*Group Processes & Intergroup Relations*
G P I R

*Article*

# Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes

*Group Processes & Intergroup Relations*
1–23

DOI: 10.1177/1368430216647189
gpir.sagepub.com
SAGE

**India R. Johnson,[1] Brandon M. Kopp,[2] and Richard E. Petty[3]**

**Abstract**

The present research compared the effectiveness of meaningful negation—"That's wrong"—and simple negation—"No"—to alter automatic prejudice. Participants were trained to negate prejudice-consistent or prejudice-inconsistent information, using either simple or meaningful negation, and completed an evaluative priming measure of racial prejudice before and after training. No significant changes in automatic prejudice in the simple negation conditions emerged. In contrast, those trained to negate prejudice-consistent information in a more meaningful way showed a significant decrease in automatic prejudice, whereas those trained to negate prejudice-inconsistent information meaningfully showed a significant increase. Study 2 revealed that these effects were driven by participants high in their motivation to control prejudiced reactions (MCPR), as they demonstrated the greatest changes in automatic prejudice following training. Contrary to research suggesting negation training is an ineffective means to reduce automatic racial prejudice, the present research suggests negation can be effective when the negation is meaningful.

**Keywords**

negation training, prejudice, stereotyping

Paper received 29 July 2015; revised version accepted 25 March 2016.

In recent decades, volumes of research in the field of social psychology has focused on techniques to reduce automatic prejudice (see Blair, 2002; Devine, Forscher, Austin, & Cox, 2012; Olson & Fazio, 2006; Turner & Crisp, 2010). One such technique is negation training—or training participants to explicitly reject stereotype-relevant beliefs. Although initial research examining negation training found it to be an effective means of reducing automatic prejudice (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000), subsequent research suggested negation training was not only ineffective but could ironically, increase automatic prejudice (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). Consequently, the present work seeks to address this inconsistency in findings and

[1]Elon University, USA
[2]United States Bureau of Labor Statistics, USA
[3]The Ohio State University, USA

**Corresponding author:**
India R. Johnson, Department of Psychology, Elon University, Campus Box 2337, Elon, NC 27244, USA.
Email: ijohnson5@elon.edu

examines the conditions under which the negation of stereotype-relevant information can effectively reduce automatic bias. Specifically, the present research investigates if meaningful negation training—in which participants are trained to think "That's wrong!" in response to stereotype-relevant information—can successfully reduce automatic prejudice.

## Negation Training and the Reduction of Automatic Bias

Available research examining the utility of negation training is unclear on whether undermining existing negative associations via negation is an effective strategy to reduce automatic prejudice. In one of the first studies relevant to negating automatic prejudice, Kawakami et al. (2000) exposed participants to a picture of a Black face along with trait words related to the stereotype of Blacks such as *lazy*. The participants were instructed to hit a key on a button box to indicate "No" to all stereotype-consistent pairings. Within the same training task, participants would also see a Black face and a trait word counter to the stereotype of Blacks (e.g., smart) and were instructed to hit a key on a button box to indicate "Yes." Participants practiced this training for a total of 160 trials. Additionally, both before and immediately following training, all participants completed a person categorization task which served as the measure of automatic prejudice (Blair & Banaji, 1996). This task assessed the speed with which participants classified Black and White faces following the subliminal presentation of positive and negative trait words. Automatic activation of prejudice was reduced following the training exercise, consistent with the notion that negation of stereotype-consistent information can reduce automatic racial prejudice (see Kawakami et al., 2000). However, because participants were also trained to affirm stereotype-inconsistent information within the same training exercise, it is unclear whether negation alone was effective or whether affirmation of positive traits was the essential component of the procedure or whether both were necessary.

To identify which element of the training task was driving the observed reduction in automatic prejudice, Gawronski et al. (2008) extended the work of Kawakami et al. (2000) by examining negation and affirmation training separately within the same study. Gawronski and colleagues adopted a design in which participants were trained either to negate prejudice-consistent information (i.e., pressing a button to indicate "No" whenever a Black face was paired with a negative trait) or affirm prejudice-inconsistent information (i.e., pressing a button to indicate "Yes" whenever a Black face was paired with a positive trait).[1] As in the Kawakami et al. (2000) research, participants completed a measure of automatic prejudice before and after training. Only the affirmation training led to a reduction in automatic prejudice. In contrast, participants who engaged in negation training showed an *increase* in prejudice. Gawronski et al. (2008) suggested continually practicing negation would not necessarily stop negative traits from coming to mind and negation can, in fact, strengthen associations and make the link between the category and trait (e.g., Black and poor) stronger. Thus, the only study focused exclusively on saying "No" to prejudice suggests this technique is ineffective, or even worse, counterproductive. This conclusion is consistent with research in other domains on the general ineffectiveness of negations in changing existing beliefs (e.g., Gregg, Seibt, & Banaji, 2006; Petty, Briñol, Tormala, & Jarvis, 2006).[2]

Despite the research on the ineffectiveness of negation training, is there reason to believe negation of prejudice can ever be effective? One possibility is the negations tested in prior research may not have been especially salient and making them more salient might improve their effectiveness. For instance, in one study Boucher and Rydell (2012) had participants complete a learning task in which they were presented with information about a novel target. Following the completion of the learning task, attitudes towards the target were assessed using both an implicit and explicit attitude measure. During the learning task, some of the presented information was negated and the visual salience of the negation

was manipulated by presenting negation information in a standard, 24-point font size, or a much larger 48-point font size. Boucher and colleagues found implicit and explicit attitude measures only reflected the intended valence of the information presented when negations were visually salient (i.e., larger) and when participants had cognitive resources available. Although this study involved a paradigm in which attitudes were being formed rather than changed, this research nonetheless suggests highly salient negations can be more effective than less salient ones, at least when people have sufficient cognitive resources available to process. Thus, in the current research we will compare a simple negation with a potentially more powerful one. Unlike Boucher and Rydell (2012), however, we use negations to modify existing rather than novel attitudes.

In addition to the salience of the negation, the current research examines another possible moderator. Past work examining the motivation of individuals to respond without prejudice suggests negations are more likely to be impactful on automatic evaluations when people are sufficiently motivated. Specifically, a number of studies have highlighted the Motivation to Control Prejudiced Reactions Scale (MCPR; Dunton & Fazio, 1997), as having an important influence on automatic measures of prejudice (Allen, Sherman & Klauer, 2010; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Olson & Fazio, 2004). For instance, in one study Maddux, Barden, Brewer, and Petty (2005) examined the interaction of context and MCPR on automatic evaluative responses towards Blacks and Whites. They found Whites who were high in MCPR demonstrated an automatic outgroup bias in favor of Blacks over Whites in contexts in which the possibility of racial bias was salient (e.g., a photo of Blacks vs. Whites in prison). Furthermore, this result was driven by an automatic inhibition of negative responses towards Blacks as if those high in MCPR were automatically correcting for their prejudice because of considerable prior practice in negating their negative reactions (Wegener & Petty, 1997; see also, Livingston, 2008). Thus, this study points to individual differences in MCPR as possibly playing a moderating role in the effectiveness of negation training. We examine this possibility in Study 2.

## A New Look at Negation Training

The current research aimed to take a new look at the role of negation in undermining automatic prejudice. In particular, we suggest two reasons why prior research might have prematurely concluded negation training was ineffective in reducing individuals' preexisting automatic prejudice.

One reason is the negation used was not strong enough, or not meaningful enough. Past research has found processing negations requires ample cognitive resources (Boucher & Rydell, 2012; Gilbert, Tafarodi, & Malone, 1993), and when lacking, the meaning of a simple "No" can be ambiguous, thereby reducing its impact (Gilbert & Osborne, 1989). We predicted a more meaningful and powerful negation had the potential to remove any ambiguity and might be more effective in reducing automatic racial prejudice. For example, simply thinking "no" in response to a prejudiced reaction is plausibly less meaningful and impactful than thinking "That's wrong!" In the present work, we examined the utility of negating prejudice by having participants think "That's wrong!" and compared it to a simple "No." The former is not only unambiguous but also may convey a moral standard that is hard to ignore (Haidt, 2001).

As also noted earlier, a second factor possibly contributing to the ineffectiveness of negation in prior research is participants might have been insufficiently motivated to control for their prejudiced reactions. A number of studies have indicated people vary in this motivation (Dunton & Fazio, 1997; Plant & Devine, 1998), and negation training might be ineffective if people are not sufficiently motivated to reduce their prejudice. In both Studies 1 and 2 we compare the ability of negation training involving a simple versus more meaningful negation to undermine racial prejudice and in Study 2 we also examine the role of individual differences in motivation to control prejudice.

