Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application

Article *in* Journal of Experimental Social Psychology · November 2017

DOI: 10.1016/j.jesp.2017.06.003

3 authors:

Mason D. Burns
University of Indianapolis
12 PUBLICATIONS **329** CITATIONS

SEE PROFILE

Margo Monteith
Purdue University
52 PUBLICATIONS **4,113** CITATIONS

SEE PROFILE

Laura Parker
University of Houston - Downtown
5 PUBLICATIONS **53** CITATIONS

SEE PROFILE

Journal of Experimental Social Psychology 73 (2017) 97–110

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application

Mason D. Burns, Margo J. Monteith*, Laura R. Parker

Purdue University, Department of Psychological Sciences, 703 Third Street, West Lafayette, IN 47906-2081, United States

ARTICLE INFO

Keywords:
Prejudice
Stereotyping
Implicit bias
Self-regulation
Counterstereotyping

ABSTRACT

A pressing issue concerns how to reduce stereotypic responses and discriminatory outcomes resulting from the operation of implicit biases. One possibility is that cognitive retraining, such as by repeatedly practicing counterstereotypes, can reduce implicit bias so that stereotype application will be reduced in turn. Another possibility involves motivated self-regulation, where people's awareness of their proneness to biased responses heightens negative self-directed affect, which in turn facilitates monitoring for biases and reduces stereotype application. These possibilities were tested across three experiments. In all experiments, participants who completed counterstereotype training subsequently scored lower on a measure of implicit bias, relative to untrained participants. In Experiments 1 and 2, counterstereotyping did not reduce subsequent stereotype application; in Experiment 3, counterstereotyping did reduce stereotype application, but this effect was not mediated by implicit bias scores. Participants in the motivated self-regulation condition (Experiments 2 & 3) were primed with their proneness to respond in biased ways, which increased negative self-directed affect among participants more internally motivated to respond without bias. Participants' degree of negative self-directed affect was not consistently associated with implicit bias scores. However, greater negative self-directed affect was associated with reduced stereotype application (Experiment 2) and greater rejection of racist jokes (Experiment 3). These results suggest that reductions of implicit bias through counterstereotype training do not, in turn, lead to reduced stereotype application. In contrast, the results support the viability of motivated self-regulation interventions that facilitate awareness of bias and heighten negative self-directed affect, thus creating the motivation to self-regulate stereotype application.

1. Introduction

In June of 2016, the Department of Justice released a statement calling for all 28,000 of their employees to receive training to combat unconscious racial bias. In the press release, Deputy Attorney General Sally Yates argued that this training is necessary, saying “Given that the research is clear that most people experience some degree of unconscious bias, and that the effects of that bias can be countered by acknowledging its existence and utilizing response strategies, it is essential” (Kaleem, 2016). The Department of Justice is not alone; implicit bias training initiatives are increasingly common in educational, medical, and other contexts (e.g., Badger, 2016). Although it is clear that implicit preferences and stereotypes are widespread (Nosek et al., 2007) and associated with important interpersonal and discriminatory behaviors (e.g., Corell, Park, Judd, & Wittenbrink, 2002; Dovidio, Kawakami, & Gaertner, 2002; Penner et al., 2010), the best way to combat these biases and their outcomes is less clear.

Numerous programs of research have examined cognitive retraining strategies designed to reduce bias on implicit measures (e.g., Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; for reviews, see Forscher et al., 2016; Lai et al., 2014). If proneness to implicit bias can be reduced by practicing alternative associations, the reasoning goes, implicit biases will then be less likely to create biased and discriminatory outcomes. However, with rare exceptions (Dasgupta & Rivera, 2008; Kawakami, Dovidio, & van Kamp, 2005), researchers have not empirically investigated whether the reduction of bias on implicit measures achieved with cognitive retraining translates into reduced stereotype application.

The present research tests the viability of a cognitive retraining approach for reducing stereotype application and also a motivated self-regulation approach. According to the Self-Regulation of Prejudice (SRP) model (e.g., Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002; Monteith, Lybarger, & Woodcock, 2009), increased awareness that one is prone to responding in biased ways that conflict with one's personal standards and values gives rise to negative self-directed affect (e.g., guilt). This affect is critical for motivating

* Corresponding author.
E-mail address: mmonteit@purdue.edu (M.J. Monteith).

http://dx.doi.org/10.1016/j.jesp.2017.06.003
Received 21 November 2016; Received in revised form 7 June 2017; Accepted 8 June 2017

future efforts to self-regulate one's responses to reduce the application of stereotypes and prejudices. Thus, according to this approach, negative outcomes of automatically activated bias can best be countered by increased awareness of one's biases and the motivated inhibition and replacement of their otherwise deleterious consequences.

1.1. Cognitive retraining and counterstereotyping

As summarized in the Associative-Propositional Evaluation (APE) model (Gawronski & Bodenhausen, 2006), the prototypical method for changing implicit attitudes is through incremental changes to the associative structure achieved with evaluative conditioning. Among the most widely used and powerful methods for reducing bias on implicit measures is the repeated conditioning of counterstereotypic associations with the target group (Forscher et al., 2016; Lai et al., 2014). For instance, "smart" can be repeatedly paired with "Blacks," and subsequently compete with the well-learned, existing stereotype "unintelligent" for activation. Researchers consistently find that repeatedly affirming counterstereotypes reduces stereotyping and prejudice on implicit measures (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008; Kawakami et al., 2000; Kawakami et al., 2005; Woodcock & Monteith, 2013).

Should reducing stereotype activation through counterstereotyping lead to reduced stereotype application? Central to this question is the distinction between stereotype *activation* and stereotype *application* (Kunda & Spencer, 2003), and whether retraining can reduce stereotype application as a result of decreased stereotype activation. Stereotype activation, or the degree to which a given stereotype is accessible in one's mind, is a necessary precursor for stereotype application, or the reliance on stereotypes in one's inferences, judgments, and behaviors (Kunda & Spencer, 2003). Accordingly, the cognitive retraining approach posits that if people are processing others in less biased ways due to counterstereotyping, the practically important outcome should be reduced stereotype application that is mediated by a reduction in implicit bias. Although researchers have very thoroughly investigated the malleability of automatic stereotype activation (Blair, 2002) and the reduction of implicit bias (e.g., Lai et al., 2014), only rarely have the effects of these procedures on subsequent stereotype application been examined.

Kawakami et al. (2005; see also Kawakami, Dovidio, & Van Kamp, 2007) found that gender counterstereotyping did not reduce sex discrimination in a hiring case when the counterstereotyping and hiring tasks occurred consecutively. The authors argued this was because participants corrected for the presumed influence of the counterstereotyping activity on their hiring decisions. In contrast, when a filler task (15 math problems) was placed between the counterstereotyping and hiring tasks, or a cognitive load was introduced during the hiring task, hiring discrimination was reduced. The authors argued that, in these cases, correction processes were not deployed, so the reduced accessibility of stereotypes could translate into reduced discrimination. The researchers did not, however, test whether reduced stereotype activation mediated the effects of their condition differences on subsequent stereotyping. In contrast, Dasgupta and Rivera (2008) did test for mediation, but using a counterstereotyping task that involved changes in pattern activation (Gawronski & Bodenhausen, 2006) rather than conditioning. Participants initially exposed to admired gay men and lesbians subsequently showed reduced anti-gay bias on an implicit measure and also lower discriminatory voting intentions, relative to a no training condition. However, the effect of initial exposure on voting intentions was not mediated by implicit bias, but occurred independently instead. As the authors argued, this may have been due to the very explicit nature of the voting intentions measure, which likely prompted participants to respond based on their consciously held beliefs and attitudes.

In sum, we believe that research to date does not provide clear conclusions about whether counterstereotyping reduces stereotype application and, if so, whether reduced implicit bias plays a mediating role. Also, the critical dependent variables in this previous research (hiring decisions, voting intentions) may have encouraged deliberate and intentional processing, whereas measures tapping into the more spontaneous application of stereotypes (as were used in the present research) likely are better suited for examining a "trickle down" effect of implicit bias.

Given the importance of motivation for stereotype activation and application (e.g., Kunda & Spencer, 2003), we also considered participants' self-reported internal and external motivations for responding without bias (Plant & Devine, 1998) in the present research. Whereas internal motivation refers to people's personal desire to respond without prejudice due to their egalitarian self-concept and values, external motivation stems from a desire to respond without prejudice because of pressure from others and politically correct standards. Taking these explicit motivations into account when predicting stereotype application allowed us to test whether counterstereotyping produced a reduction in stereotype application above and beyond explicit motivations. In addition, including explicit measures of motivation allowed us to test whether counterstereotyping might be especially effective for certain people. For instance, the greater people's internal motivation to respond without prejudice, the more they may benefit from practicing counterstereotyping (e.g., by concentrating more on the task), which could have favorable downstream consequences for reduced stereotype application. In contrast, given external motivation can elicit backlash (Plant & Devine, 2001), more externally motivated participants may actually show greater stereotype application following counterstereotype training. In sum, considering motivations to respond without prejudice allowed us to include tests of more nuanced versions of the hypothesis that practicing counterstereotyping would reduce implicit bias and, in turn, result in reduced stereotype application.

