


# Implicit bias reduction that lasts: Putting Situational Attribution Training to the test

Tracie L. Stewart<sup>1</sup>  | Ioana M. Latu<sup>2</sup> | Tim Martin<sup>1</sup> | Seamus P. Walsh<sup>3</sup> | Allyson Schmidt<sup>1</sup> | Kerry Kawakami<sup>4</sup>

<sup>1</sup>Department of Psychological Sciences, Kennesaw State University, Atlanta, Georgia, USA

<sup>2</sup>School of Psychology, Queen's University Belfast, Belfast, UK

<sup>3</sup>Department of Psychology, University of Mississippi, Oxford, Mississippi, USA

<sup>4</sup>Department of Psychology, York University, Toronto, Ontario, Canada

## Correspondence

Tracie Stewart, Department of Psychological Science, Kennesaw State University, 402 Bartow Ave NW, Kennesaw, GA 30144, USA. Email: [stewart@kennesaw.edu](mailto:stewart@kennesaw.edu)

## Funding information

Russell Sage Foundation

## Abstract

Addressing the damaging effects of implicit stereotypes—spontaneous, awareness-independent associations between social groups and particular traits—remains a social imperative. These biases have been linked to negative outcomes in settings ranging from the workplace to medical care facilities. However, many techniques found to reduce implicit biases have been shown to yield short-lived effects. In the present experiment, we assessed the longevity of reduced implicit racial stereotyping resulting from an intensive training technique that focuses on weakening the fundamental attributional processes underlying implicit stereotyping. Specifically, we aimed to strengthen the likelihood of White participants to consider situational attributions for behaviors performed by Black men that might otherwise have been judged to reflect negative African American stereotypes. White participants were randomly assigned to complete either Situational Attribution Training (SAT), a technique comprised of intensive training (480 trials) to “consider the situation” when making judgments about stereotype-consistent behaviors performed by Black men, or a control task. Implicit stereotyping was assessed 24h later via the Person Categorization Task and found to be reduced for SAT, versus control, participants even after this delay. Implications for future antibias research and practice are considered.

## 1 | INTRODUCTION

Social psychological research has found that implicit stereotypes—spontaneous, awareness-independent associations between social groups and certain traits (Hahn & Goedderz, 2020)—can predict an array of tangible critical outcomes, ranging from inequity in employers' decisions about employee raises and promotions to the potentially tragic consequences of police officers' split-second “shoot-don't shoot” judgments (Greenwald et al., 2015; Kahn & Davies, 2011; Latu et al., 2015; Spencer et al., 2016; Streeter, 2019). These spontaneous biases can also impact majority group members' likelihood of engaging in antidiscrimination and antiviolence activism

(Stewart et al., 2013). Thus, it is not surprising that a goal of many intergroup relations researchers and practitioners is to identify effective and long-lasting strategies to combat implicit biases in intergroup judgments and interactions (Kawakami et al., 2017).

Is it possible to reduce the activation of implicit stereotyping when these biases are tightly woven throughout our culture? And is any change achieved by bias reduction strategies sufficiently enduring to lead to meaningful change? There is ample evidence to answer “yes” to the question of whether bias reduction is possible, with some bias reduction strategies proving to be more impactful than others. In one recent meta-analysis, Forscher et al. (2019) found that strategies that seek to reduce biased associations through direct

training (e.g., repeatedly paired photos of African Americans with positive, nonstereotypic traits) or indirect training (e.g., introducing alternative means of processing information) were more effective than strategies that targeted feelings of threat or guilt as inducement for bias reduction. However, in general, this meta-analysis found that bias reduction effects were relatively weak and unrelated to changes in explicit biases or behavior. Perhaps even more concerning, other studies have shown that some antibias strategies can lead to *increased* bias. For example, when participants are trained to suppress existing stereotypes, unintended suppression-related “rebound effects” may occur, such that suppressed stereotypes become stronger over time (Macrae et al., 1994). In addition, reactance may occur if participants feel forced to undergo antibias training programs, which can lead to backlash effects (Devine et al., 2000; Dobbin & Kalev, 2016).