## Overview of the Present Research

In sum, in two studies, we compared the efficacy of two types of negation—a simple "No" versus a more meaningful "That's wrong!"—in altering automatic racial attitudes. Additionally, in each study we manipulated the type of information participants were instructed to negate. Half of the participants were instructed to negate prejudice-consistent information such as when a picture of an African American was paired with a prejudice-consistent term (e.g., lazy). The other half of participants were instructed to negate prejudice-inconsistent information, such as when a picture of an African American was paired with a prejudice-inconsistent term (e.g., smart). Although potentially training individuals to increase their automatic racial prejudice is not ideal from a societal point of view, we determined including such a condition was necessary for two reasons. First, the addition of this condition extends past research by examining an avenue through which individuals may potentially reaffirm their prejudicial beliefs. Second, and more conceptually relevant, it is possible simply having participants think "That's wrong!" in the context of a study about race would sensitize them not to be prejudiced toward Blacks and it would not matter what type of information they were negating. If sensitization to prejudice was involved, then it is possible thinking "That's wrong!" in the context of a study on race would reduce prejudice regardless of the direction of training. On the other hand, if meaningful negation works in both directions (socially desirable and undesirable), the results are more confidently attributed to negation rather than sensitization to prejudice. Consequently, we included conditions in which participants negated each type of information—prejudice-consistent and prejudice-inconsistent.

Based on prior negation research, we expected the simple "No" training conditions would be ineffective in altering automatic attitudes or might even produce the reverse effect. In contrast, we expected those assigned to the meaningful negation conditions to demonstrate a significant change in automatic racial attitudes. Specifically, we predicted those who meaningfully negated prejudice-consistent information would show a decrease in automatic prejudice following training, whereas those who meaningfully negated prejudice-inconsistent information would show an increase in prejudice.

## Study 1

### *Method*

*Participants and design.* One hundred and twenty-five undergraduates (70 female; 53 male, two unidentified) at a large Midwestern university in the US enrolled in introductory psychology participated in a study on "learning and memory" and received partial course credit in exchange for their involvement. Mean age of participants was 18.50 ($SD = 0.79$) years. All participants indicated they spoke English as a first language and the majority of our sample identified as White/Caucasian (83.2%).[3]

The core procedures were adapted largely from Gawronski et al. (2008), and was a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) × 2 (time of measurement: before vs. after) mixed-model factorial, where the first two variables were between-subjects and the latter was within-subjects. Sample size was determined based on previous research examining negation training and prejudice reduction (see Gawronski et al., 2008).

*Training task.* Upon arrival, participants were instructed to sit at a personal computer station of their choice, separated by a partition divider to provide participants privacy. Informed consent was obtained via the computer and all participants were informed no identifying information would be tied to any of their responses and they could leave the study at any time without loss of credit. Participants then read a short description of the experiment indicating they would be performing several tasks related to Black and White cultural stereotypes.

Participants in the prejudice-consistent training conditions were instructed to negate any face–trait combinations for which a Black face was paired with a negative trait word or a White

face paired with a positive trait word. For combinations that were prejudice-inconsistent, participants were instructed to do nothing and wait for the next face–trait combination to appear. Those in the simple negation conditions were told to hit the space bar and think "NO!" for prejudice-consistent combinations, while those in the meaningful condition were told to think "THAT'S WRONG!" (see Appendix A for exact instructions).

In contrast, participants in the prejudice-inconsistent training conditions were instructed to negate any face–trait combinations for which a Black face was paired with a positive trait word or a White face paired with a negative trait word—and do nothing for combinations that were prejudice-consistent. Again, those in the simple negation conditions were told to hit the space bar and think "NO!" for prejudice-inconsistent combinations, while those in the meaningful condition were told to think "THAT'S WRONG!" (see Appendix A for exact instructions).[4]

Participants in all conditions were presented with a total of 200 face–trait pairings, comprised of both prejudice-consistent and prejudice-inconsistent pairings. These pairings included 40 combinations of: (a) a Black face paired with a positive (prejudice-inconsistent) trait word, (b) a Black face paired with a negative (prejudice-consistent) trait word, (c) a White face paired with a positive (prejudice-consistent) trait word, and (d) a White face paired with a negative trait word (prejudice-inconsistent). Pairings of Black and White faces with positive and negative trait words were randomized by the computer for each participant. Ten Black and 10 White male faces were used as stimuli. Positive and negative trait words were the same as those used in Gawronski et al. 2008 (Study 2); see Appendix B–D for exact trait words and faces).

For each trial, participants were first presented with a picture of a Black or a White face in the center of the screen. After 500 milliseconds, a positive or negative trait word appeared just below the picture. Trait words were either related to a positive evaluation of White people or a negative evaluation of Black people. When participants correctly pressed the space bar, the stimuli disappeared and the next trial started. If participants incorrectly pressed the space bar in response, the stimuli were replaced by the message "ERROR!" which appeared in the middle of the screen for 1,500 milliseconds. If participants did not respond to a given combination, the stimuli disappeared after 2,500 milliseconds and the next trial started. The time between trials for all responses was 1,000 milliseconds. Consistent with Gawronski et al. (2008), all training tasks consisted of five blocks of 40 trials each, resulting in a total of 200 training trials. After each block, participants were asked to take a moment to relax before moving on to the next block.

*Automatic prejudice.* To assess automatic evaluations of Blacks and Whites, we used a subliminal evaluative priming task (see Fazio, Jackson, Dunton, & Williams, 1995) as it was adapted by Gawronski et al. (2008). In this task, each trial began with a fixation cross ("+") which was presented for 1,000 milliseconds in the center of the screen. The prime word "Black" or "White" was then presented for 15 milliseconds; followed by a masking stimulus ("XXXXX") for 250 milliseconds. The masking stimulus was then replaced by a positive or negative target word which remained on the screen until participants responded (see Appendix E for full list of target words).

Participants were instructed to press the "a" key as quickly as possible when they saw a positive word and the "l" key when they saw a negative word. Each of the 40 target words was presented twice with each of the two prime words, resulting in a total of 160 trials. Consistent with Gawronski et al. (2008), in order to assure independence of automatic prejudice and automatic stereotyping at the measurement level, positive and negative nouns were used as target words (e.g., candy, dirt) rather than the positive and negative trait words (e.g., friendly, lazy) used in the training task. Order of trials was randomized individually for each participant. Incorrect responses were indicated with the word "ERROR!" appearing in red for 1,000 milliseconds in the middle of the screen. Once the

"ERROR!" message left the screen or participants responded correctly, the fixation cross appeared and a new trial began.

The priming measure was scored by creating an index reflecting the overall automatic preference for Whites over Blacks (Gawronski et al., 2008). Raw latencies were used to create the index. The overall index was calculated by first subtracting the mean response latency to positive words after White primes from the mean response latency to positive words after Black primes (higher scores indicate stronger activation of positivity for White as compared to Black), and by subtracting the mean response latency to negative words after Black primes from the mean response latency to negative words after White primes (higher scores indicate stronger activation of negativity for Black as compared to White). Negativity scores were then added to positivity scores, resulting in an index of automatic preference, with higher scores indicating a stronger preference for Whites over Blacks. Participants completed the evaluative priming task twice; once immediately before training and once immediately after training.

Finally, participants completed demographic measures, were thoroughly debriefed, thanked, and dismissed.[5]

## *Results*

*Training task.* To ensure participants were successfully completing the training task, both reaction times and participants' errors across all five training blocks were examined. Participants' reaction time to negate a face–trait pairing was averaged across all trials to create a mean reaction time for each block. Percentage of errors made per block was also calculated. An error was defined as either negating a combination on a trial in which the participant was not to give any response, or by failing to hit the space bar for combinations which were to be negated. For example, for those participants assigned to the simple or meaningful prejudice-consistent conditions, an error would be classified as either hitting the space bar for combinations inconsistent with the cultural stereotype of Blacks and Whites, or failing to hit the space bar for combinations consistent with the cultural stereotype. Because each block consisted of 20 trials in which participants were instructed to hit the space bar in response and 20 trials in which they were instructed to do nothing, participants could potentially make anywhere from zero to 40 errors per block. Consequently, the number of errors of both types were totaled and divided by 40, and the resulting decimal converted to a percent to calculate the percentage of errors per block.