Although counterstereotype practice may reduce stereotype application with implicit bias playing a mediating role, there are also reasons to question whether this would be the case. Practicing counterstereotypes affects only a fraction of the full set of multifaceted associations that can contribute to stereotype activation and application (Casper, Rothermund, & Wentura, 2010; Gawronski & Bodenhausen, 2006; Kunda & Thagard, 1996; Wittenbrink, Judd, & Park, 2001). For instance, conditioning "smart" with Black people may well cause "smart" rather than "unintelligent" to be activated when one is primed with a cropped photograph of a Black face (i.e., reduced bias on an implicit measure). However, if one sees a young Black man standing on a street corner in a neighborhood with ambiguous socioeconomic cues, will "smart" be activated, or will the well-learned negative stereotypes "unintelligent" along with "criminal" and "unmotivated" be activated (see Kunda & Spencer, 2003), leading one to apply these negative stereotypes to the target? We suspected the latter outcome, so that stereotype application would result even after counterstereotyping practice reduced bias on an implicit stereotyping measure.

1.2. Motivated self-regulation

A different strategy for reducing the negative outcomes associated with implicit bias involves motivated self-regulation. According to the Self-Regulation of Prejudice (SRP) model (Monteith, 1993; Monteith, Mark, & Ashburn-Nardo, 2010; Monteith et al., 2002), when people become aware of their stereotypic and prejudiced responses that conflict with their personal standards for responding (i.e., awareness of *prejudice-related discrepancies*), a variety of consequences may follow. Especially to the extent that people's discrepant responses violate their personal motivation to respond in non-biased ways, negative self-directed affect (e.g., guilt) will be experienced (Devine, Monteith, Zuwerink, & Elliot, 1991; Monteith, Devine, & Zuwerink, 1993; Monteith & Voils, 1998). This guilt is critical for triggering subsequent regulatory processes. Specifically, through activity of the behavioral inhibition system (Gray, 1987; Gray & McNaughton, 2000) and associated conflict-detection activity (Amodio, Master, Yee, & Taylor, 2008), ongoing behavior will be briefly interrupted, and people will naturally build associations among the features surrounding their discrepant responses (e.g., the nature of the bias and to whom it was directed; the context in which it occurred), the biased response itself, and their negative affect. This retrospective activity serves to establish cues for control, which lay down the tracks for detecting biased responses in the future. Specifically, when a situation arises in which biased responses may occur again, the presence of cues for control (e.g., a member of a particular race) can trigger prospective reflection, thereby interrupting automatic processing that otherwise could give rise to a discrepant response, and enabling alternative, non-biased responses to be generated (e.g., approaching intergroup contact, individuating, generating a replacement response).

The central role that negative self-directed affect plays in instigating the motivated self-regulation of prejudiced responses has been demonstrated in various ways (e.g., Amodio, Devine, & Harmon-Jones, 2007; Monteith et al., 2002). For instance, Amodio et al. (2007) found that participants who learned they had generated racially biased responses that conflicted with their personal standards for responding subsequently experienced heightened negative self-directed affect, relative to baseline. This guilt initially predicted cortical activity indicative of reduced approach motivation, which is consistent with the retrospective activity described in the SRP model. However, the experimenters then introduced an opportunity for reparation, at which point participants' negative self-directed affect predicted interest in learning about strategies for reducing prejudice and cortical activity associated with approach motivation. In sum, the extent to which people experience negative self-directed affect after responding in biased ways is critical to subsequent self-regulatory efforts and outcomes.

In contrast to the counterstereotyping account, self-regulation should not reduce stereotype application through an immediate reduction of implicit bias. Rather, awareness of one's biases can create negative self-directed affect particularly among people who are personally motivated to respond without bias, which is essential to instigating processes that will facilitate less stereotypic and prejudiced responding in the future. Thus, the motivated self-regulation account predicts that the greater people's negative self-directed affect following heightened awareness of their proneness to biases, the less they should subsequently engage in stereotype application.

1.3. *Overview of experiments and hypotheses*

We began by testing the effects of counterstereotyping in Experiment 1. Some participants received extensive practice associating a set of counterstereotypes with Blacks, whereas other participants did not receive training. Participants went on to complete an implicit measure of stereotyping, and then a task for measuring stereotype application. Based on past findings (Gawronski et al., 2008; Kawakami et al., 2005; Woodcock & Monteith, 2013), we expected counterstereotype training to reduce the activation of stereotypes that were targeted in the training on the implicit measure, relative to the no training condition. The cognitive retraining approach predicts that counterstereotyping should also reduce subsequent stereotype application, and that this effect should be mediated by the reduction of implicit stereotyping. We also tested the more nuanced possibilities that counterstereotyping would be particularly effective at reducing stereotype activation and/or application as participants' internal motivation to respond without prejudice increased, or as their external motivation decreased.

However, we additionally considered the possibility that counterstereotyping would be altogether ineffective at prompting reduced stereotype application. According to this logic, counterstereotyping increases the accessibility of certain traits running counter to stereotypes. However, it leaves intact a multifaceted web of stereotypic associations (Casper et al., 2010; Kunda & Thagard, 1996; Wittenbrink et al., 2001) that can be applied to targets.

Experiment 1 also included a condition in which participants were warned that they may unwittingly rely on stereotypes during the study, and they were asked to avoid stereotyping. We included this condition because our stereotype application measure was intended to be subtle and not obviously involve stereotyping, in which case a mere warning not to rely on stereotypes would be insufficient to reduce stereotype application. Although the warning may heighten motivations to respond in nonstereotypic ways, we did not expect it to trigger sensitivity to subtle stereotyping contexts and the need for self-regulation. Other research has similarly used a warning not to rely on racial cues to establish that mere conscious intention cannot drive down subtle stereotyping effects (e.g., Payne, Lambert, & Jacoby, 2002).

Experiments 2 and 3 likewise included counterstereotyping and no training conditions, but additionally included a condition for examining the effects of motivated self-regulation on stereotype activation and application. Specifically, participants completed the Should-Would Discrepancy Questionnaire (Monteith & Voils, 1998), which involves rating how one *should* respond in situations involving the stereotyped group, followed by how one *would* respond in these situations. This procedure was used to prime people's proneness to prejudice-related discrepancies, and therefore to activate negative self-directed affect. That is, past research has shown that participants who rate their *would* responses as more prejudiced than their *shoulds* subsequently experience negative self-directed affect, particularly to the extent that they hold low-prejudiced attitudes and egalitarian goals (Devine et al., 1991; Monteith & Voils, 1998; Monteith et al., 1993). Replicating this past research, we expected that participants with larger should-would discrepancies would experience greater negative self-directed affect, and that this effect likely would be exaggerated among people who reported being more internally motivated to respond without prejudice.

More importantly, we expected that the participants experiencing greater negative self-directed affect would subsequently show reduced stereotype application. That is, in line with prior research establishing the importance of negative self-directed affect for the instigation of self-regulatory processes (e.g., Amodio et al., 2007; Monteith et al., 2002), our central prediction was that negative self-directed affected would be negatively related to stereotype application.

In sum, Experiments 1–3 tested the hypothesis that counterstereotyping would reduce bias on an implicit stereotyping measure, along with competing predictions about whether this reduced bias would in turn lead to reduced stereotype application. Experiments 2 and 3 additionally tested the motivated self-regulation hypothesis that negative self-directed affect resulting from awareness of one's proneness to biases would not help participants to evade stereotype activation, but it would be associated with reduced stereotype application.

2. Experiment 1

2.1. *Method*

2.1.1. *Power analyses*

We conducted power analysis simulations in SAS version 9.4 (see Lane & Hennes, in press) to estimate the number of participants needed to detect significant effects of interest. For the effect of counterstereotype training on implicit bias as assessed with the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), previous research finds effect sizes ranging from a bit below medium ($d = 0.34$–$d = 0.43$; Lai et al., 2016, Experiment 2; Woodcock & Monteith, 2013, Experiment 2) to large ($\eta^2 = 0.14$; Gawronski et al., 2008, Experiment 2). Accordingly, we assumed a medium effect size ($d = 0.50$) for this effect. We also assumed medium effect sizes ($d = 0.50$) for the effect of counterstereotyping on stereotype application (see Kawakami, Dovidio, et al., 2007) and for the effect of stereotype warning on stereotype application (see Monteith, Spicer, & Tooman, 1998). Finally, we simulated interactions between internal motivation and experimental condition and between external motivation and experimental condition. Because prior research has not investigated these interactions, we defaulted to assume medium effect sizes. Results indicated that we needed 155 participants to detect all relevant effects with 80% power.

