Forging links with the self to combat implicit bias

Anna Woodcock¹ and Margo J. Monteith¹

Abstract

Two experiments examined whether Whites’ implicit biases toward Blacks could be reduced by conditioning links between Blacks and the self. Via a computer-mediated experience, White participants were assigned to the same (minimal) group as several Black individuals and practiced classifying photographs as “MY GROUP” or “OTHER GROUP.” Subsequent performance on implicit prejudice and stereotyping measures was compared to a control condition and another condition involving extensive counterstereotype conditioning. Across experiments, the link to self strategy significantly reduced implicit prejudice, relative to the control condition, and to the same extent as the counterstereotype conditioning condition. Process dissociation analyses revealed that these effects corresponded with a reduction in the automatic activation of biased associations. Counterstereotype conditioning also reduced implicit stereotyping, but the link-to-self strategy did not. These findings extend prior work on the reduction of implicit biases and highlight the importance of comparing implicit bias strategies across different types of bias measures.

Keywords

implicit bias, prejudice reduction, stereotyping, diversity

The measurement and impact of implicit or automatic biases has been a central focus of stereotyping and prejudice research over the past decade. The often pernicious effects of implicit biases have been recognized well beyond the laboratory. Indeed, implicit biases are identified as obstacles to equality in many domains. For example, implicit gender bias has been implicated as one of eight core social and environmental factors contributing to women’s chronic underrepresentation in the science, technology, mathematics, and engineering professions (Hill, Corbett, & St. Rose, 2010). Similarly, research concerning racial health disparities indicates that physicians’ implicit race bias has a profound impact on the diagnosis and treatment of African Americans (Green et al., 2007). Clearly there is a pressing need for strategies that can be applied to the reduction of these biases.

Social psychological research and theory have advanced our knowledge of the existence and

¹Purdue University, USA

Corresponding author:
Anna Woodcock, Psychology Department, California State University, San Marcos, 333 S. Twin Oaks Valley Road, San Marcos, CA 92069, USA.
Email: woodcock@csusm.edu
nature of implicit stereotypes and prejudices, but we are just beginning to make inroads into the effective reduction of these biases. The present research tested the effectiveness of a novel strategy for reducing implicit biases that involved conditioning a link between the self and outgroup members by having participants repeatedly identify specific members of an outgroup as belonging to “my group.” We argue that such a self-linking situation is inherent in real-world settings where society’s rising levels of diversity may increasingly result in diverse groups, such as in school and work settings. In addition, we compared the self-linking strategy to another previously established implicit bias reduction strategy (i.e., counterstereotyping, see Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000) so that potentially unique effects on different types of implicit biases could be examined. That is, implicit intergroup bias can take the form of implicit prejudice (i.e., the automatic activation of negative attitudes in relation to the outgroup) and also implicit stereotyping (i.e., the automatic activation of specific traits associated with the outgroup), and different implicit bias reduction strategies may work on reducing different forms of bias to a greater or lesser extent. In sum, our goals were to test the effectiveness of the self-linking strategy in research designed to illuminate whether different implicit bias reduction strategies uniquely reduce different types of implicit intergroup bias.

Reducing Implicit Biases

Originally thought to be resistant to change, there is a growing body of research suggesting that implicit biases in the form of trait (i.e., stereotypes) and evaluative (i.e., prejudice) associations are in fact malleable (for reviews see Blair, 2002; Lenton, Bruder, & Sedikides, 2009), at least in the short term. For example, imagining counterstereotypic exemplars such as a strong woman (Blair, Ma, & Lenton, 2001), exposure to counterstereotypic exemplars (Dasgupta & Asgari, 2004), and extensive conditioning of counterstereotypes (Kawakami et al., 2000) have been shown to reduce bias on implicit stereotyping tasks. Implicit prejudice has been reduced through enrollment in a diversity seminar (Rudman, Ashmore, & Gary, 2001), reading about multiculturalism (Park, Felix, & Lee, 2007), exposure to Blacks in positive rather than negative contexts (Wittenbrink, Judd, & Park, 2001), and the use of approach behaviors (Kawakami, Phillips, Steele, & Dovidio, 2007).

An approach for potentially reducing implicit biases that has not been investigated previously involves forging an association or link between the self and outgroup members. This strategy is particularly important because it corresponds to naturally occurring situations that people can experience in everyday life. For example, the U.S. population is increasingly ethnically and racially heterogeneous, and legislation and norms are driving diversity efforts in schools, organizations, and professions. This changing landscape should promote heterogeneous groups and opportunities to associate former outgroup members with the self. Can the repeated association between the self and outgroup members encourage the reduction of preexisting and negative implicit intergroup biases?

Linking the Self With Outgroup Members

The idea that positive intergroup attitudes can be fostered by linking the self to outgroups through common ingroup membership has a long history in social psychology. The topic is relevant to many specific approaches to reducing prejudice, such as classic work on intergroup contact (e.g., Allport, 1954) and more contemporary work on the common ingroup identity model (Gaertner & Dovidio, 2000). Whereas this important research focuses on particular conditions of self-linking (e.g., common ingroup identity research involves forming a superordinate identity and working cooperatively under that identity), our interests are more basic. Specifically, we are interested in whether implicit intergroup biases can be reduced by merely linking the self to outgroup members
by defining a group to which the self and others belong.

Most relevant to our purposes is research based on the minimal group paradigm (e.g., Tajfel, Billig, Bundy, & Flament, 1971), where unfamiliar others are categorized into one’s ingroup based on arbitrary criteria. This categorization leads people to have more positive evaluations of members of their ingroup, relative to an arbitrarily defined outgroup. These effects can be understood in terms of positive feelings about the self generalizing to the ingroup members (Cadinu & Rothbart, 1996). In other words, because evaluations of the self are typically positive (Taylor & Brown, 1988) and people readily link ingroup members to the self (Smith & Henry, 1996), ingroup members become imbued with positivity (Greenwald et al., 2002).