First looking at mean reaction times for each block, a 2 (training direction: prejudice- consistent vs. inconsistent) x 2 (negation type: simple vs. meaningful) × 5 (training block) mixed-model ANOVA, where the first two variables are between-subjects, revealed a significant main effect of block, $F(4, 484) = 47.86$, $p < .0001$, $\eta^2 = .28$, such that reaction times were slowest in the first block and became much faster after that (see Figure 1). A significant interaction of Block × Negation, $F(4, 484) = 2.85$, $p = .02$, $\eta^2 = .02$, also emerged. Examining this interaction revealed those in meaningful conditions were slower across all five blocks ($M = 905.50$, $SD = 234.31$), relative to those in the simple negation conditions ($M = 889.91$, $SD = 239.09$). This suggests participants were following instructions since it should take a bit longer to think "That's wrong" than "No." The Block × Training, $F(4, 484) = 1.69$, $p = .15$, $\eta^2 = .01$, and the Block × Training × Negation interactions, $F(4, 484) = 0.22$, $p = .92$, $\eta^2 = .002$, were nonsignificant (see Supplemental File A for means and standard deviations of reaction times across all blocks for each condition).

We next examined participants' errors across blocks. Participants' mean percentage of errors made for each block was submitted to a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) × 5 (training block) mixed-model ANOVA, where the first two variables are between-subjects, while the latter is within-subjects. Results revealed only a significant main effect of block, $F(4, 484) = 19.91$, $p < .001$, $\eta^2 = .14$, with participants making fewer errors each subsequent block, from Block 1 to

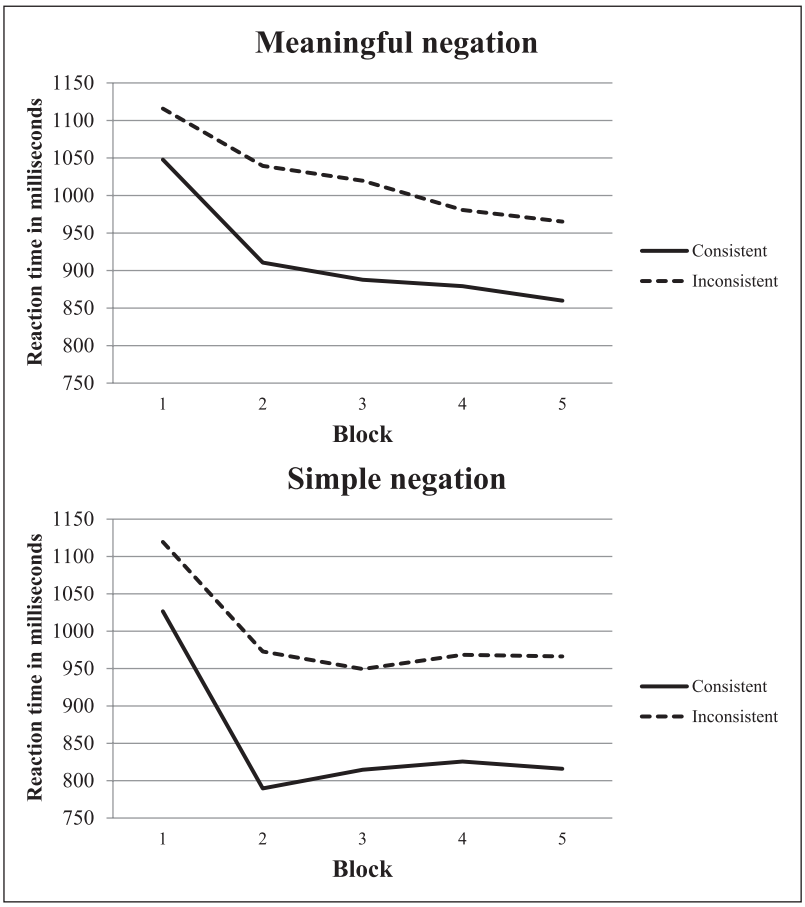


**Figure 1.** Reaction time in milliseconds for each block of the training task by training direction (prejudice-consistent vs. inconsistent) and negation type (meaningful [top panel] vs. simple [bottom panel]; Study 1).

Block 5. Errors made across all five learning blocks by condition are illustrated in Figure 2. No other significant differences in errors in Blocks 2–5 emerged and number of errors did not differ as a function of Block × Prejudice, $F(4, 484) = 0.22$, $p = .93$, $\eta^2 = .002$, Block × Negation, $F(4, 484) = 0.65$, $p = .63$, $\eta^2 = .005$, or Block × Training × Negation, $F(4, 484) = 0.34$, $p = .85$, $\eta^2 = .003$, across blocks (see Supplemental File B for means and standard deviations of percentage of errors across all blocks for each condition).

Taking participants' reaction times and errors in conjunction, despite the meaningful conditions being somewhat slower overall, given the number of errors did not differ significantly across type of negation suggests all participants learned the task equally well.[6]

*Automatic prejudice.* Common in analyzing data from evaluative priming tasks, prior to analyses, outliers were excluded by discarding responses faster than 300 ms (0.1% at Time 1; 0.6% at Time 2) or slower than 1,000 ms (7.8% at Time 1; 11.8% at Time 2; Fazio et al., 1995; Gawronski et al., 2008; Neely, 1977). Error trials were excluded from analyses as well (1.4% at Time 1; 1.3% at Time 2).

Turning to our primary hypotheses, participants' automatic prejudice scores were submitted to a 2 (training direction: prejudice-consistent vs.

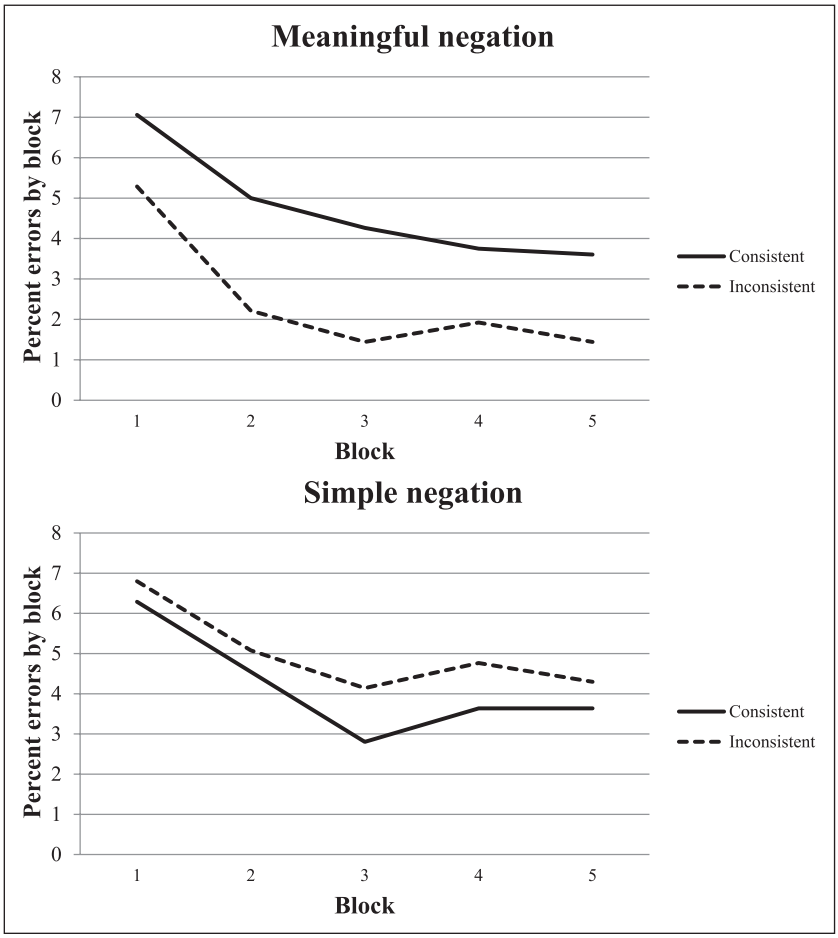


**Figure 2.** Mean percentage of errors for each block of the training task by type of training direction (prejudice-consistent vs. inconsistent) and negation type (meaningful [top panel] vs. simple [bottom panel]; Study 1).

inconsistent) x 2 (negation type: simple vs. meaningful) × 2 (time of measurement: before vs. after) mixed-model ANOVA, where the first two variables are between-subjects, while the latter is within-subjects. Results revealed a significant three-way interaction of Training × Negation × Time, $F(1, 121) = 13.54$, $p < .001$, $\eta^2 = .10$.

To examine the impact of meaningful versus simple negation, we decomposed this three-way interaction as a function of type of negation. Examining the meaningful versus simple negation conditions separately revealed a significant two-way interaction of Training × Time of Measurement, $F(1, 58) = 19.02$, $p < .001$, $\eta^2 = .25$, for the meaningful negation conditions (see top panel of Figure 3), that was absent for the simple negation conditions, $F(1, 63) = 0.20$, $p = .88$, $\eta^2 = .02$ (see bottom panel of Figure 3).