2.1.2. Participants

Participants were 182 Amazon Mechanical Turk (MTurk) workers who were compensated $0.75. We removed 19 participants who identified as Black because the study assessed stereotyping in relation to Blacks. Data from three participants who took over 1 h and 45 min to complete the experiment (> 3 *SD*s from mean time) and from two participants who had missing data on all primary dependent variables were removed. The final sample included 158 participants (67% female; 86.1% White, 6.3% Asian, 5.7% Hispanic, and 1.9% "other"; $M_{\text{age}} = 35.49$, $SD_{\text{age}} = 11.80$).

2.1.3. Design

Participants were randomly assigned to a counterstereotyping, stereotype warning, or no training condition. Internal and external motivations to respond without prejudice (IMS and EMS, Plant & Devine, 1998) varied continuously.

2.1.4. Procedure

After providing consent, participants were routed to complete the experiment through the online survey platform, SocialSci. Participants were informed that they would complete several different tasks related to information processing about people, categorizing information, and reasoning. Note that all measures, manipulations, and exclusions in this and the subsequently reported experiments have been disclosed.

2.1.4.1. Experimental conditions. Participants in the counterstereotyping condition read that their task was to identify counterstereotypes, or "words that represent the opposite of stereotypes." As in previous research (e.g., Kawakami, Dovidio, et al., 2007) the counterstereotyping task included six blocks of 80 trials, for 480 trials total. On each trial, a photograph was randomly selected from a pool of 12 White and 12 Black male faces and paired with two randomly selected words (one stereotypic and the other counterstereotypic). They were instructed to use a keystroke to select the counterstereotypic word on every trial. In so doing, participants repeatedly affirmed the positive counterstereotypes (Gawronski et al., 2008; Kawakami et al., 2005; Kawakami, Dovidio, et al., 2007). Prior to the task, participants viewed the list of counterstereotypes for Blacks (achiever, ambitious, educated, intelligent, motivated, productive, responsible, reliable, smart, and wealthy) and for Whites (complainer, deadbeat, dumb, stupid, lazy, pathetic, unreliable, failure, poor, unemployed). Participants were encouraged to respond as quickly as possible but to try to limit the number of errors. When errors were made, a red "X" appeared until the participant responded correctly.

In the stereotype warning condition participants read that in prior research, some people appeared to rely on stereotypes about Black people as they completed some of the tasks, perhaps without meaning to do so. The instructions went on to note that "we would like to collect data in which people are NOT letting stereotypes influence their responses. Thus, as you complete the tasks in the study today, please do your best to avoid thinking about Blacks in biased or stereotypic ways."

Participants in the no training condition received no special instructions.

2.1.4.2. IAT. Next participants completed an IAT (Greenwald et al., 1998) to assess the ease with which they could pair "motivated" words and "unmotivated" words with pictures of Whites and Blacks. In all experiments reported herein, the IAT used the same words as the counterstereotyping task, but different cropped photos of Blacks and Whites as stimuli (three photos each of Black and White men, taken from the Race IAT found on Project Implicit, https://implicit.harvard.edu/implicit/). The IAT had seven blocks of trials: 1) White/Black; 2) Motivated/Unmotivated; 3) practice: White-Motivated/Black-Unmotivated; 4) test: White-Motivated/Black-Unmotivated; 5) Black/White; 6) practice: Black-Motivated/White-Unmotivated; 7) test: Black-Motivated/White-Unmotivated. Forty practice and 100 test trials were used for the dual categorization blocks and 20 trials were used for the other blocks, with one exception: In Experiment 1 only, 10 trials were included for block 5 due to a programming error. Participants received one of two IAT orders to counterbalance whether stereotype compatible (White-Motivated/Black-Unmotivated) preceded or followed stereotype incompatible categorizations (White-Unmotivated; Black-Motivated).

2.1.4.3. Filler task. Participants then completed a filler task supposedly concerning logical reasoning. Across 20 items, participants identified the underlying relationship between pairs of words, and then selected a word to extend this relationship with another word (e.g., Unique is to Copy as Accident is to…? A) Mistake B) Intend C) Occur D) Incident E) Injury).

2.1.4.4. Stereotypic inferences. Participants learned that the next task assessed their ability to form inferences about people based on a single photograph and brief description. To illustrate, participants were provided with an example showing a picture of a man with the description, "This person can be found in a theater." Participants were instructed to type their inference in the box provided, such as "movie fan" or "actor." Embedded within 40 trials were 6 critical trials with images of Black men and descriptions that could yield either stereotype-consistent or nonstereotypic inferences. For example, "This person can be found on the streets" could yield a stereotypic (e.g., homeless) or nonstereotypic (e.g., tourist) inference. The remaining critical prompts were "This person uses needles for recreation," "This person depends on money from the government," "This person can be found behind bars," "This person deals with a lot of drugs," and "This person is good at getting into locked doors." Importantly, participants did not choose a response from options that we provided. They generated their own inferences and typed each one into a response box on each trial.

We conducted a pilot study to establish construct validity for use of this task to assess racial stereotyping among a separate sample of 122 non-Black MTurk participants (64% female; 85.2% White, 5.7% Asian, 5.7% Hispanic or Latino, 2.5% Native American, 0.8% Middle Eastern (Non-Arab); $M_{\text{age}} = 38.34$, $SD_{\text{age}} = 12.82$), who were paid $1.00 for participating. Participants first completed the stereotype inference task, and participants' responses were later coded to determine the percent of stereotypic responses (intercoder agreement = 98%; $M = 0.57$ or 57% stereotypic responses, $SD = 0.25$; range = 0–1.00). Participants then completed the stereotyping IAT (as described above; scored according to Greenwald, Nosek, & Banaji's, 2003 algorithm), the IMS and EMS scales (1 = strongly disagree, 9 = strongly agree; IMS $\alpha = 0.89$; EMS $\alpha = 0.90$), and the Symbolic Racism 2000 Scale (Henry & Sears, 2002; 1 = strongly disagree, 7 = strongly agree; $\alpha = 0.88$). Either before or after completing the IAT, participants also completed the Should-Would Discrepancy Scale (Monteith & Voils, 1998) and their current affect. These latter measures are relevant to interpreting results in Experiment 2; therefore, we will return to a discussion of these measures and related results in the context of Experiment 2.

As shown in Table 1, participants were significantly more likely to apply stereotypes during the photo-inference task as their IMS scores decreased, and as their Symbolic Racism and IAT scores increased. Using IAT, IMS, EMS, and Symbolic Racism scores to predict stereotypic inferences simultaneously, we found unique contributions for IAT scores, $t(113) = 2.22$, $B = 0.13$, $SE = 0.06$, $\beta = 0.20$, $p = 0.03$, and for the IMS $t(113) = 2.53$, $B = -0.04$, $SE = 0.02$, $\beta = -0.25$, $p = 0.01$. Although we did not investigate responses to these prompts when paired with identical pictures except with White targets, these

Table 1
Descriptives and correlations among measures, stereotypic inferences pilot study.

	M	*SD*	Stereotypic inferences	IMS	EMS	Symbolic racism
Stereotypic inferences	0.57	0.25	–			
IMS	7.50	1.76	− 0.31***	–		
EMS	4.20	2.40	0.14	− 0.22*	–	
Symbolic racism	3.73	1.01	0.28**	− 0.52**	0.33***	–
IAT	0.40	0.39	0.26**	− 0.15	0.21*	0.23*

Note. Ns = 122 except for IAT analyses, where N = 118.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

findings support our contention that the stereotype application task does indeed tap into racial bias, with unique relations to both an implicit measure of stereotyping and to an explicit measure of internal motivation to respond without prejudice.

2.1.4.5. IMS and EMS. For the last task in Experiment 1, participants were informed that the researchers were interested in attitudes and beliefs about people, and they completed the IMS and EMS scales (Plant & Devine, 1998). The five IMS items assess the extent to which people are motivated to respond without prejudice due to personal values and standards, and the five EMS items assess social pressure and societal norms as the impetus for responding without prejudice. Ratings were made on 7-point scales.

At the end of the experiment, participants were thanked and debriefed.

2.2. Formation of indexes

2.2.1. IMS and EMS

Ratings for IMS items were averaged to form an index ($M = 5.69$, $SD = 1.28$, $\alpha = 0.86$) as were ratings for EMS items ($M = 3.49$, $SD = 1.72$, $\alpha = 0.89$). Scores on these indexes did not vary according to participants' experimental condition, IMS $F(2,\ 155) = 0.61$, $p = 0.54$, $\eta_p^2 = 0.01$; EMS $F(2, 155) = 1.42$, $p = 0.24$, $\eta_p^2 = 0.02$.