The favorable effects of minimal group membership that have been observed with explicit measures extend to favorable implicit evaluations. Specifically, Ashburn-Nardo, Voils, and Monteith (2001) randomly assigned participants to an arbitrary ingroup that could be contrasted with an arbitrary outgroup. Following group assignment, participants demonstrated ingroup favoritism on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) such that they more easily associated pleasant words with their ingroup and unpleasant words with the outgroup. Ashburn-Nardo et al. argued that these results pointed to the preparedness of the human mind for forging connections that favor people who are associated with the self over outgroups.

Can this preparedness be capitalized on to encourage the reduction of preexisting implicit intergroup biases? That is, if outgroup members are linked to the self (e.g., Black people become members of a White person’s ingroup), will negative implicit intergroup biases be reduced? Unlike the favorable biases generated toward members of experimentally and arbitrarily created ingroups in the laboratory, implicit biases held in the real world are often the result of long-standing semantic and evaluative associations (Baron & Banaji, 2006). Such implicit associations are particularly resistant to change, although they can be altered through a “slow learning” process of repeatedly pairing new associations with the attitude object (Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Smith & DeCoster, 1998).

Given this past research, we reasoned that forging a connection between the self and outgroup members about which negative implicit biases already exist may reduce implicit intergroup bias when those associations are repeatedly practiced. That is, repeatedly identifying outgroup members as belonging to one’s own group should result in the transfer of positive self-associations to outgroup members and the conditioning of those associations, thereby reducing implicit intergroup bias.

It is important to note that this self-linking strategy differs from many previously investigated strategies for shifting intergroup biases in that it does not involve efforts to influence the traits or feelings people associate with the outgroup directly (e.g., strengthening the link between “Black” and “intelligence”). Also, the self-linking strategy we used did not blur traditional group boundaries by having Whites focus on similarities between Whites and Blacks. Such a strategy was recently used by Hall, Crisp, and Suen (2009), who found that writing about overlapping characteristics between Whites and Blacks. Such a strategy was recently used by Hall, Crisp, and Suen (2009), who found that writing about overlapping characteristics between ingroups and outgroups reduced implicit intergroup bias. Instead, the self-linking strategy we investigated involved building and conditioning a mere association between the self and outgroup members.

Comparing Strategies and Type of Implicit Intergroup Bias

A perusal of the extant implicit bias reduction literature shows a common approach where researchers devise a bias reduction strategy and examine its effects on one form of implicit bias (i.e., the automatic activation of either implicit stereotypes or implicit prejudice). This approach begs the question of whether different types of
implicit bias reduction strategies work similarly in reducing both implicit stereotyping and implicit prejudice.

Current theorizing (Gawronski & Bodenhausen, 2006; Greenwald et al., 2002) holds that evaluative associations (such as implicit prejudice) can depend on the sum total of the valence of activated semantic associations (such as stereotypic traits) (but see Amodio & Devine, 2006; Wolsko, Park, Judd, & Wittenbrink, 2000, for different perspectives). Accordingly, changes in semantic associations should result in corresponding changes to evaluative associations. For example, in a counterstereotyping paradigm (Kawakami et al., 2000), participants might repeatedly activate traits like intelligent and motivated in relation to Blacks. This activity should reduce implicit stereotyping for relevant traits. In addition, the overall implicit evaluation of Blacks may become more positive because the activated traits are positive. Indeed, researchers have found that counterstereotyping reduces both implicit stereotyping and implicit prejudice (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008), although not in a single study holding constant various procedural variables between experiments. We included a counterstereotyping condition in the present research, and we expected it to reduce both implicit stereotyping and prejudice relative to a control condition.

In contrast to conditioning alternative trait associations that are more positive than traditional stereotypes of Blacks, our self-linking strategy involved having White participants repeatedly categorize a few Black individuals as members of “MY GROUP.” We reasoned that conditioning a link between the (positively evaluated) self and the outgroup members should effectively condition positive evaluations of the group members and, therefore, reduce implicit prejudice, relative to a control condition. We further expected the conditioning-to-self procedure to reduce implicit prejudice to the same extent as the counterstereotyping procedure. However, the self-linking strategy does not work specifically on altering the patterns of trait associations or stereotypes. We therefore expected less success at changing specific implicit stereotypes than in the counterstereotyping condition.

**Experiment 1**

**Method**

**Participants.** One hundred and eighty White introductory psychology students (61% female) from a large Midwestern public university participated in return for partial course credit. Participants were preselected on the basis of ethnicity. Participants also completed the Negative Prejudice Scale (Czopp & Monteith, 2006) during mass testing at the beginning of the semester, so we could examine the possible moderating role of explicit bias.

**Design.** This study followed a 2 (implicit bias: stereotyping vs. prejudice) × 3 (strategy condition: counterstereotyping vs. link to self vs. control) mixed model design with repeated measures on the first factor.

**Procedure.** Participants signed up for a study called “Brain Power: Cognitive Training to Enhance Decision Making.” They came to the lab in groups of up to eight, were greeted by a non-Black experimenter, and were randomly assigned by session to one of the three strategy conditions (counterstereotyping, or CSA; link to self, or SELF; or the control condition). After giving written consent, participants were told that they would first complete a computer training exercise. Participants completed their assigned strategy task (i.e., CSA, SELF, or control), a short pencil-and-paper word-completion filler task, and a measure of implicit stereotyping followed by a measure of implicit prejudice. Lastly, participants answered an open-ended question that asked them to describe the purpose of the study, and they rated their interest in the strategy task on a 1 (not at all interesting) through 7 (extremely interesting) scale. Participants were then debriefed, thanked, and dismissed.

**Strategy conditions.** Each strategy task was administered on a personal computer and involved 408 trials, divided into six blocks. In each block participants saw a stimulus screen that showed a target picture in the center of the screen and two response words (one at the bottom left- and
the other at the bottom right-hand corner of the screen). Participants responded to each trial by pressing a key on the keyboard (either the “A” key for left or the “5” key on the number pad for right) that corresponded to the correct word choice as quickly as possible. In each strategy condition, the word “ERROR” appeared in red letters on the screen if the incorrect response was selected and remained until the correct response was selected.