Differences in racial prejudice from Time 1 to Time 2 were then assessed by condition. A paired-samples $t$ test looking within condition from Time 1 to Time 2 demonstrated a significant decrease in automatic racial prejudice for the meaningful prejudice-consistent training condition, $t(33) = 3.95$, $p < .001$, 95% CI [16.19, 50.53], and a significant increase in automatic racial prejudice for the meaningful prejudice-inconsistent condition, $t(27) = -2.34$, $p = .03$,

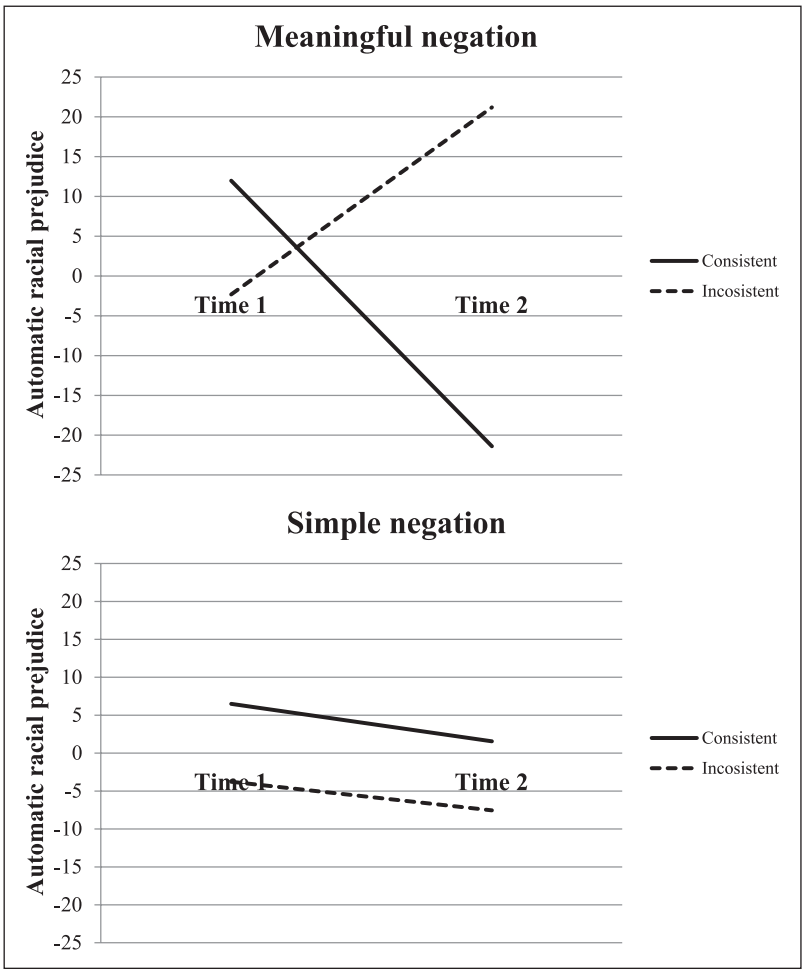


**Figure 3.** Mean scores of automatic racial prejudice as a function of training direction (prejudice-consistent vs. inconsistent) and negation type (meaningful [top panel] vs. simple [bottom panel]; Study 1).

95% CI [−44.17, −2.85]. In contrast, paired-samples *t* tests for the simple prejudice-consistent, $t(32) = 1.16$, $p = .25$, 95% CI [−3.73, 13.62], and the simple prejudice-inconsistent condition, $t(31) = 0.53$, $p = .59$, 95% CI [−10.67, 18.21] revealed no significant differences in automatic racial prejudice from Time 1 to 2.

Finally, to ensure this result was not due to participants differing in their levels of automatic racial prejudice at Time 1 (i.e., a failure of random assignment), we submitted participants' automatic racial prejudice scores at Time 1 to a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) ANOVA. The main effect of training, $F(1, 121) = 1.09$, $p = .29$, $\eta^2 = .009$, negation, $F(1, 121) = 0.34$, $p = .56$, $\eta^2 = .003$, and the interaction of the two, $F(1, 121) = 0.11$, $p = .73$, $\eta^2 = .001$, were all nonsignificant, demonstrating participants did not significantly differ at Time 1.

In contrast, examining participants' level of automatic racial prejudice at Time 2 revealed a significant main effect of training, $F(1,121) = 6.21$, $p = .014$, $\eta^2 = .05$, such that those who engaged in the negation of prejudice-inconsistent

**Table 1.** Mean scores of automatic racial prejudice as a function of type of negation (meaningful vs. simple) and training (prejudice-consistent vs. inconsistent) before (top) and after (bottom) training (Study 1).

| Training | Negation | | | |
|---|---|---|---|---|
| | Simple | | Meaningful | |
| | *M* | *SD* | *M* | *SD* |
| **Time 1: Before training** | | | | |
| Prejudice-consistent | 6.50 | 25.47 | 11.97 | 35.01 |
| Prejudice-inconsistent | −3.77 | 31.78 | −2.32 | 39.16 |
| **Time 2: After training** | | | | |
| Prejudice-consistent | 1.56 | 33.71 | −21.39 | 37.32 |
| Prejudice-inconsistent | −7.54 | 28.82 | 21.19 | 49.31 |

information had higher levels of automatic racial prejudice ($M = 5.33$, $SD = 41.55$) relative to those that negated prejudice-consistent information ($M = -10.08$, $SD = 37.12$). The effect of negation was nonsignificant, $F(1,121) = 0.19$, $p = .66$, $\eta^2 = .002$. However, of most importance, these main effects were qualified by a significant two-way interaction of Training × Negation, $F(1, 121) = 14.80$, $p < .001$, $\eta^2 = .11$, demonstrating that the prejudice-consistent versus inconsistent training influenced Time 2 automatic attitudes only for the meaningful but not for the simple negation conditions (see Table 1 for means and standard deviations of automatic prejudice scores for each condition).

## *Discussion*

The present findings are informative in at least two ways. First, consistent with our hypothesis and prior research about the ineffectiveness of simple negations, participants in the simple negation conditions showed no change in levels of automatic racial prejudice from Time 1 to 2. In stark contrast, those who engaged in meaningful negation showed significant learning effects, as evidenced by a significant change in automatic racial prejudice following training. Critically, while those who meaningfully negated prejudice-consistent information showed a significant decrease in automatic racial prejudice, those who meaningfully negated prejudice-inconsistent information showed a significant increase. This latter finding is the first evidence regarding the negation of prejudice-inconsistent information and illustrates a potential avenue by which some individuals could reaffirm prejudicial beliefs. More importantly, at a conceptual level, it also suggests saying "That's wrong!" in the context of a study about race is not sufficient to produce a reduction in prejudice. Together, these findings not only demonstrate the power of meaningful negation to influence automatic attitudes, they also point to a potential method to reduce automatic prejudice.

The goals of Study 2 were twofold. Foremost, we wanted to replicate the effects of Study 1 because our meaningful negation conditions and results were novel. In contrast, we observed no changes in automatic prejudice for the simple negation conditions. Because some prior research had indicated that simple negations can backfire rather than have no effect (Gawronski et al., 2008), we wanted to once again compare the impact of meaningful and simple negations.

A second goal of Study 2 was to examine the role of participants' motivation to control prejudiced reactions (MCPR; Dunton & Fazio, 1997). As noted earlier, negation in racial contexts might be more effective when people are especially motivated to attend to racial stimuli. For example, individuals high in MCPR are more likely to be interested in the task and thus be more involved in the training exercise. They might also be more motivated and more attuned

to cues in a domain where race is salient and consequently, may be more sensitive to prejudice-related training. One possible result is that those high in MCPR might be especially amenable to the prejudice reduction training but resistant to the prejudice enhancement training since their goal is to control their prejudiced reactions and this training is consistent with their wishes. On the other hand, if those high in MCPR are more interested in and attentive to race-relevant information and feedback of any kind, they might be more susceptible to the training regardless of the direction (i.e., whether in a prejudice-inconsistent or consistent direction). Finally, a third possibility is that because individuals high in MCPR may already be well practiced in negating their prejudice (e.g., see Maddux et al., 2005), they would be less susceptible to the antiprejudice training than those low in MCPR due to a ceiling effect (see also Livingston, 2008). These possibilities were examined in Study 2.