2.2.2. IAT scores

IAT scores were computed using the scoring algorithm recommended by (Greenwald et al., 2003). Nine participants' data were missing because they had response latencies of < 300 ms on > 10% of their trials, indicating careless responding. More positive IAT scores reflect greater ease of pairing Whites with the "motivated" words and Blacks with the "unmotivated" words, compared to the reverse pairings.

2.2.3. Stereotypic inferences

Two coders categorized each participant's six responses from the photo-description inference task as either stereotypic or nonstereotypic (inter-coder agreement = 99%). Four participants completed the task incorrectly (e.g., typing responses such as "good to know" or "yes"), resulting in missing data. An index reflecting the percentage of stereotype-consistent responses was formed; overall, nearly half of the responses provided by participants were stereotypic ($M = 0.46$ or 46%, $SD = 0.31$; range = 0–1.00).

3. Results and discussion

Interrelations among Experiment 1's variables are shown in Table 2. Replicating the pilot study, stereotypic inferences correlated negatively with the IMS and positively with the IAT.

Table 2
Correlations among measures, Experiment 1.

	IMS	EMS	IAT
IMS	–		
EMS	− 0.12	–	
IAT	− 0.15†	0.01	–
Stereotypic inferences	− 0.25*	0.10	0.23*

† $p < 0.07$.
* $p < 0.01$.

3.1. Data analytic approach

Unless otherwise noted, hierarchical regression analyses were used in this and the subsequently reported experiments using the following steps. We entered and assessed IMS and EMS (mean centered) on Step 1. The main effect for experimental condition was assessed on Step 2 by evaluating the increment in R^2 observed when two dummy codes capturing experimental condition (DC1 and DC2) were entered as a set. Dummy coding was accomplished as follows: no training: DC1 = 0, DC2 = 0; counterstereotyping: DC1 = 0, DC2 = 1; stereotype warning (Experiment 1) or discrepancy salience (Experiments 2 and 3): DC1 = 1, DC2 = 0. Thus, the DC1 carries the comparison between the no training and counterstereotyping condition, and DC2 carries the comparison between the no training and stereotype warning condition. The coding was modified as appropriate for making other comparisons (e.g., between counterstereotyping and discrepancy salience conditions). To examine whether any effects of experimental condition were moderated by IMS or EMS, DC1 X IMS and DC2 X IMS were entered as a set and assessed on Step 3, and DC1 X EMS and DC2 X EMS were entered as a set and assessed on Step 4.

Our research questions were not pertinent to testing whether IMS interacted with EMS, nor did we predict 3-way interactions between IMS, EMS, and experimental condition. Inclusion of these terms in analyses yielded only two significant effects across all dependent variables and experiments that did not qualify other results reported below. Descriptions of these results can be found in the supplemental materials.

3.2. IAT performance

IMS marginally predicted IAT performance, $t(146) = 1.87$, $B = -0.04$, $SE = 0.02$, $\beta = -0.15$, $p = 0.06$. Participants demonstrated less IAT bias as their IMS increased. More important to our main hypotheses, DC1 and DC2 together produced a significant increment in R^2, $\Delta R^2 = 0.02$, $F(2, 144) = 3.65$, $p = 0.03$, indicating a significant main effect for experimental condition. As shown in Table 3, participants in the counterstereotyping condition had significantly lower IAT scores than participants in the no training condition, $t(144) = 2.33$, $B = -0.15$, $SE = 0.06$, $\beta = -0.22$, $p = 0.02$. IAT scores of participants who were warned not to stereotype Blacks were comparable to scores of no training participants, $t(144) = 0.03$, $B = -0.002$, $SE = 0.06$, $\beta = -0.003$, $p = 0.97$, and were significantly greater than the scores in the counterstereotyping condition, $t(144) = 2.43$, $B = 0.15$, $SE = 0.06$, $\beta = 0.24$, $p = 0.02$. Importantly, the main effect

Table 3
IAT and stereotypic inferences as a function of experimental condition, Experiment 1.

	IAT			Stereotypic inferences		
	N	Mean	*SD*	N	Mean	*SD*
No training	46	0.32	0.36	50	0.49	0.32
Stereotype warning	61	0.31	0.25	59	0.44	0.32
Counterstereotype training	42	0.16*	0.25	45	0.44	0.30

* Cell mean differs from no training and stereotype warning conditions, $p < 0.05$.

of experimental condition was not moderated by IMS or EMS, *ps* > 0.35, indicating that counterstereotype training reduced implicit bias regardless of participants' consciously held motivations to respond without prejudice.

3.3. Stereotypic inferences

There was a main effect of IMS on stereotypic inferences. As IMS increased, participants were less likely to generate stereotypic responses, $t(151) = 3.08$, $B = -0.06$, $SE = 0.02$, $\beta = -0.24$, $p = 0.002$. More relevant to our research hypotheses, the effect of experimental condition was not significant, $\Delta R^2 = 0.01$, $F(2, 149) = 0.36$, $p = 0.70$. As shown in Table 3, participants in the counterstereotyping, stereotype warning, and no training conditions were all equally likely to rely on stereotypes. Furthermore, this was equally true regardless of participants' motivations to respond without prejudice, as experimental condition did not interact with IMS or EMS, *ps* > 0.30. Thus, even though implicit race bias was reduced through counterstereotyping, this training did not reduce the likelihood of labeling Black targets as homeless, drug addict, poor, criminal, and the like. Even the warning not to stereotype Blacks did not affect participants' likelihood of characterizing Blacks in stereotypic ways, we assume because a simple warning does not help people to realize that they may make subtle stereotypic inferences in the context of the task.

4. Experiment 2

In Experiment 1, cognitive retraining did not reduce participants' application of negative stereotypes to Blacks, despite significantly reducing stereotyping on an implicit measure. In addition, warning participants that stereotypes could influence their responses did not prompt reduced stereotype application. Experiment 2 expanded on Experiment 1 by contrasting the counterstereotype training strategy with the motivated self-regulation strategy. Non-Black participants assigned to a discrepancy salience condition considered and reported how they should respond in various situations involving Blacks, and then how they would respond in these situations, by completing the Should-Would Discrepancy Questionnaire (Monteith & Voils, 1998). We expected participants whose *woulds* were more prejudiced than their *shoulds* to experience heightened negative self-directed affect, particularly to the extent that their discrepancies violated their consciously held egalitarian goals (Devine et al., 1991; Monteith et al., 1993; Monteith & Voils, 1998). Importantly, people vary in the extent to which they are prone to should-would discrepancies, and in the extent to which discrepancies elicit negative self-directed affect. Thus, we did not expect that merely being in the discrepancy salience condition would lead to less stereotypic responding. Rather, given the significance of negative self-directed affect for triggering self-regulatory processes, our prediction was that participants who experienced more negative self-directed affect after completing the discrepancy questionnaire would subsequently show reduced stereotype application.

4.1. Power analyses

We conducted power analysis simulations in SAS version 9.4 (see Lane & Hennes, in press) to estimate the number of participants needed to detect significant findings for the effects of theoretical interest. We used the effect size for counterstereotyping versus no training from Experiment 1 ($d = 0.39$) for estimating the anticipated effect on IAT scores. We used a medium effect size ($d = 0.50$) based on Kawakami, Dovidio, et al. (2007) for estimating the number of participants needed to detect a significant effect of counterstereotyping on stereotype application. To estimate the number of participants needed to detect a significant effect of negative self-directed affect on stereotype application within the discrepancy salience condition, we averaged effect sizes from Monteith et al. (2010; $r = 0.20$) and Czopp, Monteith and Mark (2006; $r = 0.29$). We assumed medium effect sizes to determine the sample size needed to detect significant two-way interactions between IMS and experimental condition and EMS and experimental condition. Results indicated that we needed at least 195 participants in order to detect all main effects and interactions of theoretical interest with 80% power.

4.2. Participants

Participants were 246 MTurk workers who were compensated $0.75 for their participation. As in Experiment 1, we removed participants who identified as Black ($n = 25$). We collected page progression data that allowed us to identify and exclude data from seven participants who took long breaks (> 30 min) between critical experimental tasks (e.g., between counterstereotype training and IAT completion). Data from two participants with missing data on the primary dependent variables and from three participants who had recently completed another MTurk experiment using the stereotypic inference task were also excluded. The final sample included 209 participants (59% female; 85.2% White, 9.1% Asian, 4.3% Hispanic, and 1.4% "other"; $M_{\text{age}} = 35.26$, $SD_{\text{age}} = 12.71$).

4.3. Design

Participants were randomly assigned to a counterstereotyping, discrepancy salience, or no training condition. IMS and EMS varied continuously.