As to whether bias reduction effects are enduring, in many cases this question cannot be assessed given that measures of implicit bias are administered only immediately after the antibias intervention. In Forscher et al. (2019) meta-analysis of studies that did include a delay in bias assessment, a discouraging pattern of effects was obtained suggesting that bias reduction effects are short-lived. Another recent study compared the longevity of reduced automatic racial biases for nine interventions that had previously yielded strong reductions in automatic racial biases and found that none of the interventions continued to yield positive effects 24 h after training (Lai et al., 2016). The authors posit two explanations for the short temporal impact of these effects. First, the interventions might need to be longer and more intensive to produce enduring effects. Second, the present interventions simply might not be tapping into the most effective mechanisms for change.

A noteworthy exception to this pattern of short-lived bias reduction effects is a series of studies by Forscher et al. (2017) which tested implicit bias reduction 2 weeks or 2 months (Devine et al., 2012) after an antibias intervention. This intervention was approached through a framework of implicit biases as “habits” to be broken and comprised a “semi-interactive slide show” in which participants were informed of their personal levels of implicit bias and received education about causes, consequences, and evidence-based strategies to reduce implicit biases. In the 2017 study, participants were also asked to write an essay about benefits of the slideshow for potential future participants. The Implicit Association Test (Greenwald et al., 1998) and a measure of discrepancies between participants’ personal antibias goals and behaviors were administered before and after the intervention, with posttest assessments administered every 2 days for 14 days (Forscher et al., 2017) or every month for 2 months (Devine et al., 2012) after the first session.

Findings of the long-term effectiveness of this intervention were mixed across the two studies, with a significant long-term reduction in implicit bias found for the experimental compared with the control participants in the earlier experiment (Devine et al., 2012), but no differences in implicit bias reduction found for experimental and control participants in the later experiment (Forscher et al., 2017). The authors discuss various reasons for the potential discrepancy,

ranging from effects of increased testing in one experiment (Forscher et al., 2017) to the possibility of a false positive in the other experiment (Devine et al., 2012), ultimately concluding that clarifying the reason for the discrepancy remains a question for future research. Despite the inconsistency across studies, we believe that Devine and Forscher’s contributions to the antibias literature are valuable. We concur with their conceptualization of bias as a habit ingrained by society and see habit-breaking as a promising framework for bias reduction strategies. We have all had experience with trying to break bad habits, with goals ranging from correcting hand positions when playing piano to trying to avoid saying “um” in presentations. And we have all learned that accomplishing these goals often requires a great deal of practice. We posit that breaking the stereotyping “habit” similarly requires intensive practice.

## 1.2 | Situational Attribution Training

In the present research, we investigated the long-term effectiveness of a bias-reduction strategy that incorporates key elements suggested by prior research to be critical for enduring change: the training targets the implicit bias “habit” through intensive practice, resists rebound effects associated with stereotype suppression, and is an indirect training approach targeting the fundamental attributional pillars that underlie automatic stereotyping, rather than the stereotype itself. We predicted that our antibias approach would create lasting change and would even generalize more broadly to stimuli not included in the training.

In Situational Attribution Training (SAT), we target the stereotype-perpetuating tendencies to underestimate situational factors and overestimate dispositional factors in explaining negative behaviors of outgroup members. These tendencies have been labeled the ultimate attribution error (UAE; Byrd & Ray, 2015; Pettigrew, 1979). For example, White participants are inclined to attribute a negative action by an African American actor (e.g., dropping out of college before earning a degree) to genetic dispositional factors (e.g., he was not smart enough) rather than situational factors (e.g., he could no longer afford tuition after it was increased). These types of attributions, in turn, tend to reinforce and perpetuate this negative stereotype. Consider how the UAE might contribute to inequality in the workplace. If a White employee’s tardiness is attributed to unavoidable traffic issues and their raised voice attributed to confidence, whereas a Black employee’s similar tardiness and raised voice are attributed respectively to irresponsibility and aggression, then differences in performance evaluation and workplace advancement are likely to follow for these employees despite their similar records.

In SAT, we train participants to overcome this stereotype-enforcing tendency through intensive practice making situational rather than dispositional attributions for negative stereotype-consistent behaviors of African American men. In prior research, we found SAT had a significant impact in reducing racial bias, compared to control conditions (Stewart et al., 2010). Because SAT targets the

attributional pillars on which stereotyping stands, we posit that SAT has not only immediate effects but also the potential to facilitate a long-lasting reduction of automatic stereotyping. Furthermore, this targeting of a stereotype-perpetuating procedure, versus specific stereotype content, positions this technique to generalize beyond the specific stereotypic attributes targeted during training.