## Study 2

### *Method*

*Participants and design.* Two-hundred and thirteen undergraduates (114 female; 99 male) at a large Midwestern university in the US enrolled in introductory psychology were recruited to participate in a study on "learning and memory" and received partial course credit in exchange for their involvement. Introductory psychology students are prevented from reenrolling in studies once having received partial course credit for a study, thus no Study 2 participants took part in Study 1. Mean age of participants was 18.86 ($SD$ = 2.17) years. All participants indicated they spoke English as a first language and majority of our sample identified as White/Caucasian (84.5%).[7]

Consistent with Study 1, the design was a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) × 2 (time of measurement: before vs. after) mixed-model factorial, where the first two variables were between-subjects and the latter was within-subjects. MCPR was also assessed and used in the analysis as a continuous variable. Sample size was again determined based on previous research examining negation training and prejudice reduction (see Gawronski et al., 2008).

*Training task.* The procedure and instructions for the training task were identical to Study 1. One change was participants also completed the Motivation to Control for Prejudiced Reactions Scale (MCPR; Dunton & Fazio, 1997) after completing the evaluative priming task for the second time. We chose to administer the MCPR scale following the experimental manipulation to avoid the MCPR questions from potentially contaminating the training task and making the study of race overly salient. Finally, participants completed demographic items and were thoroughly debriefed, thanked, and dismissed.

*Automatic evaluation.* Consistent with Study 1, a subliminal evaluative priming task was used to assess automatic evaluations of Blacks and Whites (see Fazio et al., 1995; Gawronski et al., 2008). The general procedure, number of trials, and scoring were identical to Study 1. Again, participants completed the priming task immediately before and after training.

### *Results*

As expected, relatively few errors were made in judging adjectives as positive or negative (4.0% at Time 1; 5.6% at Time 2), and these trials were excluded from further analyses. Prior to analyses, outliers were excluded by discarding response times lower than 300 ms (0.8% at Time 1; 1.3% at Time 2) or higher than 1,000 ms (4.6% at Time 1; 5.3% at Time 2).

As in Study 1, in our first analysis participants' automatic prejudice scores were submitted to a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) × 2 (time of measurement) mixed-model ANOVA, where the first two variables were between-subjects and the latter was within-subjects. Results revealed a significant three-way interaction of Training × Negation × Time, $F(1, 209) = 7.26$, $p = .008$, $\eta^2 = .03$.

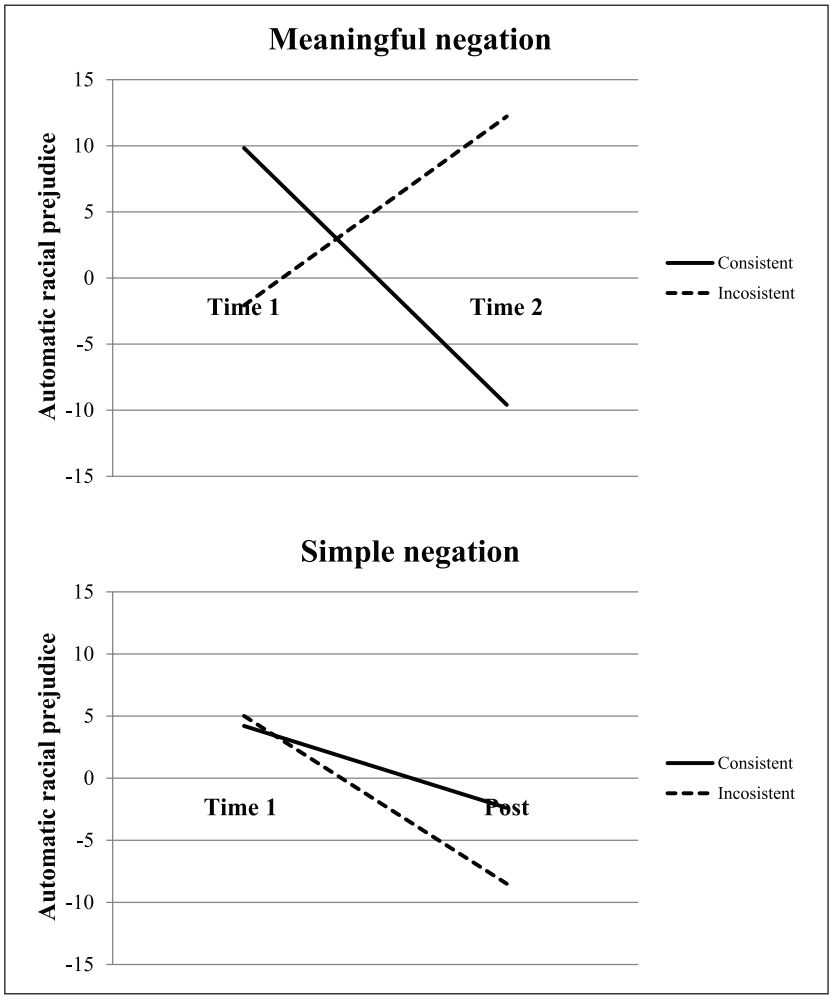


**Figure 4.** Mean scores of automatic racial prejudice as a function of training direction (prejudice-consistent vs. inconsistent) and negation type (meaningful [top panel] vs. simple [bottom panel]; Study 2).

Decomposing this three-way interaction as a function of negation revealed a significant two-way interaction of Training × Time of Measurement, $F(1, 124) = 14.73$, $p < .001$, $\eta^2 = .11$, for the meaningful negation conditions (see top panel of Figure 4), that was absent for the simple negation conditions, $F(1, 85) = 0.285$, $p = .59$, $\eta^2 = .003$ (see bottom panel of Figure 4). This three-way interaction is very similar to that observed in Study 1 (see Figure 3).

We also assessed differences in automatic racial prejudice from Time 1 to Time 2 by experimental condition. A paired-samples $t$ test looking within condition from Time 1 to Time 2 demonstrated a significant decrease in automatic racial prejudice for the meaningful prejudice-consistent training condition, $t(67) = 3.08$, $p = .003$, 95% CI [6.85, 32.01], and a significant increase in automatic racial prejudice for the meaningful prejudice-inconsistent condition, $t(57) = -2.38$, $p = .02$, 95% CI [−26.29, −2.30]. In contrast, paired-samples $t$ tests for the simple prejudice-consistent, $t(46) = 0.78$, $p = .43$, 95% CI [−10.31, 23.51], and the simple prejudice-inconsistent

**Table 2.** Mean scores of automatic racial prejudice as a function of type of negation (meaningful vs. simple) and training (prejudice-consistent vs. inconsistent) before (top) and after (bottom) training (Study 2).

| Training | Negation | | | |
|---|---|---|---|---|
| | Simple | | Meaningful | |
| | *M* | *SD* | *M* | *SD* |
| **Time 1: Before training** | | | | |
| Prejudice-consistent | 4.20 | 42.68 | 9.83 | 34.37 |
| Prejudice-inconsistent | 5.01 | 38.21 | −2.08 | 30.76 |
| **Time 2: After training** | | | | |
| Prejudice-consistent | −2.39 | 42.28 | −9.60 | 46.34 |
| Prejudice-inconsistent | −8.49 | 41.17 | 12.22 | 34.23 |

condition, $t(39) = 1.35$, $p = .20$, 95% CI [−6.69, 33.71], revealed no significant differences in automatic racial prejudice from Time 1 to 2.

Finally, we again examined levels of automatic racial prejudice at Time 1 and Time 2 separately to examine differences as a function of training direction and negation type. We submitted participants' automatic racial prejudice scores at Time 1 to a 2 (training direction: prejudice-consistent vs. inconsistent) × 2 (negation type: simple vs. meaningful) ANOVA. Consistent with Study 1, the main effect of training, $F(1, 209) = 1.23$, $p = .27$, $\eta^2 = .006$, negation, $F(1, 213) = .02$, $p = .88$, $\eta^2 < .001$, and the interaction of the two, $F(1, 209) = 1.57$, $p = .21$, $\eta^2 = .007$, were all nonsignificant, demonstrating participants did not significantly differ at Time 1. In contrast, examining participants' level of automatic racial prejudice at Time 2 revealed only a significant interaction of training and negation, $F(1, 209) = 5.84$, $p = .01$, $\eta^2 = .027$. The main effects of training, $F(1, 209) = 1.83$, $p = .17$, $\eta^2 = .009$ and negation, $F(1, 209) = 1.35$, $p = .24$, $\eta^2 = .006$, were nonsignificant (see Table 2 for means and standard deviations of automatic prejudice scores for each condition). The interaction of training and negation at Time 2 indicated the direction of training influenced automatic racial attitudes when the training involved meaningful negation but had no impact on attitudes in the simple negation conditions.