4.4. Procedure

Experiment 2 was identical to Experiment 1 with a few exceptions. Most importantly, we replaced the stereotype warning condition with the discrepancy salience condition. These participants completed the Should-Would Discrepancy Questionnaire (Monteith & Voils, 1998), which involves initially rating (1 = strongly disagree; 7 = strongly agree) how they personally believe they *should* respond across 16 situations involving Blacks (e.g., "I should react to all my supervisors the same, regardless of their race"). Next, participants rated (1 = strongly disagree; 7 = strongly agree) how they actually *would* respond across 16 parallel situations involving Blacks (e.g., "I would feel awkward having a Black supervisor"). Immediately after the discrepancy questionnaire, participants in the discrepancy salience condition rated the extent to which 29 affect items applied to their current feelings (1 = does not apply at all; 7 = applies very much). We were particularly interested in whether completing the discrepancy questionnaire activated negative self-direct affect (e.g., guilt, disappointment with the self), which has been associated with increased efforts and success at regulating biased responses (e.g., Amodio et al., 2007; Monteith et al., 2002). Note that we included the affect measure in the discrepancy salience condition only, as this is the only condition in which affect was theoretically relevant.

Remaining aspects of the procedure included two minor changes from Experiment 1. First, we shortened the counterstereotyping task to 400 trials to reduce tedium. Second, we substituted "This person often handles other people's money" for "This person deals a lot with drugs" in the stereotype application task.

4.5. Formation of indexes

4.5.1. IMS and EMS

Ratings for IMS items were averaged to form an index ($M = 5.68$, $SD = 1.41$, $\alpha = 0.88$) as were ratings for EMS items ($M = 3.80$, $SD = 1.55$, $\alpha = 0.84$). IMS and EMS did not vary according to participants' experimental condition, IMS $F(2, 206) = 0.06$, $p = 0.94$, $\eta_p^2 = 001$; EMS $F(2, 206) = 1.32$, $p = 0.27$, $\eta_p^2 = 0.01$.

4.5.2. IAT

IAT scores were computed as in Experiment 1. Data from four participants were missing due to careless responding (latencies < 300 ms on > 10% of their trials).

4.5.3. Stereotypic inferences

Coding of stereotypic responses was performed (inter-coder agreement = 99%; missing data for two participants who completed the task incorrectly), and the percentage of stereotype-consistent responses was computed ($M = 0.49$, $SD = 0.31$, range = 0–1.00).

4.5.4. Discrepancy scores and affect

With data from the discrepancy salience condition ($n = 82$), we used standard methods (e.g., Monteith & Voils, 1998) for computing a discrepancy score for each participant. Specifically, each *should* rating was subtracted from the corresponding *would* rating, and we then averaged across the 16 difference scores ($M = 0.80$, $SD = 0.67$, $\alpha = 0.69$). Thus, larger discrepancy scores reflect a greater tendency to respond to Blacks in ways that are more stereotypical than participants' personal standards suggest are appropriate.

Also following past research (Devine et al., 1991; Monteith, 1993; Monteith et al., 1993; Monteith & Voils, 1998; Monteith et al., 2002), we formed an index of negative self-directed affect, or *Negself*, by averaging the following items: helpless, shameful, angry at myself, embarrassed, disgusted with myself, self-critical, guilty, and regretful ($M = 2.43$, $SD = 1.38$, $\alpha = 0.93$). (For details concerning other affect indexes from Experiments 2 and 3 that were not as relevant to our hypotheses, see the Supplemental materials.)

4.6. Results and discussion

Interrelations among all measures are shown in Table 4.

4.6.1. Discrepancy scores and negative self-directed affect (discrepancy salience condition only)

Although not the main focus of the present research, we first set out to test whether participants' discrepancies between *should* and *would* ratings had affective consequences in the form of Negself feelings that would replicate past findings (e.g., Devine et al., 1991; Monteith et al., 1993; Monteith & Voils, 1998). Hierarchical regression analyses were performed to predict Negself, entering and assessing IMS, EMS and discrepancy score (all mean centered) on Step 1, 2-way interactions on Step 2, and the 3-way interaction on Step 3. Consistent with past findings, as participants' discrepancy score increased, they reported greater Negself, $t(76) = 4.08$, $B = 0.89$, $SE = 0.22$, $\beta = 0.43$, $p < 0.001$; in addition, the interaction between discrepancy and IMS was significant, $t(73) = 3.65$, $B = 0.57$, $SE = 0.15$, $\beta = 0.35$, $p < 0.001$. As shown in Fig. 1, participants who reported relatively low levels of internal motivation to respond without prejudice experienced little Negself, regardless of their discrepancy scores, $b = 0.24$, $se = 0.27$, $p = 0.38$. However, at relatively high levels of IMS, participants with larger discrepancy scores reported significantly greater Negself than participants with smaller discrepancy scores, $b = 1.69$, $se = 0.30$, $p < 0.001$. No other effects were significant.

Table 4
Correlations among measures, Experiment 2.

	Full sample ($n = 209$)		
	IMS	EMS	IAT
IMS	–		
EMS	− 0.15*	–	
IAT	0.06	0.18*	–
Stereotypic inferences	− 0.21**	0.18*	0.30**

	Discrepancy salience condition ($n = 82$)				
	IMS	EMS	IAT	Stereotypic inferences	Discrepancy total
Discrepancy total	− 0.23*	0.28*	0.15	0.15	–
Negative self-directed affect	0.05	0.19	0.16	− 0.20	0.43**

* $p < 0.05$.
** $p < 0.01$.

4.6.2. IAT performance

We used the same hierarchical regression approach to analyzing IAT scores as described in Experiment 1, regressing IAT scores on IMS and EMS (centered); dummy coded experimental condition; and interactions between experimental condition and IMS and EMS. The main effect for EMS was significant, such that IAT scores increased as EMS increased, $t(196) = 2.73$, $B = -0.04$, $SE = 0.02$, $\beta = 0.19$, $p = 0.01$. More importantly, the set of dummy codes capturing experimental condition was significant, $\Delta R^2 = 0.06$, $F(2, 194) = 6.40$, $p = 0.002$. As shown in Table 5, participants in the counterstereotyping condition showed significantly less implicit race bias than participants in the no training condition, $t(194) = 3.40$, $B = -0.20$, $SE = 0.06$, $\beta = -0.28$, $p = 0.001$. Unexpectedly, participants in the discrepancy salience condition also showed significantly less implicit bias compared to the no training condition, $t(194) = 2.79$, $B = -0.15$, $SE = 0.06$, $\beta = -0.23$, $p = 0.01$, and the discrepancy salience and counterstereotyping conditions did not differ from one another, $t(194) = 0.82$, $B = 0.04$, $SE = 0.05$, $\beta = 0.07$, $p = 0.41$. Of note, as in Experiment 1, we did not find evidence that IMS or EMS interacted with experimental condition to predict IAT performance, $ps > 0.17$.

Although we had expected participants in the counterstereotyping condition to manifest less implicit bias on the IAT stereotyping measure than no training participants, the finding that discrepancy salient participants also showed less IAT bias than no training participants was both surprising and theoretically unexpected. Using the results from the pilot study reported in connection with Experiment 1 (see *Stereotypic Inferences* section of *Procedure*, Experiment 1), we were able to test whether this effect replicated by determining whether completing the Should-Would Discrepancy Scale and affect prior to the IAT, compared to after the IAT, affected IAT performance. In this separate sample of participants, we found that IAT scores were slightly but not significantly higher when the Should-Would Discrepancy scale and affect measures were completed before the IAT ($M = 0.46$, $SD = 0.41$) rather than after the IAT ($M = 0.34$, $SD = 0.34$), $t(116) = 1.75$, $SE = 0.07$, $p = 0.083$. These results run contrary to Experiment 2 findings. In addition, as will be seen, we did not replicate the finding that discrepancy salient participants had lower IAT scores than no training participants in Experiment 3. Thus, we believe that the most parsimonious interpretation of the unexpectedly low IAT scores in the discrepancy salience condition in Experiment 2 is that they were coincidental and not replicated.

4.6.3. Stereotypic inferences

The hierarchical regression analysis predicting stereotypic inferences revealed a significant main effect for IMS, $t(198) = 2.72$, $B = -0.04$, $SE = 0.02$, $\beta = -0.19$, $p = 0.01$, with stereotypic responses decreasing as internal motivation increased. EMS also had a significant main effect, $t(198) = 2.13$, $B = 0.03$, $SE = 0.01$, $\beta = 0.15$, $p = 0.03$, with stereotypic responses increasing as external motivation increased.