Some researchers have suggested that analyses of antibias interventions (e.g., Lai et al., 2016) may not show lasting effects because of a focus on a limited type of intervention, with most being quite brief and administered online (e.g., Kawakami et al., 2017). Examining an alternative, more intensive approach that targets a novel mechanism, such as encouraging situational attributions, is therefore a recommended avenue for further research.

SAT instructs participants to practice situational attributions for stereotype-consistent behaviors but does not ask participants to suppress alternative judgments. Specifically, participants are presented simultaneously with both stereotypic and situational explanations for specific actions and asked to consistently choose the situational explanation. At no time are participants instructed to suppress or reject stereotypic explanations for the behavior. Because this approach circumvents tendencies to consciously suppress stereotypic attributions, it avoids rebound effects in which suppressed stereotypes become stronger over time through non-conscious monitoring (Macrae et al., 1994).

Besides reducing the likelihood of stereotype rebound effects, SAT is also less likely to induce reactance. Because participants are not directly asked to inhibit stereotypes in this task, but rather to choose situational attributions, the SAT is less likely to lead to resistance or backlash, a significant challenge for current diversity training programs (Dobbin & Kalev, 2016, 2019). Therefore, we posited that SAT can be a uniquely effective, long-lasting, generalizable tool in reducing implicit stereotyping, either when used as a stand-alone intervention (Stewart et al., 2010) or as an option in more comprehensive interventions that offer participants a range of approaches from which to choose (e.g., Devine et al., 2012).

When testing occurred immediately following training, two experiments indicated that participation in SAT was associated with reduced implicit negative racial stereotyping for White participants, relative to participants randomly assigned to a control group (Stewart et al., 2010). This reduction in implicit stereotyping generalized beyond the specific negative African American stereotypic traits targeted in SAT to new negative African Americans stereotypic traits not targeted in the training task. Moreover, it was shown that these effects stemmed, at least in part, from an increase in the automatic activation of situational explanations for stereotype-consistent behaviors by racial outgroups. However, despite this initial experimental support for the effectiveness of SAT, the longevity of SAT's positive effects is not known. In previous experiments using SAT the dependent measures were administered immediately after the training in a single experimental session. In the present experiment, we sought to address the critical question of the persistence of SAT effects.

### 1.3 | The present experiment

We predicted that the reduction of implicit racial stereotyping through SAT would continue up to 24 h after the initial training session. We expected our effects to be longer lasting than many other interventions targeting implicit racial biases (Lai et al., 2016) because SAT is more intensive (480 trials), circumvents attempts at stereotype suppression, and is directed toward the fundamental attributional pillars that underlie automatic stereotyping rather than the surface stereotypic associations themselves. We further predicted that effects of this intensive training would generalize beyond the specific stimuli used in the training.

## 2 | METHODS

### 2.1 | Overview

We assessed the longevity of SAT's effectiveness in reducing automatic racial stereotyping of Black men across a delay of at least 24 h. White participants were randomly assigned to either the SAT condition or one of two control conditions: a "grammar training" task (Grammar Control) in which participants were presented with the same photographs and stereotype-consistent behaviors displayed to the SAT participants but were asked to make grammar, rather than attributional, judgments about the behavior sentences (e.g., How many verbs were in this sentence?) or a No Training-Control condition, in which participants only responded to the dependent measures. Automatic racial stereotyping was assessed using a modified Person Categorization Task (PCT; Banaji & Hardin, 1996). We predicted that participants who completed the SAT would exhibit less automatic racial stereotyping compared to control participants 1 day after training.

### 2.2 | Participants

A total of 130 White undergraduate students from two state universities in the southeastern United States participated in this IRB-approved study as one means to earn course credit. Participants were randomly assigned to either the SAT condition ( $N = 68$ ) or a control condition ( $N = 62$ ). The present sample size exceeds that of previous experiments documenting the effectiveness of SAT in reducing negative African American stereotypes ( $Ns = 32$  and  $40$  in Stewart et al., 2010; Experiments 1 and 2). Our sample size was chosen to be "approximately 100–150" with a goal for the sample to be larger than that of these prior experiments. An additional 27 participants (52% in SAT condition) dropped out of the study before completing the dependent measures, a relatively low level of attrition for a two-part study.