*Motivation to control prejudiced reactions.* Participants' motivation to control prejudiced reactions (MCPR) scores were calculated according to Dunton and Fazio (1997) to examine if individual differences on this measure moderated automatic racial prejudice following training. The mean score for our sample was $M = 4.25$, $SD = 0.57$ and despite following the experimental manipulation, mean scores did not differ significantly by training direction, $F(1, 209) = 2.03$, $p = .16$, $\eta^2 = .01$, 95% CI [−.049, 0.35], negation type, $F(1, 209) = 0.12$, $p = .91$, $\eta^2 < .001$, 95% CI [−0.17, 0.26], or the interaction of the two, $F(1, 209) = 0.22$, $p = .63$, $\eta^2 = .001$, 95% CI [−0.39, 0.24]. In order to assess if MCPR differentially affected participants' changes in automatic racial prejudice from Time 1 to Time 2, negation and training were dummy coded (i.e., Meaningful = 1, Simple = 0; Consistent = 1, Inconsistent = 0), whereas time was effects-coded (i.e., Time 1 was coded as 1 and Time 2 was coded as −1) and participants' MCPR scores were centered and treated as a continuous variable. Additionally, given we again only found significant effects of training direction among the meaningful negation conditions, we opted to examine the impact of MCPR on simple versus meaningful conditions separately.

Looking first at the simple negation conditions, we submitted participants' raw automatic racial prejudice scores to a Time × Training × MCPR regression analysis using hierarchical linear modeling; no significant effects emerged. Mirroring our repeated measures ANOVA, no significant effects of time, training, or the interaction of the two were found (all $p$s < .19). Likewise,

the effect of MCPR, and its interaction with time, training, and the interaction of all three, were all nonsignificant (all *p*s < .13; see Supplemental File C for full regression statistics).

Turning to the meaningful negation conditions, participants' raw automatic racial prejudice scores pre- and posttraining were submitted to a Time × Training × MCPR regression analysis, using hierarchical linear modeling. In addition to a two-way Time × Training interaction, $B = 32.25$, 95% CI [14.97, 49.54], $SE = 8.82$, $t(119) = 3.66$, $p < .001$, replicating the repeated measures ANOVA, a significant three-way interaction of Time × Training × MCPR, $B = 31.62$, 95% CI [0.63, 62.61], $SE = 15.81$, $t = 2.00$, $p = .04$, also emerged. To understand the nature of this three-way interaction, automatic racial prejudice at Time 1 and Time 2 by condition were investigated for people who were relatively high and low in MCPR. To keep MCPR as a continuous variable, these analyses were performed as a regression, with MCPR examined at one standard deviation above and below the mean. Decomposing this three-way interaction into two, two-way interactions of Time × Training for each level of MCPR, we find a significant effect of Time × Training among those people high in their MCPR, $B = 51.13$, 95% CI [26.78, 75.48], $SE = 12.42$, $t(119) = 4.12$, $p < .001$ (see top panel of Figure 5), and a nonsignificant interaction among those low in MCPR, $B = 13.37$, 95% CI [−12.88, 39.63], $SE = 13.39$, $t(119) = 1.00$, $p = .32$ (see bottom panel of Figure 5).

To further decompose the Time × Training interaction among those high in MCPR, we examined the effect of time for the prejudice-consistent and prejudice-inconsistent conditions separately. Looking first at the prejudice-consistent condition, we found a significant effect of time, $B = 27.15$, 95% CI [8.80, 45.50], $SE = 9.36$, $t(65) = 2.90$, $p = .005$. Similarly, looking at the prejudice-inconsistent condition, we also found a significant effect of time, $B = -23.98$, 95% CI [−39.61, −8.35], $SE = 7.97$, $t(55) = -3.01$, $p = .004$, suggesting the training significantly altered levels of automatic prejudice in both meaningful conditions among those high in MCPR.

### *Discussion*

Consistent with the results of Study 1, Study 2 again illustrated meaningful negation can be an effective tool to alter automatic attitudes. As predicted, those participants who engaged in meaningful negation of prejudice-consistent information demonstrated a decrease in levels of automatic racial prejudice, whereas those who engaged in meaningful negation of prejudice-inconsistent information showed an increase in levels of automatic racial prejudice. Also, no significant differences were found in automatic prejudice levels for those who completed simple negation training. We also found the effectiveness of meaningful negation training was moderated by participants' motivation to control for prejudiced reactions as measured by the MCPR scale (Dunton & Fazio, 1997). Participants high in their motivation to control prejudiced reactions appeared to be driving the observed changes in automatic racial prejudice from Time 1 to 2, for both meaningful conditions. Because those high in their motivation to control for prejudice were more impacted by both the negation of prejudice-consistent and prejudice-inconsistent information, it suggests MCPR is playing a moderating role because participants are more attentive to and interested in racial stimuli and feedback and thus show greater learning regardless of the direction of that learning.

## General Discussion

The primary goal of these studies was to examine the conclusion of prior research that negation is an ineffective means of reducing automatic prejudice. The present research demonstrated negation can be an effective tool to reduce automatic prejudice if the negation is more meaningful and powerful than that used in prior research. In support of this idea, in two studies we demonstrated prejudice reduction training using meaningful negation could effectively alter

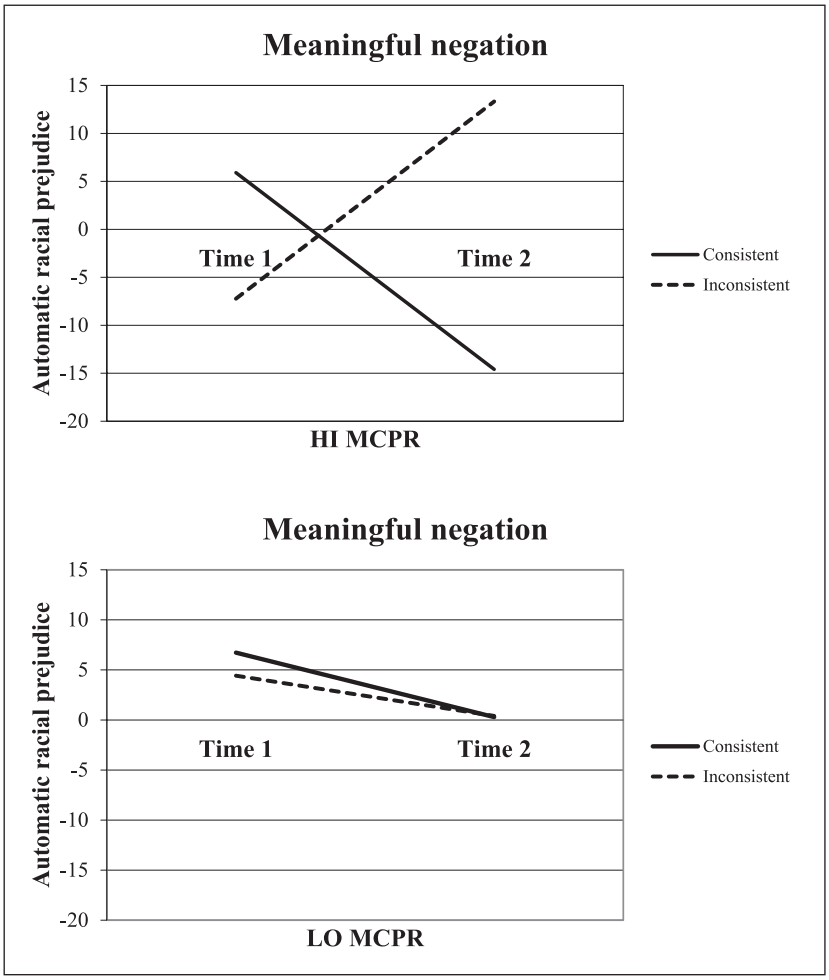


**Figure 5.** Mean scores of automatic racial prejudice as a function of training direction (prejudice-consistent vs. inconsistent), negation type (simple vs. meaningful), and MCPR. MPCR scores were mean-centered and examined at both one standard deviation above (high; top panel) and below (low; bottom panel) the mean (Study 2).

automatic racial attitudes whereas simple negation was completely ineffective.

In Study 1, we extended the work of Gawronski et al. (2008) and compared the impact of negating prejudice-consistent versus prejudice-inconsistent information using simple and meaningful negations. We found only those assigned to the meaningful negation training conditions showed a significant change in automatic racial attitudes. Specifically, we found participants who meaningfully negated prejudice-consistent information demonstrated a significant decrease in their automatic racial prejudice from Time 1 to Time 2, whereas those who meaningfully negated prejudice-inconsistent information demonstrated a significant increase in their automatic racial prejudice from Time 1 to 2. This latter finding demonstrates the meaningful negation of prejudice-inconsistent information is a potential means by which individuals might reinforce their prejudicial beliefs. Of conceptual importance, this novel result also suggests simply saying "That's

wrong!" in a study about racial prejudice is not enough to reduce automatic prejudice. Rather, what is meaningfully negated is of critical importance. Finally, in Study 1, we found no significant changes in automatic prejudice for those who completed simple negation training.