The CSA strategy involved the repeated conditioning of counterstereotypic associations of Blacks and Whites, specifically targeting the stereotype of Whites being more intelligent than Blacks, and Blacks being more athletic and rhythmic. Participants were told that they were completing a computer task designed to give them practice at associating counterstereotypic words with members of specific social groups. They were given a list of 20 counterstereotypic words to review. Half of the words were White counterstereotypes (such as ATHLETIC and DANCE) and the other half were African American counterstereotypes (such as BRAINY and MATH). All of the words were intentionally positive to avoid a possible confound of presenting positively valenced Black counterstereotypes and negatively valenced White counterstereotypes. Following Kawakami, Dovidio, and van Kamp (2005), a photograph randomly selected from a pool of 38 White and 38 African American male faces was paired with two randomly selected words (one stereotypic and the other counterstereotypic of either African Americans or Whites). For each trial, participants were instructed to select the word that was not culturally associated with the ethnic group of the person in the photograph as quickly as possible. For example, when a picture of an African American was on the screen with the words “SCIENTIST” and “BASKETBALL” the correct response would be “SCIENTIST.”

The SELF strategy involved two phases: the formation of a diverse group, and the reinforcement of the link between participants and members of their new ingroup, the majority of which were Black. Following Ashburn-Nardo et al. (2001), participants first rated their liking of a series of paintings with two supposedly distinctive artistic styles. They were given fixed feedback to indicate that their strong preference for one style of art over the other indicated that they had a “bottom-up” cognitive processing style. Participants were told that what type of processor they were was important for completing the tasks in the rest of the experiment. Participants were then led to believe that the study was being conducted at a number of colleges and the researchers were amassing a large database of participants categorized by their processing style. To enhance the credibility of the cover story, participants were told that they would be asked later for their permission to be photographed and added to the database.

The participants were shown an array of 70 photographs of supposed former participants in the database and watched the computer conduct a “matching” process that selected five people who were supposedly also “bottom-up” processors. Each participant was assigned to a group of two Black females, two Black males and one White male (six members including the participant). Participants were informed that they would be completing decision-making tasks later, and that their scores would be combined with their group members’ scores to draw conclusions about the effects of cognitive style on decision-making. To strengthen the link between the participants and the members of the diverse group, they were instructed to study the photographs of their group members to be able to quickly and accurately identify them. Participants were told that being able to quickly and accurately identify the other members of their group would be important for their performance on the tasks they would complete later. Next, membership of the group was reinforced by having participants’ complete 408 computer trials where they saw a series of both White and Black faces on the computer screen and had to decide whether each person was a member of their group as quickly and accurately as possible. This training resulted in participants selecting “MY GROUP” for the Black group members 163 times, and for the White member of their group the remaining 41 times. Participants also selected the “OTHER GROUP” option for
149 trials containing White faces and 55 trials with Black faces.

The control condition was designed to be similar to the other strategy conditions in terms of number of trials and level of interest, but unrelated to race. Participants were told they were completing “environmental orientation training” to assess whether they were more oriented to natural or built items in their physical environment. Participants were presented with a series of photographs of natural items (such as a bird or a tree), built items (such as a brick wall or a bridge), and ambiguous items (such as a nest or a stone structure) with the words “BUILT” or “NATURE” appearing on the bottom left and right corners of screen. Participants were instructed to identify whether the images on the screen were built or natural objects as quickly and accurately as possible.

Implicit bias tasks. Implicit prejudice was assessed with an Implicit Association Test (IAT; Greenwald et al., 1998) designed to measure implicit racial evaluative preferences for Whites versus Blacks (referred to as the pIAT hereafter). The IAT is a dual categorization task that assesses implicit bias by the degree to which participants respond more quickly to stimuli presented in “congruent” trials (where the categories are Black/unpleasant and White/pleasant) compared with “incongruent” trials (where the categories are Black/pleasant and White/unpleasant). Following the standard IAT procedure, participants were presented with faces of African American and White males and words relating to pleasant concepts such as “RAINBOW” and “DIAMOND” and unpleasant concepts such as “VOMIT” and “FILTH” and required to categorize them into their appropriate category—Black, White, pleasant, or unpleasant. Stimuli were presented across 200 trials in seven blocks that included warm-up trials to become familiar with the process. Of specific interest are 20 practice and 40 test trials where participants were instructed to quickly and accurately categorize the stimuli into stereotypically congruent paired categories (Black/unpleasant vs. White/pleasant) and incongruent paired categories (Black/pleasant vs. White/unpleasant).

Implicit stereotyping was assessed with a stereotyping IAT (sIAT). Following Amodio and Devine (2006), participants were presented with words relating to physicality, specifically tapping the concepts of athleticism and dance such as “AGILE” and “JUMP,” and words relating to mental or academic concepts such as “MATH” and “BRAINY.” The sIAT used the same African American and White male faces as the pIAT and was procedurally identical to the pIAT with the exception of the category labels and stimuli words. On the congruent trials of the sIAT participants classified stimuli words and faces into either the Black/physical or the White/mental category. On the incongruent trials, participants classified stimuli into either the Black/mental or the White/physical category.

Results

Preliminary Analyses

We conducted preliminary analyses to verify that our strategy conditions did not differ significantly in terms of participants’ level of interest, comprehension of the stated reason for the task, or arousal of suspicion. Examination of the interest ratings indicated a moderate level of overall task interest (M = 4.02, SD = 1.44) and no strategy task effect, F(2, 168) = 0.11, p = .89. No participants were dropped from analysis for reporting suspicion or failing to comprehend the stated purpose of the training task. Additionally, none of the results varied significantly by participant sex, so sex was dropped from subsequent analyses.