Replicating Experiment 1, we found that the main effect for experimental condition was not significant, $\Delta R^2 = 0.001$, $F(2, 196) = 0.15$, $p = 0.86$. As shown in Table 5, participants in the counterstereotyping, discrepancy salience, and no training conditions were all equally likely to rely on stereotypes when making inferences about

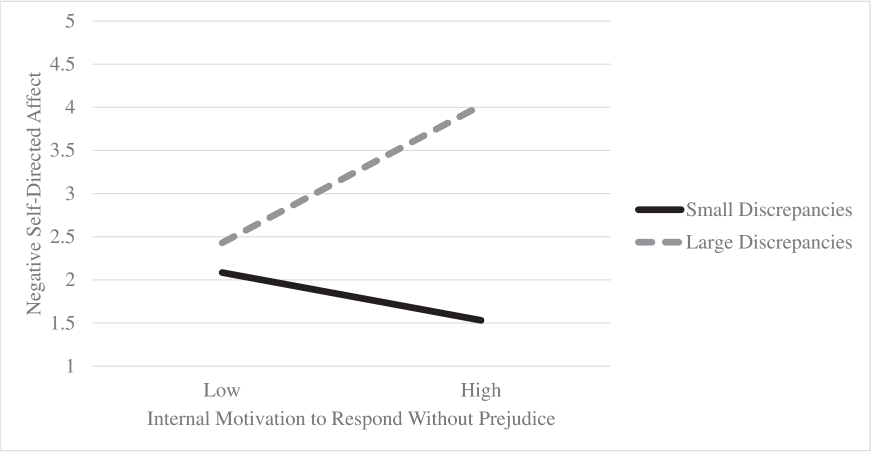

Fig. 1. The effect of IMS and should-would discrepancies on negative self-directed affect, Experiment 2.

Table 5
IAT and stereotypic inferences as function of experimental condition, Experiment 2.

	IAT			Stereotypic inferences		
	N	Mean	*SD*	N	Mean	*SD*
No training	58	0.41	0.32	59	0.50	0.31
Discrepancy salience	80	0.28*	0.34	80	0.48	0.33
Counterstereotype training	61	0.22*	0.30	62	0.50	0.29

* Cell means differ from no training condition at $p < 0.05$.

Blacks. Thus, as in Experiment 1, participants who had initially practiced counterstereotypes were not less likely to apply stereotypes during the stereotypic inference task. Furthermore, neither IMS nor EMS interacted with experimental condition to predict stereotypic inferences, *ps* > 0.14, indicating that the effect of training on stereotype application was not dependent on participants' motivations to respond without prejudice.

Next, we tested our hypothesis that heightened motivation to self-regulate biased responses, evidenced by discrepancy-associated Negself feelings, would be associated with greater avoidance of biased responses on the photo-description task. Thus, Negself, along with IMS and EMS, were used to predict stereotypic inferences within the discrepancy salience condition. Consistent with results reported above, IMS was negatively related to stereotypic responding, $t(74) = 1.88$, $B = -0.05$, $SE = 0.03$, $\beta = -0.21$, $p = 0.06$, and EMS was positively related, $t(74) = 2.49$, $B = 0.05$, $SE = 0.02$, $\beta = 0.28$, $p = 0.01$. More importantly, the analysis also revealed a significant main effect for Negself, $t(74) = 2.16$, $B = -0.06$, $SE = 0.03$, $\beta = -0.23$, $p = 0.03$. Participants who reported greater guilt and disappointment with the self over their awareness of their own biased tendencies generated significantly fewer stereotypic inferences. This finding supports our prediction that activating the motivation to self-regulate, as reflected in discrepancy-associated negative self-directed affect, would be related to a reduction in stereotype application.[1]

However, perhaps participants who reported greater negative self-directed affect in our experiment would have generated relatively few stereotypic inferences even if this affect had not been recently activated. To establish that this is not the case, we return again to the pilot study reported in Experiment 1 (see *Stereotypic Inferences* section of *Procedure*, Experiment 1). Recall that participants in this study always completed the stereotypic inference task prior to completing the Should-Would Discrepancy scale and affect measure. Thus, if Negself is unrelated to stereotypic inferences, we can conclude that people higher on Negself do not characteristically generate relatively few stereotypic inferences. Indeed, the relation between Negself and stereotypic inferences in the pilot study was not significant, $r\ (122) = -0.04$, $p = 0.62$; controlling for IMS and EMS, $r\ (118) = -0.05$, $p = 0.59$. This finding supports our motivated self-regulation account, suggesting that the recent experience of Negself in relation to one's biased responses predicts people's ability to regulate and reduce the subsequent application of stereotypes.

In sum, whereas practicing counterstereotypes did not reduce participants' likelihood of using stereotypes when making inferences about Blacks, activating the motivation to self-regulate, as operationalized by the experience of negative self-directed affect, did.

5. Experiment 3

Across Experiments 1 and 2, counterstereotype training reduced implicit bias but did not affect the likelihood of generating stereotypic inferences of Blacks. In contrast, the experience of discrepancy-related negative self-directed affect was related to fewer stereotypic inferences. Experiment 3 was designed to replicate and extend Experiment 2 with a different stereotype application task. Similar to the stereotypic inference task used in Experiments 1 and 2, our intention was to use a task that could tap into rather spontaneous stereotyping. We decided to have participants evaluate jokes that played on stereotypes of Blacks. There can be little doubt that racial humor can be used deliberately to communicate intergroup antipathy and also consciously suppressed to abide with salient non-prejudiced norms (Crandall, Eshleman, & O'Brien, 2002; Experiment 3). However, research also reveals that people can be biased by stereotypic portrayals and react favorably to such jokes before “thinking twice” about laughing (Monteith & Voils, 1998, Experiment 3), particularly if the motivation to self-regulate biases has not been recently activated (Monteith, 1993, Experiment 2). Thus, the use of racial jokes that play on stereotypes seemed an appropriate task for assessing whether reduced stereotype activation achieved through counterstereotyping or heightened motivation for self-regulation would decrease stereotype application. Furthermore, because disparagement humor is associated with greater tolerance for discrimination bias and “releases” biases among higher prejudiced individuals (Ford & Ferguson, 2004; Ford,

[1] Although our main motivated self-regulation hypotheses concerned the relation between Negself and stereotypic responding, our data also allowed us to conduct a more elaborate moderated mediation analysis to test whether the interaction between IMS and discrepancy scores indirectly reduced stereotype application through Negself. Using Hayes's (2013) PROCESS (Model 8) analysis with 5,000 bootstraps we predicted stereotype application with discrepancy scores as the independent variable, IMS as the moderator, Negself as the mediator, and EMS as a covariate. At high levels of IMS, the indirect effect of discrepancy on stereotypic inferences, through Negself, was significant, 95% CI [−0.25, −0.04]. At low levels of IMS, as expected, the indirect effect was not significant, 95% CI [−0.08, 0.02]. These results support the motivated self-regulation account and demonstrate that increased bias awareness resulted in Negself, in particular for internally motivated participants. This experience of Negself, in turn, predicted a reduction of stereotypic inferences.

Richardson, & Petit, 2015), determining effective ways to reduce the acceptance of racist jokes is of practical interest.

5.1. Method

5.1.1. Power analyses

We again conducted power analyses in SAS version 9.4 (Lane & Hennes, in press). Averaging across Experiments 1 and 2, we assumed an effect size of $d = 0.44$ for the effect counterstereotype training versus no training condition for the IAT. We used $d = 0.40$ from Experiment 2 for the effect of the discrepancy salience versus no training for predicting IAT performance, although we did not expect this effect to be significant. We assumed $d = 0.50$ (Kawakami, Dovidio, et al., 2007) for counterstereotyping versus no training when predicting stereotype application, and $r = 0.24$ (Experiment 2) for the effect of negative self-directed affect on stereotype application. Additional power analyses to determine the sample size needed to detect significant interactions between IMS or EMS and experimental condition were run assuming medium effect sizes. Results indicated that we needed 180 participants to detect all of these effects with 80% power.

5.1.2. Design

Participants were randomly assigned to a counterstereotyping, discrepancy salience, or no training condition. IMS and EMS varied continuously.

5.1.3. Participants

Participants were 249 MTurk workers who were compensated $1.00 for their participation. We removed 26 Black participants. Based on page progression data, eight participants were removed for exceptionally long (e.g., 56 min to complete the IAT) or short (e.g., < 7 min to complete entire experiment) times. The final sample included 215 participants (61% female; 85.6% White, 7% Asian, 7% Hispanic, and 1% "other").

5.1.4. Procedure

Participants were told that the experiment investigated how people categorize places, objects, people, and social information. Participants first completed a filler (facial expression categorization) task. The experimental manipulation was introduced next exactly as in Experiment 2 (with the exception that we removed one *should* item from the discrepancy scale concerning laughing at stereotypic jokes), and was followed by the IAT. Participants then completed a second filler task (generation of objects to fit specified categories), the joke evaluation task, and finally the IMS and EMS.