In Study 2, we replicated the results of Study 1 and also examined the role of participants' motivation to control prejudiced reactions (MCPR; Dunton & Fazio, 1997). Specifically, we addressed the possibility negation training with respect to racial prejudice would be more impactful for individuals who are high in MCPR. This study had two key results. First, we replicated the finding of Study 1; only meaningful negation training led to a significant change in automatic racial prejudice levels and observed no significant changes for those who completed simple negation training. Second, this study highlighted the role of MCPR in that we found the influence of meaningful negation training on participants' change in automatic racial prejudice from Time 1 to Time 2 was driven primarily by those participants who were high in MCPR. Those high in MCPR became less prejudiced after they practiced meaningfully negating prejudice-consistent information, but they also became more prejudiced after meaningfully negating prejudice-inconsistent information, relative to those low in MCPR.

There are at least two possibilities for why those high in MCPR would show a greater impact of training regardless of its direction (pro or antiprejudice). The first possibility is those high in MCPR performed better during the learning trials than those low in MCPR and their reduced prejudice was a result of greater task performance. The second possibility is everyone learned to perform the task equally well, but the task was more meaningful for those high in MCPR so the training carried over to influence attitudes for them but not for those low in MCPR. That is, it could be everyone was able to perform the learning task equally well (i.e., push this button when the stimulus appears), but if the task is a sterile exercise for those low in MCPR but had substantive meaning for those high in MCPR, the latter group would generalize the task to more global attitudes. In essence, just as individuals performed the task equally well whether the response was "That's wrong!" or "No!" but the former is presumably more meaningful than the latter, so too might those high and low in MCPR perform the task equally well, but "That's wrong!" would be more meaningful for the former than the latter group.

To examine these possibilities, we analyzed performance in the training task in Study 2 as a function of MCPR. Critically, although the changes in automatic prejudice were clearly driven by those high in MCPR, we did not find any differences in participants' performance on the training task. Specifically, when we examined participants' errors and reaction times on the training task as a function of block (i.e., 1–5), their assigned training task and motivation to control for prejudice reactions, we found no main or interaction effects involving MCPR (i.e., $p$s > .30). This suggests while both high and low MCPR participants performed equally well on the training task, the task was simply more meaningful for those high in MCPR than those low in this individual difference, just as "That's wrong!" is more meaningful than a simple "No!" Taken together, these studies provide the first evidence that negation training can serve as a useful tool to alter individuals' automatic racial prejudice—if the negations are meaningful and one is motivated to avoid being prejudiced.

The present work clearly suggests negations using "That's wrong" versus a simple "No" carry some added value. Indeed, the key difference between the present research and previous work on negating prejudice was the replacement of "No" with "That's wrong." Yet, this seemingly slight modification yielded a very different result. Consequently, these findings raise the question of why "That's wrong," is impacting automatic attitudes whereas a simple "No" is not. We suggest two potential reasons why "That's wrong" is more impactful.

As mentioned earlier, one potential reason why "That's wrong" was more impactful is perhaps "That's wrong" is less ambiguous than

"No." Intuitively, negating prejudice-consistent information should lead an individual to be less prejudiced; but perhaps in the context of the simple negation training, the meaning of "No" was unclear. Previous work has suggested processing negations is effortful (Boucher & Rydell, 2012; Gilbert et al., 1993), and successful negation can take extensive time and practice (e.g., Deutsch, Gawronski, & Strack, 2006). Consequently, in the absence of these factors, negation can be ambiguous. Perhaps by attaching a clear meaning to the negation, the effort needed to successfully negate associations was reduced allowing the more meaningful negation to effectively modify automatic prejudice. Related to this possibility is negation training using "That's wrong" may carry some moral significance. In other words, one potential means through which the negation was made clear was by adding a moral sentiment to the meaning of the negation. Previous research has found attitudes held with strong moral conviction are more likely to modify behavior reliably, compared to attitudes held without moral conviction (Skitka, Bauman, & Sargis, 2005). Assuming most individuals disapprove of morally forbidden behaviors, perhaps using "That's wrong!" made this norm more salient and thus made the negation training more impactful.

A second potential reason is even if "That's wrong" and "No" carry the same initial meaning, the former might be more likely to be stored for later retrieval or be more accessible. According to the meta-cognitive model (MCM; Petty, Briñol, & DeMarree, 2007) of attitude structure, social categories (e.g., Black) might not only be associated with evaluations (good/bad), but also these evaluations can be further linked to validity tags. If tagging an evaluation along the right/wrong dimension is more memorable or more quickly retrievable than tagging it along a yes/no dimension, this could account for why "That's wrong" is more effective in undermining automatic prejudice. Indeed, if a negation is to be effective in undermining automatic prejudice, it should be retrieved quickly and automatically. Thus, in addition to the present findings suggesting meaningful negation training is an effective tool to reduce automatic racial prejudice, the present work is consistent with the idea that negations, at least following extensive practice, can be stored along with evaluative associations in memory for later retrieval (Maddux et al., 2005; Mayo, Schul, & Burnstein, 2004).[8]

In the present research, participants were trained to negate prejudice-consistent information meaningfully and this produced a reduction in automatic racial prejudice. One issue worth noting is the negation training for the current studies and others (e.g., Gawronski et al., 2008; Kawakami et al., 2000) required participants to negate both Black faces paired with negative traits *and* White faces paired with positive traits. This, of course, is done because racial prejudice can potentially stem from a "Black–bad" and a "White–good" mentality, and targeting both associations has been shown to effectively reduce automatic bias as noted in a recent meta-analysis by Lai et al. (2014). However, different people can show relative prejudice against Blacks versus Whites mostly because of anti-Black sentiment or pro-White sentiment. One direction for future research is to examine if the negation of both face–trait combinations is necessary in order for a significant decrease in automatic racial prejudice with respect to Blacks versus Whites to be observed or if negating one or the other is sufficient. Although the present research intentionally adopted the methods utilized in previous research to best address if meaningful negation could effectively reduce automatic prejudice whereas simple negation would fail, future research is needed to examine if negating the association of Black–negative or White–positive alone is sufficient to influence automatic prejudice, and if so does this differ as a function of participants' initial racial prejudice.

Past research by Livingston and Drwecki (2007) found nonracially biased individuals, as measured by the racial Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and Attitudes Towards Blacks (ATB; Brigham, 1993) Scale, were less likely to ascribe negativity to novel attitude objects, though racially biased and nonracially biased individuals were equally

likely to ascribe positivity to novel attitude objects. Consequently, if nonracially biased individuals are less likely to assign negativity to novel attitude objects, perhaps meaningful negation training only negating *Black* paired with negative traits would be enough to reduce automatic racial prejudice, but only for those participants already low in racial bias. In any case, now that it is clear negation training, when meaningful, can work to reduce automatic prejudice, future research should address these issues.

A final important direction for future research is to determine how meaningful negation training influences downstream social judgments and behavior. A host of social psychological research has demonstrated the impact of prejudice and stereotypes on social judgments (Macrae, Milne, & Bodenhausen, 1994) and behavior (Jellison, McConnell, & Gabriel, 2004; Wheeler & Petty, 2001), but far fewer have delineated how changes in automatic racial prejudice impact other judgments and behavior (Blair, 2002; Kawakami et al., 2000; Monteith, 1993). As past research has documented, racial prejudice is far more predictive of discrimination than is stereotyping (Dovidio & Gaertner, 1996; Esses, Haddock, & Zanna, 1993; Stangor, Sullivan, & Ford, 1991), it is critical for future research to investigate the long-term effects of meaningful negation training on other judgments (e.g., jury verdicts) and behavior (e.g., hiring decisions).