Implicit Prejudice and Stereotyping

IAT effect scores (D scores) were computed for both the sIAT and the pIAT using the improved procedure provided by Greenwald, Nosek, and Banaji (2003), whereby higher D scores indicate greater negative implicit bias toward African Americans compared with Whites. The sIAT and pIAT D scores were not significantly correlated
\( r(175) = .091, p = .231 \) (see Table 1 for correlation matrix). A 2 (IAT type: pIAT vs. sIAT) \( \times 3 \) (strategy condition: CSA, SELF, and control) mixed model ANOVA with repeated measures on IAT type revealed a main effect for strategy, \( F(2, 172) = 4.01, p = .02 \), partial \( \eta^2 = .05 \), such that less bias was found in the CSA (\( M = .23 \)) and SELF (\( M = .25 \)) conditions than in the control condition (\( M = .35 \)). The Strategy \( \times \) IAT Type interaction was not significant, \( F(2, 172) = 0.20, p = .82 \). Given our focused predictions that both the CSA and SELF conditions would differ significantly from the control condition for the pIAT but only the CSA condition would differ significantly from the control condition for the sIAT, we conducted specific tests of cell differences despite the fact the Strategy \( \times \) IAT Type interaction was not significant. However, given the overall interaction was not significant, these analyses are admittedly exploratory and should be interpreted with caution. The means and standard deviations of the implicit bias scores as a function of strategy are shown in Table 2. Cell comparisons revealed that pIAT scores were lower in both the CSA and SELF conditions than in the control condition, with the effects just missing conventional levels of significance CSA versus control, \( t(116) = 1.91, p = .056, d = .35 \); SELF versus control, \( t(117) = 1.90, p = .06, d = .35 \). Hence, both an experience that involved the repeated affirmation of counterstereotypic associations of Blacks, and an experience that involved the formation and conditioning of a common ingroup including Blacks, appeared to reduce implicit negative evaluations of Blacks relative to Whites at a marginally significant level.

As expected, sIAT scores were lower in the CSA condition compared with the control condition, with a marginally significant effect, \( t(116) = 1.94, p = .054, d = .36 \). Thus, practicing a strategy that was specifically geared toward reducing the stereotypic associations targeted in the IAT task reduced the level of those associations to a degree approaching statistical significance. In contrast, forging links between outgroup members and the self did not significantly affect physical/mental stereotypic associations, as evident in a nonsignificant difference between the SELF and control conditions, \( t(116) = 1.23, p = .22, d = .23 \).

### Table 1. Bivariate correlations between implicit and explicit bias measures (Experiment 1).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Total</th>
<th>Counterstereotype affirmation</th>
<th>Control</th>
<th>SELF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pIAT</td>
<td>sIAT</td>
<td>pIAT</td>
<td>sIAT</td>
</tr>
<tr>
<td>sIAT</td>
<td>.091</td>
<td>.141</td>
<td>.033</td>
<td>.014</td>
</tr>
<tr>
<td>Explicit Prejudice</td>
<td>.194†</td>
<td>.101</td>
<td>.251†</td>
<td>.217†</td>
</tr>
</tbody>
</table>

*Note. sIAT = stereotype IAT; pIAT = prejudice IAT.*

†= \( p < .01 \).

### Table 2. Prejudice and stereotyping IAT D scores as a function of strategy condition (Experiment 1).

<table>
<thead>
<tr>
<th>Strategy condition</th>
<th>pIAT</th>
<th>sIAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N )</td>
<td>Mean</td>
</tr>
<tr>
<td>Counterstereotype Affirmation (CSA)</td>
<td>58</td>
<td>.26*</td>
</tr>
<tr>
<td>Control</td>
<td>60</td>
<td>.37b</td>
</tr>
<tr>
<td>Link With Self (SELF)</td>
<td>59</td>
<td>.26*</td>
</tr>
</tbody>
</table>

*Note. Means within a column that do not share a common subscript differ at \( p < .05 \).
Moderation Based on Explicit Prejudice Toward Blacks

Participants’ explicit prejudice scores obtained during the semester’s mass testing were decently distributed for examining their potential moderating effect ($M = 3.21$, $SD = .95$, on a 1–7 scale). We were interested in whether the effects of our implicit bias reduction strategies depended on the extent to which participants’ consciously held negative attitudes in relation to Blacks. For example, perhaps only participants whose explicit attitudes were relatively low in prejudice would be responsive to the CSA and SELF strategies. This issue was examined in regression analyses using explicit prejudice scores (centered), strategy condition (effects coding: SELF: $-1$, $-1$, CSA: $1$, $-1$, vs. control: $0$, $2$), and their interaction to predict implicit prejudice and implicit stereotypes. Regression analyses revealed no moderating effects of explicit prejudice toward Blacks on the impact of our strategies on either pIAT or sIAT $D$ scores. Thus, the two strategies appeared to be effective regardless of participant’s prejudice-related attitudes.

Discussion

The primary aim of Experiment 1 was to test the efficacy of the formation of a diverse group in the reduction of implicit bias. We compared the magnitude and nature of implicit bias reduction to a well-established bias reduction strategy—counterstereotype affirmation conditioning. Looking first at the impact of forging links between members of the outgroup (in this case Blacks with Whites) with the self on implicit bias reduction, in support of our hypotheses, the repeated conditioning of this link significantly reduced implicit prejudice relative to the control condition (although the effect was marginal, $p = .06$). In contrast, implicit stereotyping of Blacks as “physical” and Whites as “mental” was not significantly reduced in the SELF condition compared to the control condition. This effect was anticipated given the SELF strategy does not target these specific associations.

Replicating past research, participants who practiced counterstereotyping showed lower levels of implicit stereotyping than participants in a control condition (with $p = .054$). Interestingly, the extensive and intensive conditioning of counterstereotypical associations of Blacks and Whites and intelligence and physicality, respectively, also reduced implicit prejudice (with $p = .056$), and as much as in the SELF condition. This finding is consistent with the argument that evaluative associations reflect the sum total of specific semantic associations (Greenwald et al., 2002). The finding extends prior research showing the efficacy of counterstereotyping in the reduction of the implicit stereotypical associations targeted by the training (e.g., Kawakami et al., 2000), and suggests that targeting specific stereotypic associations also has an effect on evaluative associations (see also Gawronski et al., 2008).