For the joke evaluation task, participants viewed 45 jokes one-at-a-time and selected a "Boo!" response or from one HA! to seven HA!s. Three of the jokes played on stereotypes about Blacks (e.g., "What can a pizza do that a Black man can't do? Feed a family of four"). In addition, two other ethnic stereotype jokes (one about Chinese people: "How do they name Chinese babies? They throw silverware down the stairs until they hear something they like," and one about Mexican people: "What is the difference between a Mexican and a book? A book has papers.") were also included.

All filler and stereotypic jokes were culled from popular websites. Jokes were pilot tested to ensure that the stereotypic jokes yielded sufficiently favorable and variable ratings for use in the present research. Twenty-two participants rated the jokes on a 1 (Boo!) to 8 scale (seven HA!s), and the average of ratings of the stereotypic jokes were analyzed with a one-sample t-test with 1 (Boo!) as the test value. The average stereotypic joke ratings ($M = 2.76$, $SD = 1.48$) differed significantly from 1, $t(21) = 5.55$, $p < 0.001$, and yielded acceptable variability ($SD = 1.52$). Examination of each stereotypic joke individually also supported use in the present research.

5.1.5. Formation of indexes

5.1.5.1. IMS and EMS.

Ratings for IMS items were averaged to form an index ($M = 5.71$, $SD = 1.38$, $\alpha = 0.89$) as were ratings for EMS items ($M = 3.37$, $SD = 1.61$, $\alpha = 0.86$). Neither IMS nor EMS varied systematically with experimental condition, IMS $F(2, 212) = 0.08$, $p = 0.92$, $\eta_p{}^2 = 0.001$; EMS $F(2, 212) = 2.31$, $p = 0.10$, $\eta_p{}^2 = 0.02$.

5.1.5.2. IAT.

IAT scores were computed as in Experiments 1 and 2. Data from 13 participants were missing due to careless responding (latencies < 300 ms on > 10% of their trials).

5.1.5.3. Joke evaluations.

We scored the joke evaluations in two ways. First, we computed a continuous measure ranging from 1 (i.e., selection of "Boo!") to 8 (i.e., selection of "HA!HA!HA!HA!HA!HA!HA!"), averaging across the three jokes about Blacks ($M = 2.63$, $SD = 2.04$, $\alpha = 0.86$). Although participants' average ratings ranged from 1 to 8, note that 80 participants (37% of the sample) selected "Boo!" for all three jokes about Blacks, so the continuous representation of the data was somewhat skewed (skewness = 1.22). Also, note that a "Boo!" response could be considered categorically different from a "HA!" response, so that "Boo!" may not necessarily psychologically represent one scale point lower than "HA!" Given these considerations, our second approach to scoring joke evaluations involved counting and summing the number of "Boo!" responses across the three jokes about Blacks. This yielded a considerably less skewed distribution (skewness = − 0.14) with participants' responses ranging from 0 to 3 ($M = 1.61$, $SD = 1.26$; $\alpha = 0.79$).

5.1.5.4. Discrepancy scores and affect.

Discrepancy scores were computed as in Experiment 2 ($M = 0.81$, $SD = 0.76$, $\alpha = 0.65$), as was the Negself index ($M = 2.14$, $SD = 1.09$, $\alpha = 0.92$).

5.2. Results and discussion

Interrelations among all measures are shown in Table 6.

5.2.1. Discrepancy scores and negative self-directed affect

Regression analyses as described in Experiment 2 were performed to examine the affective consequences of prejudice-related discrepancies among participants in the discrepancy salience condition.

Participants higher in EMS reported greater Negself, $t(72) = 3.43$, $B = 0.24$, $SE = 0.07$, $\beta = 0.35$, $p = 0.001$. Of greater importance, the anticipated main effect of discrepancy when predicting Negself was significant, $t(72) = 2.95$, $B = 0.44$, $SE = 0.15$, $\beta = 0.31$, $p = 0.004$, indicating that Negself increased as participants' proneness to discrepancies increased. The interaction between discrepancy and IMS was not significant, $t(69) = 1.67$, $B = 0.20$, $SE = 0.12$, $\beta = 0.18$, $p = 0.10$. Nonetheless, the pattern of the interaction replicated Experiment 2, such that the effect of discrepancy was not significant at low levels of IMS scores, $b = 0.10$, $se = 0.25$, $p = 0.68$, but was significant at high levels of IMS, $b = 0.65$, $se = 0.19$, $p = 0.001$.

5.2.2. IAT performance

IAT scores were predicted using the same hierarchical regression approach as in Experiments 1 and 2. Participants higher in IMS had lower IAT scores, $t(199) = 2.93$, $B = -0.05$, $SE = 0.02$, $\beta = -0.20$, $p < 0.001$. The set of dummy codes capturing experimental condition was also significant, $\Delta R^2 = 0.13$, $F(2, 197) = 15.86$, $p < 0.001$. As show in Table 7, participants in the counterstereotyping condition had significantly lower IAT scores than participants in the no training condition, $t(197) = 4.60$, $B = -0.27$, $SE = 0.06$, $\beta = -0.35$, $p < 0.001$. Unlike Experiment 2, but consistent with expectations and the pilot study results summarized earlier, IAT scores in the discrepancy salience condition were comparable to the no training condition, $t(197) = 0.42$, $B = 0.02$, $SE = 0.06$, $\beta = 0.03$, $p = 0.68$, and significantly greater than in the counterstereotyping condition, $t(197) = 5.19$, $B = 0.30$, $SE = 0.06$, $\beta = 0.35$, $p < 0.001$.

Table 6
Correlations among measures, Experiment 3.

	Full sample (n = 215)			
	IMS	EMS	IAT	Number of "Boo!" ratings
IMS	–			
EMS	0.01	–		
IAT	− 0.20**	0.03	–	
Number of "Boo!" ratings	0.48**	− 0.19**	− 0.15*	–
Continuous joke ratings	− 0.58**	0.11	0.23**	− 0.82**

	Discrepancy salience condition (n = 76)					
	IMS	EMS	IAT	Number of "Boo!" ratings	Continuous joke ratings	Discrepancy total
Discrepancy total	− 0.21†	0.20	0.15	− 0.10	0.16	
Negative self-directed affect	− 0.13	0.41**	0.06	− 0.06	0.04	0.39**

† $p < 0.07$.
* $p < 0.05$.
** $p < 0.01$.

Replicating Experiments 1 and 2, the interaction effects were not significant, *ps* > 0.49, thus indicating that the effect or lack thereof of experimental condition was not moderated by consciously held motivations.

5.2.3. Racial joke evaluations

5.2.3.1. Continuous 1–8 ratings. We first analyzed the continuous representation of racial joke ratings ranging from 1 to 8 in our hierarchical regression model. We controlled for participants ratings of the neutral jokes, which was a significant covariate, $t(209) = 7.18$, $B = 0.61$, $SE = 0.09$, $\beta = 0.36$, $p < 0.001$. Participants' IMS scores were strongly, negatively related to racial joke evaluations, $t(209) = 11.49$, $B = -0.84$, $SE = 0.07$, $\beta = -0.57$, $p < 0.001$. The dummy codes representing experimental condition added a marginally significant increment to the model, $\Delta R^2 = 0.01$, $F(2, 207) = 2.53$, $p = 0.08$. As shown in Table 7, participants in the discrepancy salience condition evaluated the racial jokes significantly less favorably than participants in the no training condition, $t(207) = 2.25$, $B = -0.55$, $SE = 0.25$, $\beta = -0.13$, $p = 0.03$. In contrast, participants in the no training and counterstereotyping conditions provided similar racial joke evaluations, $t(207) = 1.23$, $B = -0.31$, $SE = 0.26$, $\beta = -0.07$, $p = 0.22$. In sum, participants who completed the discrepancy and affect measures, as a whole, provided less favorable joke evaluations than other participants.

Next we tested our more specific hypothesis that feelings of negative self-directed affect would be related to greater self-regulation in relation to the jokes. We used Negself, IMS, EMS, and neutral joke ratings to predict racial joke evaluations. As in the analysis reported above, IMS and the neutral joke ratings significantly predicted continuous joke ratings (*ps* < 0.001). Although in the expected direction, the effect of Negself was not significant, $t(71) = 1.49$, $B = -0.20$, $SE = 0.13$, $\beta = -0.12$, $p = 0.14$.