## 2.3 | Materials and procedure

### 2.3.1 | Conditions

All study sessions were conducted face-to-face in campus research labs. Participants in the SAT condition were informed that the study examined ways in which individuals explain the behavior of other people. They were told that they had been randomly assigned to a condition in which they would be judging behaviors performed by African American men and that their goal was to choose the “situational,” versus “dispositional,” explanation for each displayed behavior. Participants first completed six computerized practice trials in which they were given feedback about the accuracy of their responses in choosing situational explanations.

Participants then began the primary computerized training trials, comprised of 6 blocks of 80 trials, with 6 additional practice trials with feedback administered between each block. On each of the trials, a photograph of a Black man and the label “African American” were displayed on a computer screen. Below the photograph and label, a sentence describing a behavior was presented (e.g., “Failed to get his work done for the day”). Forty behaviors were presented twice per block, randomly assigned to a different photograph of a Black man each time. Each behavior implied a trait consistent with negative stereotypes of African American men. Specifically, four behaviors were constructed, and confirmed via pretests, to be indicative of each of the following 10 negative stereotypic traits of African American men: loud, criminal, unintelligent, unreliable, irresponsible, violent, dishonest, dangerous, lazy, and promiscuous.

After 3000 ms, the words “I Choose:” appeared below the behavior description and two potential explanations for the behavior were presented underneath on the left and right sides of the screen. One of the explanations offered a dispositional and stereotype-consistent attribution for the behavior (e.g., He is an unreliable worker), whereas the other explanation offered a situational explanation of the behavior (e.g., His office was being painted, so he could not access his work materials). The location of the attributions on the bottom of the screen was counterbalanced across trials, such that half of the time the situational attribution to be chosen appeared on the left side and the other half on the right side. Participants were instructed to type a key labeled “L” if the situational explanation was on the left and a key labeled “R” if the situational explanation was on the right.

The remaining participants were randomly assigned to one of two control groups. Based on prior research (Stewart et al., 2010), both control groups were expected to demonstrate implicit racial stereotyping. The inclusion of two control groups addressed competing conceptualizations of the more appropriate comparison with SAT: A control condition comprised of a training paradigm highly similar to SAT but with no attributional judgments made or a true no-training control condition. We, therefore, established two types of baseline for comparison with the experimental group, but expected to find evidence of implicit stereotyping in both control groups.

In the Grammar Control condition, participants saw the same photograph, label, “I Choose” text, and behavior descriptions as participants in the SAT condition. However, rather than choose an attribution for each behavior, they were instructed to make decisions about whether the behavior descriptions contained “2 or under 2 nouns” (or “verbs” on some trials) or “over 2 nouns (verbs).” The locations of these two response options were counterbalanced across trials. Participants in the No Training Control condition completed only the dependent measures.

Initial analyses confirmed that, as expected based on prior research, there was no difference in the degree of automatic negative racial stereotyping observed for participants who completed Grammar Control Training and participants who were in the No Training Control condition,  $F(1, 60) < 1$ ,  $\eta^2 = 0.002$ . Therefore, in accordance with prior research, data from the control conditions were combined in the primary analyses.

## 2.4 | Person Categorization Task

All participants completed a measure of implicit stereotyping, the PCT (Banaji & Hardin, 1996). We used Cronbach's  $\alpha$  to compare even and odd trials for each trait prime type used in this task to assess the measure's split-half reliability. The split-half reliability for these conditional means ranged between  $\alpha = .597$  to  $\alpha = .888$ . No Training Control participants completed only the PCT. Participants in the SAT and Grammar Training Control conditions completed the PCT the following day after the training session (delay range = 24–29 h). To encourage participants to return, they were offered bonus credit as an incentive. The PCT portion was presented as a separate experiment and participants were told that this study investigated their ability to categorize faces quickly by race and that they had been randomly assigned to a condition in which a distractor word would be presented briefly, before presentation of the faces. Among these distractor words were our experimental trait primes.