## Author Note

This research article is based on a thesis submitted by India R. Johnson to The Ohio State University Graduate School in partial fulfillment of the requirement for the master's degree of Social Psychology.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. In the Gawronski et al. (2008) research, participants were only presented with negative traits stereotypically associated with Blacks and positive traits stereotypically associated with Whites. Consequently, we use the terms "prejudice-consistent" and "prejudice-inconsistent" in lieu of stereotype-consistent and stereotype-inconsistent, respectively.
2. Other research suggests that negations that are present at the time of initial learning of something new rather than after learning can be effective (e.g., Boucher & Rydell, 2012; Peters & Gawronski, 2011) as can negations that are very highly practiced over an extended period of time (e.g., Deutsch et al., 2006; Maddux et al., 2005).
3. The remainder of participants identified as Black/African American (1.6%), Native American (0.8%), Middle Eastern (0.8%), Asian (7.2%), Hispanic/Latino (4.8%) or other (1.6%). The results of analyses conducted excluding African Americans did not differ from those which included all participants. Consequently, we report results for each study including the entire sample.
4. Because prior research on prejudice negation included White as well as Black faces in the training procedure (e.g., see Gawronski et al., 2008; Kawakami et al., 2000), we also included them so we could directly compare our results to previous research.
5. In addition to explaining our purpose and expected findings, our debriefing provided participants with an opportunity to ask any questions or express any concerns they had regarding the experimental session. All questions and concerns were addressed immediately following the experimental session and participants were provided with the primary investigator's contact information if they would like to follow up further. No participants chose to do so.
6. The pattern for the number of errors and participants' reaction times for the training task were consistent across both studies, and thus these results will not be discussed for Study 2.
7. The remainder of participants identified as Black/African American (1.8%), Middle Eastern (2.3%), Asian (6.6%), Hispanic/Latino (3.6%) or other (0.9%).
8. It is also possible during negation of prejudice training, instead of encoding an association of Blacks with a negative term and a negative term with an invalidity tag, participants engaged in an on-line translation process whereby Black was eventually encoded with the opposite of the negative term presented (i.e., Black–lazy–that's wrong

was translated into Black–industrious). Some research suggests this is most likely when there is an easily reversible negation (e.g., hot–cold; Mayo et al., 2004). In any case, regardless of the precise mechanism, the current work still provides the first evidence negation training, when meaningful, can effectively undermine automatic prejudice, though the precise manner in which this is accomplished is a matter for future research.

## References


Allen, T. J., Sherman, J. W., & Klauer, K. C. (2010). Social context and the self-regulation of implicit bias. *Group Processes & Intergroup Relations*, *13*, 137–149. doi:10.1177/1368430209353635

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

Blair, I. V., & Banaji, M. H. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*, 1142–1163. doi:10.1037/0022-3514.70.6.1142

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin*, *38*, 1329–1342. doi:10.1177/0146167212450464

Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied and Social Psychology*, *23*, 1933–1967. doi:10.1111/j.1559-1816.1993.tb01074.x

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology*, *91*, 385–405. doi:10.1037/0022-3514.91.3.385

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice-habit breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267–1278. doi:10.1016/j.jesp.2012.06.003

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. doi:10.1037/0022-3514.82.5.835

Dovido, J. F., & Gaertner, S. L. (1996). Affirmative action, unintentional racial biases, and intergroup relations. *Journal of Social Issues*, *52*, 51–75. doi:10.1111/j.1540-4560.1996.tb01848.x

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*, 316–326. doi:10.1177/0146167297233009

Esses, V. M., Haddock, G., & Zanna, M. P. (1993). Values, stereotypes, and emotions of determinants of intergroup attitudes. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 137–366). San Diego, CA: Academic Press.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027. doi:10.1037/0022-3514.69.6.1013

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*, 370–377. doi:10.1016/j.jesp.2006.12.004

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, *57*, 940–949. doi:10.1037/0022-3514.57.6.940

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*, 221–233. doi:10.1037/0022-3514.65.2.221

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464. doi:10.1037/0022-3514.74.6.1464

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*, 1–20. doi:10.1037/0022-3514.90.1.1

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814. doi:10.1037/0033-295X.108.4.814

Jellison, W. A., McConnell, A. R., & Gabriel, S. (2004). Implicit and explicit measures of sexual orientation attitudes: Ingroup preferences and related behaviors and beliefs among gay and straight men. *Personality and Social Psychology Bulletin*, *30*, 629–642. doi:10.1177/0146167203262076

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*, 871–888. doi:10.1037/0022-3514.78.5.871

Lai, C. K., Marini, M., Lehr, S., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing racial implicit preferences: A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785. doi:10.1037/a0036260

Livingston, R. (2008). Why are some individuals not racially biased? Susceptibility to affective conditioning predicts nonprejudice towards Blacks. *Psychological Science*, *18*, 816–823. doi:10.1111/j.1467-9280.2007.01985.x

Livingston, R. W., & Drwecki, B. B. (2007). Why are some individuals not racially biased? Susceptibility to affective conditioning predicts nonprejudice toward Blacks. *Psychological Science*, *18*, 816–823. doi:10.1111/j.1467-9280.2007.01985.x

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality & Social Psychology*, *66*, 37–47. doi:10.1037/0022-3514.66.1.37

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, *41*, 19–35. doi:10.1016/j.jesp.2004.05.002

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*, 433–449. doi:10.1016/j.jesp.2003.07.008

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*, 469–485. doi:10.1037/0022-3514.65.3.469

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 226–254. doi:10.1037/0096-3445.106.3.226

Olson, M. A., & Fazio, R. H. (2004). Trait inferences as a function of automatically activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology*, *26*, 1–11. doi:10.1207/s15324834basp2601_1

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

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*, 557–569. doi:10.1177/0146167211400423

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Implications for attitude measurement, change and strength. *Social Cognition*, *25*, 657–686. doi:10.1521/soco.2007.25.5.657

Petty, R. E., Briñol, P., Tormala, Z. L., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, *90*, 21–41. doi:10.1037/0022-3514.90.1.21

Plant, A. E., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832. doi:10.1037/0022-3514.75.3.811

Skitka, L., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, *88*, 895–917. doi:10.1037/0022-3514.88.6.895

Stangor, C., Sullivan, L. A., & Ford, T. E. (1991). Affective and cognitive determinants of prejudice. *Social Cognition*, *9*, 359–380. doi:10.1521/soco.1991.9.4.359

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

Wegener, D. T., & Petty, R. E. (1997).The flexible correction model: The role of naive theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141–208). San Diego, CA: Academic Press.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin*, *127*, 797–826. doi:10.1037/0033-2909.127.6.797

## Appendix A

### *Training Instructions*

Prejudice-consistent conditions (*changes for those in the meaningful conditions are in parentheses*):

> Your task is to respond "NO! [THAT'S WRONG!]" each time you see a combination that is CONSISTENT with the cultural stereotype of Blacks and Whites. Specifically, you are asked to respond "NO! [THAT'S WRONG!]" with the space bar each time you see a picture of a BLACK man and a NEGATIVE trait or a picture of a WHITE man and a POSITIVE trait. Please attend particularly to combinations that are CONSISTENT with the cultural stereotype of Blacks and Whites! For combinations that are inconsistent with the cultural stereotype of Blacks and Whites, you do not have to do anything. Again, please respond "NO! [THAT'S WRONG!]" with the space bar each time you see a combination that is CONSISTENT with the cultural stereotype of Blacks and Whites. Please try to respond as quickly and accurately as possible!

Prejudice-inconsistent conditions (*changes for those in the meaningful conditions are in parentheses*):

> Your task is to respond "NO! [THAT'S WRONG!]" each time you see a combination that is INCONSISTENT with the cultural stereotype of Blacks and Whites. Specifically, you are asked to respond "NO! [THAT'S WRONG!]" with the space bar each time you see a picture of a BLACK man and a POSITIVE trait or a picture of a WHITE man and a NEGATIVE trait. Please attend particularly to combinations that are INCONSISTENT with the cultural stereotype of Blacks and Whites! For combinations that are consistent with the cultural stereotype of Blacks and Whites, you do not have to do anything. Again, please respond "NO! [THAT'S WRONG!]" with the space bar each time you see a combination that is INCONSISTENT with the cultural stereotype of Blacks and Whites. Please try to respond as quickly and accurately as possible!

## Appendix B

### *Stimuli*

Trait words used in the training task in Studies 1 and 2.

Trait words used that were consistent with prejudice toward Blacks: *poor*, *dishonest*, *complaining*, *violent*, *shiftless*, *superstitious*, *lazy*, *threatening*, *dumb*, *hostile*.

Trait words used that were inconsistent with prejudice toward Blacks: *intelligent*, *successful*, *ambitious*, *industrious*, *educated*, *responsible*, *wealthy*, *ethical*, *smart*, *friendly*.

## Appendix C

Black faces used on training task:

## Appendix D

White faces used on training task:

## Appendix E

Positive and negative target words used in the evaluative priming task in Studies 1 and 2.
Positive target words: *paradise*, *summer*, *harmony*, *freedom*, *honesty*, *honor*, *health*, *cheer*, *pleasure*, *heaven*, *friend*, *sunrise*, *love*, *relaxation*, *peace*, *vacation*, *happy*, *lucky*, *miracle*, *gift*.
Negative targets words: *evil*, *sickness*, *vomit*, *bomb*, *murder*, *abuse*, *prison*, *death*, *assault*, *cancer*, *rotten*, *accident*, *grief*, *poison*, *stink*, *cockroach*, *virus*, *disaster*, *ugly*, *terror*.