We think it is also important to note that the strategies’ effects on implicit bias reduction were independent of participants’ levels of explicit prejudice in relation to Blacks. These results suggest that relatively high- as well as low-prejudice people are responsive to the bias-reducing effects of counterstereotype affirmation and the creation of links between the outgroup and the self.

Experiment 2

Experiment 2 was designed to replicate and extend Experiment 1. We considered a replication particularly important because, although the SELF strategy was not associated with a significant reduction in implicit stereotyping, as expected, there was some trend in this direction. This likely was why the expected interaction between strategy and IAT type was not significant. In addition, we investigated the effect of the SELF and CSA strategies on underlying processes by using the process dissociation procedure (PDP, Payne, 2001).

The PDP was originally developed by Jacoby (1991) as a way to distinguish between the extent to which conscious and unconscious processes influence task performance. Any measure of task performance can include both conscious and
unconscious influences, and this is true for what social psychology often casually calls “implicit tasks,” including the IAT. Thus, IAT performance can reflect both automatic (e.g., the automatic activation of biased associations) and controlled processes (e.g., effortful processing to control biased responding), and the PDP allows us to estimate the extent each process is associated with task performance. Specifically, the PDP yields (1) an automatic estimate (A), which is an index of how much automatically activated associations bias responses, and (2) a controlled estimate (C), which is an index of participants’ success at carrying out their intentions and correctly classifying the stimuli presented.

The question of present interest is whether differences in IAT scores between our experimental conditions and the control condition correspond to different degrees of automatic and controlled processing when performing the IAT. For example, pIAT scores in the CSA condition tended to be lower than in the control condition in Experiment 1. Does this result correspond with lower automatic processing (A estimate) in the CSA condition than in the control condition, or does it correspond with greater controlled processing (C estimate) in the CSA condition than in the control condition? Some implicit bias-change strategies appear to produce their effects by increasing individuals’ ability to draw on controlled processing (e.g., the use of implementation intentions, see Stewart & Payne, 2008). In contrast, our theoretical perspective on how both the CSA and SELF strategies work is that they decrease the likelihood of automatic activation of biased associations. Thus, we expected the strategies to be associated with reduced A rather than increased C.

Because PDP analyses are performed on error data, one’s measures must yield a sufficient number of errors to be useful in analyses. The error rates for the implicit prejudice and stereotyping tasks used in Experiment 1 were too low to permit meaningful PDP analyses (3.8% and 3.2%, respectively). Thus, we introduced a response window on the implicit bias tasks in Experiment 2 to increase errors on the IAT measures and enable PDP analyses.

Method

Participants. One hundred and fifty-eight White introductory psychology students from a large Midwestern public university participated in return for partial course credit. Six participants denied permission to use their data, one participant experienced persistent computer problems, and one participant failed to complete the experiment, leaving a final sample of 150. The mean age of participants was 19, and females made up 41% of the sample. Explicit prejudice scores were again available from the mass testing based on responses to the Negative Prejudice scale (Czopp & Monteith, 2006). In addition, participants completed Plant and Devine’s (1998) Internal and External Motivation to Respond Without Prejudice scales (IMS and EMS) during the mass testing.

Design. This experiment followed a 2 (implicit bias: stereotyping vs. prejudice) × 3 (strategy condition: CSA vs. SELF vs. control) mixed model design with repeated measures on the first factor.

Procedure. The same general procedure as in Experiment 1 was followed. Exceptions were that the filler task between the strategy manipulation and implicit measures was omitted, we modified the IATs by increasing the number of test trials from 40 congruent and 40 incongruent trials to 50 of each, administering the pIAT followed by the sIAT (as the reverse order was used in Experiment 1), and introducing a response window of 650 ms on each trial. The response window was implemented such that, if participants did not respond within 650 ms, they saw the message “Too slow!” after they responded. Even response windows as short as 500 ms that cut off participants’ opportunity to respond once the window was exceeded have effectively allowed for the detection of controlled as well as automatic processes on the PDP (e.g., Schlauch, Lang, Plant, Christensen, & Donohue, 2009).
Results

Implicit Prejudice and Stereotyping

Implicit Association Tests. IAT D scores were computed using the procedure outlined in Experiment 1. The sIAT and pIAT D scores were positively related, \( r(125) = .179, p = .046 \) (see Table 3 for correlation matrix). A 2 (IAT type: pIAT vs. sIAT) × 3 (strategy condition: CSA, SELF, and control) mixed model ANOVA with repeated measures on IAT type revealed a significant main effect for strategy, \( F(2, 122) = 3.58, p = .03 \), partial \( \eta^2 = .06 \), qualified by a significant Strategy × IAT Type interaction, \( F(2, 122) = 4.46, p = .01 \), partial \( \eta^2 = .07 \). The cell means and standard deviations are shown in Table 4. Simple effects tests revealed significantly lower pIAT D scores in the CSA condition compared with the control condition, \( t(81) = 3.36, p = .001, d = 1.23 \). Similarly, pIAT D scores were significantly lower in the SELF condition than in the control condition, \( t(91) = 2.94, p = .01, d = .53 \). Note as well that the means in our two experimental conditions were similar, indicating that the procedures reduced implicit bias to the same extent.

Turning to the sIAT scores, the CSA condition mean unexpectedly did not differ significantly from the control condition, \( t(82) = 1.43, p = .16 \), although the trend was in the same direction as in Experiment 1. Furthermore, the effect size \( (d = .31) \) was similar in magnitude to that obtained in Experiment 1 \( (d = .36) \). Consistent with predictions, the magnitude of sIAT bias was comparable in the SELF and control conditions, \( D = .32 \) and .33 respectively.