Specifically, on each trial, one of 56 positive and negative trait primes were presented for 250 ms before the display of a photograph of a Black or White man. Participants were instructed to ignore the traits words and to categorize the targets according to race by pressing a key labeled “B” to indicate a photo of a Black person or a key labeled “W” to indicate a photo of a White person. All trait primes had been pretested extensively and used in prior research (Stewart et al., 2010). The target trait primes that comprised our dependent measure were 16 negative stereotypic traits of African American men: eight traits that had been implied in behaviors during SAT (e.g., “missed an important deadline at work” implying “irresponsible”) and eight “new” traits that had been neither presented nor implied during SAT. The remaining nontarget trait primes included eight positive African American-stereotypic traits; 16 negative nonstereotypic traits; and 16 positive nonstereotypic traits. No effects of SAT have been found for these nontarget traits in prior experiments and none were expected for the present experiment.

Two 56-trial blocks were completed. In each block, half of the traits from each trait prime category preceded a Black photo, and the other half preceded a White photo. Targets in the photos were all dressed in similar, casual clothing. None wore glasses or had other distinctive characteristics. The trait–photograph pairings were counterbalanced such that the traits that appeared with a photo of a Black person in one block were shown with a photo of a White person in the other block. Participants were instructed to classify the race of the person as quickly and accurately as possible, and response times were recorded. After completion of the PCT, participants were debriefed and dismissed.

Our primary dependent measure of implicit racial stereotyping was the relative speed in categorizing photographs of Black versus White targets following trait primes consistent with negative stereotypes of African American men, indicating relative association of these traits with Black and White men.

### 3 | RESULTS

#### 3.1 | Preliminary analyses

All response time data were log-transformed to limit effects of outliers. Means are reported in untransformed milliseconds in the present paper. The primary dependent measure was the relative response time for participants to categorize by race photographs of Black versus White men, following trait primes consistent with negative stereotypes of African American men, during the PCT. Stronger negative automatic stereotype activation is indicated by shorter response times categorizing photographs of Black, compared with White, men after being primed with negative African American male-stereotypic trait primes. We combined into a single dependent measure the response times for negative *old* (targeted during training) and negative *new* African American-stereotypic trait primes, given that there was no significant old/new trait main effect, indicating no overall difference between response times to old ( $M = 549.98ms$ ,  $SD = 9.28$ ) and new ( $M = 554.89ms$ ,  $SD = 9.64$ ) target traits, and no old/new trait  $\times$  condition interaction, both  $F(1, 127) < 1.0$ . These findings demonstrate that any observed training effects were not restricted to negative stereotypic traits specifically targeted during training. As expected based on prior research, separate analyses conducted for each type of nontarget trait (i.e., stereotypic positive and nonstereotypic positive and negative traits) yielded no significant effects of training condition.

#### 3.2 | Main analyses

To investigate our primary hypothesis that SAT reduces automatic negative stereotype activation 24 h after training, we conducted a 2 (participant sex)  $\times$  2 (training location: one of two Southeastern U.S. Universities)  $\times$  2 (condition: SAT vs. control)  $\times$  2 (target race: Black vs. White) mixed factorial analysis of variance, with repeated measures

on the last factor. No significant main effects or interactions were expected or observed for either participant sex or training location. Given the estimated measurement error and sample size in this study, we had a power of  $(1 - B) = 0.22$  to detect an effect size of  $\eta_p = 0.1$ ,  $(1 - B) = 0.84$  to detect an effect of  $\eta_p = 0.25$ . As expected, there was a main effect of target race,  $F(1, 122) = 15.79$ ,  $p < .001$ ,  $n^2 = 0.12$ , indicating that categorization of targets' racial group membership following a negative African American male-stereotypic trait prime was significantly faster for Black male targets ( $M = 552.00ms$ ,  $SD = 9.02$ ) than White male targets ( $M = 573.45ms$ ,  $SD = 10.76$ ). This finding is consistent with an overall pattern of automatic negative racial stereotyping of African American male targets.

