, M. R., & Baumgardner, M. H. (1988). In search of reliable persuasion effects: III. The sleeper effect is dead: Long live the sleeper effect. *Journal of Personality and Social Psychology, 54,* 203-218.

Rae, J. R., Newheiser, A. K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science, 6,* 535-543.

Richeson, J. A., & Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology, 40,* 417–423.

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review, 10,* 336–353.

Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science, 13,* 79–82.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81,* 856–868.

Schmidt, K., & Nosek, B. A. (2010). Implicit (and explicit) racial attitudes barely changed during Barack Obama's presidential campaign and early presidency. *Journal of Experimental Social Psychology, 46,* 308–314.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1-66.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115,* 314-335.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84,* 127-190.

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships: An experimental field test of the contact hypothesis. *Psychological Science, 19,* 717–723.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*), 137-150.

Sportel, B. E., de Hullu, E., de Jong, P. J., & Nauta, M. H. (2013). Cognitive bias modification versus CBT in reducing adolescent social anxiety: A randomized controlled trial. *PloS ONE, 8:* e64355.

Sritharan, R., & Gawronski, B. (2010). Changing implicit and explicit prejudice. *Social Psychology, 41,* 113-123.

Steffens, M.C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51,* 165-179.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34,* 1332–1345.

Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology, 100,* 255-270.

Thomas, J. J., Judge, A. M., Brownell, K. D., & Vartanian, L. R. (2006). Evaluating the effects of eating disorder memoirs on readers' eating attitudes and behaviors. International Journal of *Eating Disorders, 39*, 418-425.

Xiao, W. S., Fu, G., Quinn, P. C., Qin, J., Tanaka, J. W., Pascalis, O., & Lee, K. (2014). Individuation training with otherrace faces reduces preschoolers' implicit racial bias: a link between perceptual and social representation of faces in children. *Developmental Science, 18,* 655-663.

Vezzali, L., Capozza, D., Giovannini, D., & Stathi, S. (2011). Improving implicit and explicit intergroup attitudes using imagined contact: An experimental intervention with elementary school children. *Group Processes & Intergroup Relations, 15*, 203-212.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*, 101-126.

Wolsko, C., Park, B., Judd, C. M., & Wittenbrink, B. (2000). Framing interethnic ideology: effects of multicultural and color-blind perspectives on judgments of groups and individuals. *Journal of Personality and Social Psychology,* 78, 635-654.