Meta-Analysis of IAT Scores. To provide an overall summary of the effects of the CSA and SELF strategies on pIAT and sIAT across the two experiments, we meta-analyzed the data. Eight effect sizes \( (d; \text{Cohen, 1988}) \) were calculated, assessing the impact of the CSA strategy compared with the control condition, and the impact of the SELF strategy compared with the control condition, for pIAT and sIAT, and for each experiment. The weighted mean effect size was then calculated across the two experiments for each of

<table>
<thead>
<tr>
<th>Measure</th>
<th>Link with self</th>
</tr>
</thead>
<tbody>
<tr>
<td>sIAT</td>
<td>.211**</td>
</tr>
<tr>
<td>pIAT</td>
<td>.199</td>
</tr>
<tr>
<td>NP</td>
<td>.001</td>
</tr>
<tr>
<td>IMS</td>
<td>.057</td>
</tr>
<tr>
<td>EMS</td>
<td>.017</td>
</tr>
<tr>
<td>Note: sIAT = stereotype IAT; pIAT = prejudice IAT; NP = Negative Prejudice scale; IMS = Internal Motivation to Respond Without Prejudice scale; EMS = External Motivation to Respond Without Prejudice scale.</td>
<td>** = ( p &lt; .01 ); *** = ( p &lt; .001 ).</td>
</tr>
</tbody>
</table>
Table 4. Prejudice and stereotyping IAT D scores as a function of strategy condition (Experiment 2).

<table>
<thead>
<tr>
<th>Strategy condition</th>
<th>pIAT</th>
<th>sIAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N  Mean SD</td>
<td>N  Mean SD</td>
</tr>
<tr>
<td>Counterstereotype Affirmation (CSA)</td>
<td>45 .46a .35</td>
<td>42 .21a .38</td>
</tr>
<tr>
<td>Control</td>
<td>38 .74b .42</td>
<td>42 .33a .38</td>
</tr>
<tr>
<td>Link With Self (SELF)</td>
<td>55 .49a .50</td>
<td>52 .32a .43</td>
</tr>
</tbody>
</table>

Note. Means within a column that do not share a common subscript differ at p < .05.

the four effects of interest. For the pIAT, we found that the comparison of the CSA and control conditions yielded a medium effect size of \( d = .50, Z = 3.41, p < .001 \) (95% confidence interval of .21 to .81). The comparison of the SELF and control conditions on the pIAT also yielded a significant effect size of \( d = .42, Z = 2.95, p < .01 \) (95% confidence interval of .41 to .70). Turning to the sIAT, a significant effect size emerged for the CSA versus control comparison, \( d = .34, Z = 2.37, p < .01 \) (95% confidence interval of .06 to .62). Finally, the SELF versus control condition comparison for the sIAT was very small and not significant, \( d = .14, Z = 1.05, p = .15 \) (95% confidence interval of −.13 to .42). In sum, these results are consistent with our expectation that counterstereotyping would reduce both implicit stereotyping and prejudice, but that forging links between the self and outgroup members would be effective at reducing implicit prejudice only.

We also conducted tests to compare the sizes of the effects from Experiment 1 to the sizes of the effects from Experiment 2 to ensure that they did not differ significantly. For example, finding that the effect size for the comparison between the CSA and control conditions in Experiment 1 is comparable to the effect size for the comparison between these two conditions in Experiment 2 would indicate similarity across the experiments. This was accomplished by first computing eight independent \( r_s \) (i.e., for CSA vs. control and SELF vs. control for both types of IATs and in both experiments), and testing the significance of difference between \( r_s \) as a function of experiment (e.g., \( r \) for the CSA vs. control comparison on the sIAT in Experiment 1 vs. \( r \) for the CSA vs. control comparison on the sIAT in Experiment 2). None of these tests yielded \( Z \) scores that approached significance, \( .16 < Z < 1.33 \), suggesting comparable results across the studies.

Process dissociation. The relative contribution of controlled and automatic processes to IAT performance in Experiment 2 was explored using the process dissociation procedure (PDP), which decomposes responses into controlled and automatic estimates (see Payne, 2001). In the case of the IAT, the controlled estimate (C) represents correct perceptual processing and responding in accordance with intentions (i.e., successfully classifying the target object or word) and is calculated as

\[
\text{# of correct responses on congruent trials} - \frac{\text{# of incorrect responses on incongruent trials}}{1 - C}
\]

where congruent trials are when Black is paired with unpleasant and White is paired with pleasant, and incongruent trials are when Black is paired with pleasant and White is paired unpleasant. One can think of this difference score as reflecting how well participants perform when there is no incongruency to deal with minus how poorly they perform when their responses require that they overcome incongruency.

With this estimate of control, the automatic estimate (A) can be solved:

\[
\frac{\text{# of incorrect responses on incongruent trials}}{1 - C}
\]

Thus, the A estimate reflects how often participants respond incorrectly when Blacks are paired with pleasant and Whites are paired with unpleasant relative to performance when control has failed. In this way the A estimate indexes the extent to which race-based associations have influenced performance.

The overall error rate on the IATs was 16%, which was sufficient to apply the PDP. Automatic and controlled estimates for the pIAT data for participants in each strategy condition were calculated. Separate ANOVAs were conducted on the C and A components of the pIAT data to
determine the impact of our strategy manipulation. Analyses revealed no significant effect of strategy condition on C, \( F(2, 135) = 0.48, p = .62 \), partial \( \eta^2 = .007 \), but did reveal the predicted significant effect on A, \( F(2, 135) = 5.27, p < .01 \), partial \( \eta^2 = .07 \). Specifically, as shown in Table 5, automatic estimates were significantly lower in the CSA condition compared with the control condition, \( t(81) = 3.14, p < .01, d = .69 \), and also significantly lower in the SELF condition compared with the control condition, \( t(91) = 2.14, p < .05, d = .44 \). Therefore, we can conclude that both counterstereotype affirmation and forging a link between outgroup members and the self decreased the automatic activation of biased associations, rather than increasing control.

As expected, the effect of strategy condition on C estimates for the sIAT was not significant, \( F(2, 132) = 0.59, p = .61 \), partial \( \eta^2 = .009 \). The strategy condition main effect also was not significant in the analysis of the A estimates, \( F(2, 132) = 1.01, p = .37 \), partial \( \eta^2 = .015 \). Further probing showed that, as expected, the A estimate in the CSA was lower (.57) than in the control condition (.59), but this difference was not significant, \( t(82) = 1.04, p = .30, d = .23 \). The fact that the sIAT yielded lower bias effects to begin with may account for the rather weak effect of the counterstereotype training on the A estimate. This may be related to the particular stereotypes assessed in the present research, a point to which we return in the General Discussion section.