As predicted, this pattern of automatic negative racial stereotyping of Black men was pronounced and statistically significant for control participants,  $F(1, 58) = 14.13$ ,  $p < .001$ , but much smaller (one-fifth the effect size:  $n^2 = 0.20$  vs.  $0.04$ ) and not statistically significant for training participants,  $F(1, 64) = 2.67$ ,  $p = .11$ . These simple effects were examined after the overall target race  $\times$  condition interaction, although not significant, yielded a pattern consistent with our a priori directional hypotheses and replicated the pattern of findings in prior SAT studies,  $F(1, 122) = 3.64$ ,  $p = .059$  (see Table 1).

### 4 | DISCUSSION

In the present experiment, the SAT antibias training technique was found to reduce White participants' implicit racial stereotyping of Black men, relative to control participants, even 24 h after training. In addition, the bias reduction effects for SAT participants were found to generalize beyond negative stereotypic traits targeted during training to untrained negative stereotypic traits. In contrast to SAT participants, implicit racial stereotyping was pronounced for control participants who received no training, as well as for control participants who were presented with SAT stimuli but asked to make grammar judgments concerning displayed behaviors, instead of choosing situational attributions for these behaviors. Given that the same stimuli were displayed in both the SAT and grammar training conditions, the reduction in automatic stereotyping observed for SAT participants is not a function simply of the specific stimuli displayed; rather, its effects are linked to the nature of the SAT task: choosing situational attributions for stereotype-consistent behaviors. As in

**TABLE 1** Means (in milliseconds), standard deviations, mean differences, and effect sizes for categorization response times to photos of White and Black men paired with behaviors consistent with negative stereotypes of African American men

	White photo		Black photo		Mean difference	Partial $\eta^2$
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Control ( <i>n</i> = 62)	583.26	15.26	547.69	12.80	35.57	0.20
SAT ( <i>n</i> = 68)	563.63	15.16	556.32	12.71	7.31	0.04

prior research (Stewart et al., 2010), SAT effects generalized to negative-stereotypic traits not seen in training but not to any other trait type. SAT seems to be “surgical” in impacting a specific type of trait judgment, although its effects are not restricted to specific traits.

The findings of the present study reveal enduring, generalizable reduction of automatic stereotyping through SAT, an outcome of considerable applied significance. Whereas many antibias measures have been found to be quite short-lived (Forscher et al., 2019), SAT's effectiveness 24 h after training bolsters its utility as a tool to decrease automatic racial stereotyping. Given the serious negative consequences of implicit stereotyping for outcomes ranging from hiring decisions to police officers' judgments in the field (e.g., Latu et al., 2015; Spencer et al., 2016; Stewart & Branscombe, 2015), a tool that can facilitate enduring bias reduction can have substantial positive implications across a number of real-world settings.

The bias reduction observed for participants a day after completing SAT was achieved through intensive training targeting the processes underlying automatic stereotyping, versus targeting the stereotype itself. We posit that the SAT effects observed in the present study are likely due both to the targeting of fundamental attribution processes and to the intensive nature of the task. However, the study does not enable us to definitively pinpoint the mechanisms responsible for these observed effects. The present study assessed delayed effects of SAT employing the same dependent measure (the PCT) used in a prior SAT study in which implicit stereotyping was tested immediately after training (Stewart et al., 2010), allowing a more comparable cross-study comparison between these data and the present findings. However, in addition to these benefits of incorporating the PCT as our dependent measure, we also must contend with the drawback of the PCT not being well-suited to cognitive process modeling. In future research, we plan to incorporate a new measure of implicit stereotyping better suited to isolating the cognitive processes involved in implicit stereotyping and how SAT affects these processes. For example, the diffusion decision model (Ratcliff & McKoon, 2008) can be applied to response time data to estimate the rate of information accumulation, voluntary criterion for amount of information accumulation before a response, and nonddecision time, as well as variability in those parameters. Such an approach has been used to investigate processes associated with tasks such as the Implicit Association Test (Klauer et al., 2007) and allows investigators to be more specific about which mechanisms are affected by manipulations of response latency. Multinomial processing tree models (e.g., Conrey et al., 2005) might also be applied to future implicit bias measures to help isolate the relative contributions of association versus inhibition in SAT response time findings.

Of both theoretical and applied significance is the finding that SAT effects on bias reduction generalized to new negative stereotypic traits not targeted during training. This generalizability amplifies SAT's practical utility as an antibias tool. In addition, this finding hints that SAT is affecting mechanisms of implicit stereotyping and not merely deconditioning specific trained stimuli—a

distinction to be explored further in future studies. Other plans to further test SAT's generalizability include focusing on different stigmatized groups at training and test. For example, we plan to examine whether an SAT module targeting implicit African American stereotypes will reduce Latinx stereotypes. Notably, a paradigm based on SAT by independent researchers has found that the antibias effectiveness of practicing situational attributions for stereotype-consistent behaviors generalizes to different groups and settings outside of the U.S.: Levontin et al. (2013) found that practicing situational attributions reduced Israelis' bias against Arabs.

Additional plans for examination of SAT's generalizability include assessment of whether SAT will positively impact explicit, as well as implicit, bias. Although interventions targeting implicit bias reduction are generally not found to impact explicit biases (Lai et al., 2016), we theorize that SAT's focus on fundamental procedures underlying biased judgments might boost its generalizability. And, in fact, preliminary analyses of a new study suggest that SAT not only reduced implicit racial bias, but also reduced the tendency of White participants to engage in explicit dehumanization of Black targets (Stewart et al., 2022). Following SAT, White participants were significantly less likely than a control group to attribute “uniquely human” traits (e.g., cultured; refined; Haslam & Loughnan, 2012) to White than Black targets. Thus, SAT shows promise both in the longevity and generalizability of its effects.

In future research, it will, of course, be important to increase the length of time between training and measurement of implicit stereotyping to further explore how long the bias reduction effects of SAT last. Perhaps more importantly, we plan to test SAT's longevity following a series of four short training sessions rather than one extended session. Interestingly, a cognitive bias modification training approach similar to SAT in the clinical neuroscience literature has found that repeating short, intensive retraining sessions can yield long-lasting reduction in depression, anxiety, and substance abuse for clinical samples (Wiers & Wiers, 2017) and has linked these changes to specific neural processes. We predict parallel effects for repeated SAT retraining sessions.

## 5 | CONCLUSION

In the battle for racial justice in the United States, the development of evidence-based strategies to reduce implicit racial bias is critical. The present research provides new evidence that SAT is not only effective in reducing implicit biases but also that it has long-lasting, generalizable effects. These findings have both practical and theoretical importance by underscoring the malleability of implicit bias and the utility of targeting the fundamental attributional biases underlying implicit stereotyping as a bias-reduction intervention. Although additional research is needed to further explore the various characteristics of the SAT, the present findings that SAT yields positive, generalizable effects that last at least a day is an important next step in elucidating the processes and potential of this effective antibias technique.

## ACKNOWLEDGMENTS

The authors thank Ashley Myers, Amanda Culver, and Alden Treadway for outstanding research assistance and thank Becky Hagenston, Maren Stewart-Tanner, Hannah Dixon, and Niccole Marshall for invaluable feedback on previous drafts of this manuscript. This study was supported by grants awarded to Tracie Stewart from the Russell Sage Foundation and Kennesaw State University's Radow College of Humanities and Social Sciences.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author.