### Table 5. Automatic and controlled estimates for the implicit prejudice measure (Experiment 2).

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Controlled estimate (C)</th>
<th>Automatic estimate (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Counterstereotype</td>
<td>45</td>
<td>.70a</td>
<td>.14</td>
</tr>
<tr>
<td>Affirmation (CSA)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>38</td>
<td>.69a</td>
<td>.13</td>
</tr>
<tr>
<td>Link With Self (SELF)</td>
<td>55</td>
<td>.67a</td>
<td>.16</td>
</tr>
</tbody>
</table>

*Note.* Means within a column that do not share a common subscript differ at \( p < .05 \).

### Discussion

Experiment 2 provided additional evidence that practicing counterstereotyping and conditioning a link between the self and outgroup members significantly reduced implicit prejudice. Furthermore, process dissociation analysis revealed that the reduction in implicit prejudice in both strategy conditions was associated with a reduction in the automatic activation of race-based associations. We found no evidence that either strategy affected participants’ ability to exert control over activated biases. These findings support our hypothesis that the bias reduction strategies examined here—conditioning specific counterstereotypical associations and the forging of links with former outgroup members—would affect the extent to which automatic processes biased participants’ responses.

Consistent with Experiment 1, the formation of a diverse group and conditioning links between the self (White) and Blacks did not reduce White participants’ implicit stereotyping. As for
counterstereotype training, we again observed the expected trend for it to reduce implicit stereotyping relative to the control condition, and meta-analysis of the sIAT data from Experiments 1 and 2 provided statistical support for this trend. However, it is noteworthy that the magnitude of the effect of the counterstereotyping on sIAT scores revealed in the meta-analysis (.34) was weaker than the magnitude of the effect of this strategy on pIAT scores (.50). Furthermore, although process dissociation analysis of the Λ estimates revealed a trend for race-based associations to influence sIAT less in the counterstereotyping condition relative to the control condition, this effect did not reach significance. These findings are rather surprising in light of the fact that participants received considerable practice affirming the counterstereotypes that were assessed in the sIAT. Why were our effects not stronger?

The answer may lie in the specific stereotypes targeted in the counterstereotyping procedure and the sIAT. Prior research showing reduced implicit stereotyping of Blacks following counterstereotyping targeted a range of positive and negative stereotypes in the training and implicit stereotyping tasks (Kawakami et al., 2000, Study 3). However, we focused on certain positive stereotypes only (i.e., physical and mental). The use of positive stereotypes was a deliberate attempt on our part to separate stereotypes from negative valence to effectively test our hypotheses. Implicit stereotyping effects for these positive stereotypes do not tend to be particularly strong (IAT mean D scores ranging from .15–.29 in Amodio & Devine, 2006, and ranging from .21–.33 in our experiments), so we had a tendency toward a floor effect. Despite this tendency, results across the two studies do provide support for the hypothesis that forging links with the outgroup operates to reduce implicit prejudice but not implicit stereotyping, whereas practicing counterstereotypic associations reduces both implicit prejudice and implicit stereotyping.

Finally, Experiment 2 provides further support for the efficacy of our strategies to reduce implicit biases regardless of explicit prejudiced attitudes. Across three different prejudice measures, we found no evidence of any moderating effect of explicit racial bias on the reduction of implicit bias.

**General Discussion**

Does the mere but repeated association of the self with outgroup members help to reduce pre-existing implicit biases against the outgroup? Theoretically, linking the self to outgroup members should cause positive associations with the self to extend to the outgroup (e.g., Greenwald et al., 2002), thereby increasing the favorability of implicit evaluative associations with the outgroup. The results from both of our experiments were consistent with this reasoning. White participants who formed a group including several Black individuals and practiced associating themselves and the Black group members subsequently scored lower on a prejudice IAT than participants in a control condition. The self-linking strategy fared well compared with another well-established strategy for reducing implicit bias in which participants repeatedly practiced counterstereotyping (e.g., pairing a Black person with “intelligent”; Kawakami et al., 2000). Despite radical differences between these two strategies, both reduced implicit prejudice to the same extent across our two experiments. Furthermore, both strategies appeared to exert their effects on implicit bias by decreasing participants’ automatic reliance on race-based associations. That is, relative to the control condition, both strategies significantly reduced automatic activation of biased associations, as revealed in process dissociation analyses. Unlike other strategies for altering implicit bias that rely on heightened controlled processing (Gonsalkorale, Sherman, Allen, Klauer, & Amodio, 2011; Stewart & Payne, 2008), the strategies we investigated reduced the extent to which race-based automatic processes contributed to responses.

Our self strategy involved placing Whites in a mixed-race group and repeatedly conditioning the link between the self and the mixed-group members. This procedure raises two questions. First, would preexisting negative implicit biases be
similarly reduced if participants were merely placed in a heterogeneous group, and no conditioning of the self–other link occurred? Recent research from our lab suggests that this would not be the case. Specifically, Gulker and Monteith (2012) found that linking a member of an outgroup to the self was associated with implicit bias reduction when participants repeatedly practiced the self–other link (Experiment 2). However, merely linking an outgroup member to the self—without conditioning this link—had no effect on implicit biases (Experiment 1). A second issue is whether mere exposure to a series of Black and White faces would facilitate the same reduction of implicit bias. There is converging evidence that this is unlikely to be the case. For example, van Bavel and Cunningham (2009) found that mere exposure to mixed-race stimuli in the absence of the creation of a superordinate group did not decrease implicit racial bias. In addition, there is a body of work that supports the notion that IAT is uninfluenced by familiarity with the stimuli (see Dasgupta, McGhee, Greenwald, & Banaji, 2000; Ottaway, Hayden, & Oakes, 2001).

We have assumed that the self-linking strategy reduced implicit prejudice because positive self associations transferred to the outgroup. Future research can address the mechanism more precisely. In the case of counterstereotyping, we have assumed that the positivity of the practiced counterstereotypes fed into the evaluative dimension (see Gawronski & Bodenhausen, 2006) so that implicit prejudice was reduced. If this is the case, practicing negative counterstereotypes about an outgroup (e.g., Asians as unintelligent) should result in greater implicit prejudice. Future research is needed to determine whether this is indeed the case.

We also examined the effects of the self-linking and counterstereotyping strategies on implicit stereotyping. In this case, as expected, we found divergent effects. Participants trained to affirm specific counterstereotypes of Blacks as “mental” (e.g., brainy) and Whites as “physical” (e.g., agile) subsequently scored lower on a relevant stereotyping IAT, relative to the control condition. In contrast, linking outgroup members to the self did not yield a reliable effect on the stereotyping IAT, which suggests that the strength of specific trait associations is unaffected even when evaluations overall become more positive. We acknowledge that both of the experiments used relatively positive rather than negative stereotypes of Blacks, and that the implicit activation of these stereotypes was perhaps not as strong as the typical activation of Whites’ negative implicit stereotypes of Blacks. This invites the speculation that the self-linking self strategy may reduce the automatic activation of strongly negative stereotypes due to spreading of the evaluative associations to trait associations, even if weaker positive stereotypes are unaffected. Future research is needed to investigate this issue, but we suspect that the reduction in implicit stereotyping through self-linking would not be as strong as the reduction that occurs when counterstereotyping is practiced on the same trait dimensions that are included in the implicit stereotyping measure.

Overall, we believe that our approach of comparing the effects of different training strategies across both implicit prejudice and stereotyping measures can provide unique insights into the landscape of implicit bias malleability, and we encourage others to use this approach. Some strategies for reducing implicit biases may affect primarily stereotypic associations, some may affect primarily evaluative (prejudice) associations, and some may affect both. Both consequences are very important, and they should have unique effects on subsequent behavior. For example, Amodio and Devine (2006) found that implicit stereotyping is especially likely to shape judgments and impression formation, whereas implicit evaluation affects affective processes that guide interpersonal preferences and social distance.

We would not argue that bias change strategies that reduce only one type of implicit intergroup bias should be abandoned in favor of strategies that affect both implicit stereotyping and implicit prejudice. Once inroads are made with successful bias reduction on one dimension of bias, avenues may be opened for reducing bias on another dimension. For instance, with increased positivity encouraged by self-linking, people may be more likely to pursue interactions with outgroup members and be exposed to individuating and
counterstereotypic information (e.g., Dovidio, Gaertner, & Kawakami, 2003). These experiences, such as in diverse groups settings in real-world settings, may ultimately decrease implicit stereotyping.

Related are recent findings reported by Riek, Mania, Gaertner, McDonald, and Lamoreaux (2010). Democratic and Republican participants were classified as a single group, “Americans.” This common identity was emphasized by having participants wear T-shirts with an emblem of the American flag, and they performed a cooperative and interdependent task that was to be compared to that of groups from other countries. Compared to participants for whom a common identity was not emphasized, participants in the common ingroup later rated the former outgroup as less stereotypic (e.g., Democratic participants rated Republicans less stereotypically). Thus, self-linking (albeit through a fairly elaborate procedure for establishing a common group identity) encouraged less explicit stereotyping. We would expect further group experiences reinforcing new patterns of outgroup associations to result eventually in reduced implicit stereotyping as well.

In addition to the need for more “multiple strategy/multiple implicit bias measure” research, the literature on implicit bias reduction needs to move toward examining longer term change, rather than just temporary shifts in implicit bias. Although there has been recent, rapid progress in identifying various contexts and experiences that can shift implicit biases (see Blair, 2002; Lenton et al., 2009), we currently know very little about whether these are temporary shifts or may result in longer term change. Furthermore, it is important that researchers translate their experimental manipulations into strategies that individuals can employ or encounter in their everyday lives. For example, while it is unlikely that people will stop by computer labs for several hundred trials of counterstereotyping to keep their implicit stereotyping in check, they may be taught to use a counterstereotyping strategy as they walk through daily life. We are currently investigating whether people can learn to use certain experimental laboratory strategies that we have modified for real-world settings to reduce implicit intergroup bias across time.

**Conclusions**

The potential for the self-linking self strategy to curb implicit biases could be far reaching, especially with increasingly diverse school, work, and organizational settings that presumably encourage heterogeneous ingroups. Also encouraging is the finding that the strategy appears equally effective among less as well as more egalitarian people (as we found no evidence of moderation by explicit prejudice or motivations to control prejudice). However, we suspect that neither self-linking nor any other type of implicit bias change strategy alone is likely to be sufficient for leveling implicit biases and maintaining positive change. The many motivational, cognitive, and sociostructural factors that encourage implicit biases in the first place are likely to have daily influences that will need to be countered with a multistrategy approach for instigating and sustaining equality in mental representations and concomitant outcomes.

**Acknowledgements**

This research was supported by a National Science Foundation grant (BCS 0921516) to the second author. Portions of this research were presented at the 2009 European Social Cognition Network conference and the 2010 meeting of the Midwestern Psychological Association. We thank Rhutba Ahad, Kevin Brett, Lisa Graham, Matt Gifford, Hannah Gonya, Jill Gulker, Courtney Schaffer, Margaret Yagelski, and Sisi Yu, for their help with data collection. Our sincere thanks to Keith Payne for advice on using the PDP with IAT data.

**Notes**

1. Three of the African American males faces used in the CSA strategy were also used in the SELF strategy, and none used in the CSA and SELF training were used in the IATs.
2. pIAT and sIAT data are missing for some participants due to a computer problem where the computer “froze” during data collection, and because scores were excluded when participants had more than 10% of trials completed in < 300 ms (see Greenwald et al., 2003). This was also the case in Experiment 2. These missing data account for differing df across some of our analyses.
References