## ORCID

Tracie L. Stewart  <http://orcid.org/0000-0001-6654-4869>

## REFERENCES

- Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, 7(3), 136–141. <https://doi.org/10.1111/j.1467-9280.1996.tb00346.x>
- Byrd, W. C., & Ray, V. E. (2015). Ultimate attribution in the genetic era: White support for genetic explanations of racial difference and policies. *Annals of the American Academy of Political and Social Science*, 661(1), 212–223. <https://doi.org/10.1177/0002716215587887>
- Conroy, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The Quad Model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487. <https://doi.org/10.1037/0022-3514.89.4.469>
- Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, 48, 1267–1278. <https://doi.org/10.1016/j.jesp.2012.06.003>
- Devine, P. G., Plant, E. A., & Buswell, B. N. (2000). Breaking the prejudice habit: Progress and obstacles. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 185–208). Lawrence Erlbaum Associates Publishers.
- Dobbin, F., & Kalev, A. (2016, July–August). Why diversity programs fail. *Harvard Business Review*. <https://hbr.org/2016/07/why-diversity-programs-fail>
- Dobbin, F., & Kalev, A. (2019). The promise and peril of sexual harassment programs. *Proceedings of the National Academy of Sciences of the United States of America*, 116(25), 12255–12260. <https://doi.org/10.1073/pnas.1818477116>
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3), 522–559. <https://doi.org/10.1037/pspa0000160>
- Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, 72, 133–146. <https://doi.org/10.1016/j.jesp.2017.04.009>
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, 108(4), 553–561. <https://doi.org/10.1037/pspa0000016>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluation. *Social Cognition*, 38, S115–S134. <https://doi.org/10.1521/soco.2020.38.suppl.s115>
- Haslam, N., & Loughnan, S. (2012). Prejudice and dehumanization. In J. Dixon, & M. Levine *Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change* (pp. 89–104). Cambridge University Press. <https://doi.org/10.1017/CBO9781139022736.006>
- Kahn, K. B., & Davies, P. G. (2011). Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members. *Group Processes & Intergroup Relations*, 14, 569–580. <https://doi.org/10.1177/1368430210374609>
- Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology*, 55, 1–80. <https://doi.org/10.1016/bs.aesp.2016.10.001>
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, 93, 353–368. <https://doi.org/10.1037/0022-3514.93.3.353>
- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M., Burns, M., ... Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, 145, 1001–1016. <https://doi.org/10.1037/xge0000179>
- Latu, I. M., Schmid Mast, M., & Stewart, T. L. (2015). Gender biases in (inter)action: The role of interviewers' and applicants' implicit and explicit stereotypes in predicting job interview outcomes. *Psychology of Women Quarterly*, 39, 539–552. <https://doi.org/10.1177/0361684315577383>
- Levontin, L., Halperin, E., & Dweck, C. S. (2013). Implicit theories block negative attributions about a longstanding adversary: The case of Israelis and Arabs. *Journal of Experimental Social Psychology*, 49, 670–675. <https://doi.org/10.1016/j.jesp.2013.02.002>
- Macrae, C. N., Boedenhuis, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, 67, 808–817. <https://doi.org/10.1037/0022-3514.67.5.808>
- Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, 5, 461–476. <https://doi.org/10.1177/014616727900500407>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Spencer, K. B., Charbonneau, A. K., & Glaser, J. (2016). Implicit bias and policing. *Social and Personality Psychology Compass*, 10, 50–63. <https://doi.org/10.1111/spc3.12210>
- Stewart, T. L., Amoss, R. T., Weiner, B. A., Elliott, L. A., Parrott, D. J., Peacock, C., & Vanman, E. J. (2013). The psychophysiology of social action: Facial electromyographic responses to stigmatized groups predict anti-discrimination actions. *Basic and Applied Social Psychology*, 35, 418–425. <https://doi.org/10.1080/01973533.2013.823618>
- Stewart, T. L., & Branscombe, N. R. (2015). The costs of privilege and dividends of privilege awareness: The social psychology of confronting inequality. In B. Bergo & T. Nicholls (Eds.), *"I don't see color": Personal and critical perspectives on white privilege* (pp. 135–145). Penn State University Press.

- Stewart, T. L., Butts, E., Schmidt, A., Hill, S. F., DeMarco, J., Campbell, E., Marshall, N., Latu, I. M., & White, K. R. G. (2022). *Reducing explicit dehumanization and implicit stereotyping through Situational Attribution Training*. Manuscript under editorial review.
- Stewart, T. L., Latu, I. M., Kawakami, K., & Myers, A. C. (2010). Consider the situation: Reducing automatic stereotyping through alternative attribution training. *Journal of Experimental Social Psychology*, 46, 221–225. <https://doi.org/10.1016/j.jesp.2009.09.004>
- Streeter, S. (2019). Lethal force in black and white: Assessing racial disparities in the circumstances of police killings. *The Journal of Politics*, 81(3), 1124–1132.
- Wiers, C. E., & Wiers, R. W. (2017). Imaging the neural effects of cognitive bias modification training. *NeuroImage*, 151, 81–91.

**How to cite this article:** Stewart, T. L., Latu, I. M., Martin, T., Walsh, S. P., Schmidt, A., & Kawakami, K. (2022). Implicit bias reduction that lasts: Putting Situational Attribution Training to the test. *Journal of Applied Social Psychology*, 52, 1062–1069. <https://doi.org/10.1111/jasp.12912>