5.2.3.2. Number of "Boo!" Ratings. The number of "Boo!" responses to the racial jokes was predicted in our hierarchical regression model. We controlled for the number of "Boo!" responses to neutral jokes, which was a significant covariate, $t(209) = 3.87$, $B = 0.04$, $SE = 0.01$, $\beta = 0.23$, $p < 0.001$. We found that IMS showed a strong, positive relation with joke evaluations, $t(209) = 8.76$, $B = 0.46$, $SE = 0.05$, $\beta = 0.50$, $p < 0.001$, and EMS showed a significant negative relation, $t(209) = 2.73$, $B = -0.12$, $SE = 0.05$, $\beta = -0.16$, $p = 0.01$. Surprisingly, the dummy codes representing experimental condition added a significant increment to R^2, $\Delta R^2 = 0.02$, $F(2, 207) = 3.13$, $p = 0.05$. As shown in Table 7, compared to the no training condition, participants in both the discrepancy salience condition, $t(207) = 2.31$, $B = 0.40$, $SE = 0.17$, $\beta = 0.15$, $p = 0.02$, and the counterestereotyping condition, $t(207) = 2.04$, $B = 0.37$, $SE = 0.18$, $\beta = 0.14$, $p = 0.04$, provided significantly more Boo!s. Furthermore, the discrepancy salience and counterstereotyping conditions did not differ significantly from each other, $t(207) = 0.21$, $B = 0.04$, $SE = 0.17$, $\beta = 0.01$, $p = 0.84$. These unanticipated condition effects may have arisen given the nature of the stereotype application task in this experiment, which likely afforded greater cognitive control over responses than the stereotypic inference task used in Experiments 1 and 2. In other words, participants in Experiment 3's experimental conditions, having completed race-relevant tasks earlier in the study, may have deliberately adjusted their evaluations of the racial jokes to be less positive.

To test our more specific hypotheses relevant to the cognitive retraining and motivated self-regulation accounts, we proceeded to perform two other types of analyses. First, given that participants in the counterstereotyping condition provided more Boo!s than participants in the no training condition, it was important to test whether this effect was driven by reduced stereotype accessibility (i.e., IAT performance).

Table 7
IAT and racial joke ratings as a function of experimental condition, Experiment 3.

	IAT			Continuous joke ratings			Number of Boo! ratings		
	N	Mean	*SD*	N	Mean	*SD*	N	Mean	*SD*
No training	65	0.42	0.36	67	2.88	2.16	67	1.40	1.28
Discrepancy salience	74	0.44	0.30	76	2.30*	1.74	76	1.79*	1.17
Counterstereotype training	63	0.15*	0.36	70	2.73	2.20	70	1.66*	1.31

* Cell means differ from No Training condition at $p < 0.05$.

Using Hayes's (2013) PROCESS (Model 4) analysis with 5000 bootstraps, we predicted the number of Boo!s to the racial jokes with the contrast between the counterstereotyping and control conditions as the independent variable, IAT scores as the mediator, and IMS, EMS, and Boo!s to neutral jokes as covariates. This analysis did not provide support for mediation, 95% CI [-0.11, 0.12], indicating that rejection of the racial jokes was not driven by reduced accessibility of stereotypic associations. These results mirror Dasgupta and Rivera's (2008) findings that priming admired outgroup members reduced implicit bias and also biased voting intentions (i.e., stereotype application); however, the effect of priming on voting intentions was not mediated by implicit bias scores. Like Dasgupta and Rivera (2008), we conclude that our results do not support the notion that counterstereotyping affects stereotype accessibility so as to produce a reduction of stereotype application.

Second, to test the motivated self-regulation account in the discrepancy salience condition, we used Negself (along with IMS, EMS, and neutral joke Boo!s) to predict the number of Boo!s to the racial jokes. IMS, EMS, and neutral jokes were significant predictors ($ps < 0.001$). More importantly, the greater participants' experience of Negself, the greater their number of Boo!s, $t(71) = 2.27$, $B = 0.22$, $SE = 0.10$, $\beta = 0.21$, $p = 0.03$. This finding supports the hypothesis that heightened motivation to self-regulate, as manifested in the experience of negative self-directed affect induced by prejudice-related discrepancies, was related to more egalitarian responding.[2]

5.2.4. Auxiliary analyses: jokes about Chinese and Mexican people

Two other jokes played on stereotypes of other ethnic groups (stereotypes of Mexican and Chinese people), and they were analyzed using the same methods as the racial jokes to examine generalization. Using the 1–8 continuous joke evaluations as the dependent variable, IMS was positively related to these other ethnic group joke evaluations, $t(209) = 6.74$, $B = -0.57$, $SE = 0.09$, $\beta = -0.38$, $p < 0.001$, and the neutral jokes ratings were a significant covariate, $t(209) = 7.29$, $B = 0.72$, $SE = 0.10$, $\beta = 0.41$, $p < 0.001$. More importantly, the dummy codes capturing experimental condition produced a significant increment in R^2, $\Delta R^2 = 0.02$, $F(2, 207) = 3.30$, $p = 0.04$. Participants in the discrepancy salience condition evaluated the racial jokes significantly less favorably ($M = 2.62$, $SD = 1.95$) than participants in the no training condition ($M = 3.21$, $SD = 2.35$), $t(207) = 1.98$, $B = -0.57$, $SE = 0.29$, $\beta = -0.13$, $p = 0.05$. Joke evaluations were also significantly less favorable in the counterstereotyping condition ($M = 2.67$, $SD = 1.97$) than in the no training condition, $t(207) = 2.43$, $B = -0.71$, $SE = 0.30$, $\beta = -0.16$, $p = 0.02$. Next, we followed up with analyses mirroring those with the racial jokes to better understand the processes at work. First, we did not find support for the cognitive retraining account for counterstereotyping, as the effect of counterstereotyping vs. no training on joke evaluations was not mediated by IAT scores, 95% CI[-0.49, 0.04]. Second, we found support for the motivated self-regulation account in the discrepancy salience condition; that is, Negself predicted joke evaluations, $t(71) = 1.96$, $B = -0.36$, $SE = 0.18$, $\beta = -0.20$, $p = 0.05$.

Turning to the number of Boo! responses, the main effect for experimental condition was not significant, $\Delta R^2 = 0.02$, $F(2, 207) = 2.46$, $p = 0.09$, with comparable ratings across the no training ($M = 0.80$, $SD = 0.82$), discrepancy salience ($M = 1.01$, $SD = 0.90$), and counterstereotyping ($M = 1.01$, $SD = 0.87$) conditions. Testing our more specific motivated self-regulation prediction, we found that participants provided more Boo! responses as their Negself increased, $t(71) = 2.60$, $B = 0.23$, $SE = 0.09$, $\beta = 0.27$, $p = 0.01$.

In sum, analyses of the racial joke data supported a motivational account of avoiding stereotype application, particularly in the case of rejecting the jokes playing on stereotypes of Blacks by providing Boo! responses. To the extent that participants' motivation to self-regulate biases had been heightened, as indexed by their discrepancy-induced negative self-directed affect, they evaluated the jokes less favorably. Furthermore, this process generalized to evaluations of jokes concerning other ethnic outgroups. We also found that participants who practiced counterstereotyping responded to the racial and other ethnic group jokes less favorably than participants in the no training condition, but that this was not due to the reduction in implicit bias resulting from counterstereotyping. Thus, the cognitive retraining account for reduced stereotype application was not supported.

5.2.5. Meta-analyses

To provide an overall summary of the effects of the counterstereotyping and discrepancy-salience conditions on stereotype activation (IATd) and stereotype application (Experiments 1 & 2: stereotypic inferences; Experiment 3: number of Boo! responses controlling for neutral joke responses, reverse scored) across the reported experiments, we conducted meta-analyses using Cumming and Calin-Jageman's (2017) ESCI software. Analyses relevant to the counterstereotyping condition were performed including data from Experiments 1–3, and analyses for the discrepancy salience condition involved data from Experiments 2 and 3.

5.2.5.1. Stereotype activation (IAT performance). The meta-analytic comparison between the counterstereotyping and no training conditions when predicting IAT performance was, as expected, significant, $d = 0.63$, 95% CI [0.45, 0.85]. In contrast, as predicted, the comparison between the discrepancy salience and no training conditions was not reliable, $d = 0.16$, 95% CI [-0.28, 0.60].

5.2.5.2. Stereotype application (stereotypic inferences and joke evaluations). The comparison between counterstereotype and no training conditions did not yield a significant effect size when predicting stereotype application, $d = 0.03$, 95% CI [-0.17, 0.24]. Likewise, participants in the discrepancy salience condition were not, overall, less likely than no training participants to apply stereotypes, $d = 0.13$, 95% CI [-0.24, 0.50].

In addition, we meta-analytically examined the more specific hypotheses stemming from the cognitive retraining and motivated self-regulation accounts. First, we examined the indirect effect of counterstereotype training versus no training on stereotype application through IAT in each study (PROCESS, Model 4, 5000 bootstraps, Hayes, 2013). Following recommendations by Wen and Fan (2015; see also, Preacher & Kelley, 2011) we used fully standardized indirect effects to conduct the meta-analysis. We found that counterstereotyping did not reliably reduce stereotype application through reduced stereotype activation, $\beta = -0.07$, 95% CI [-0.18, 0.04]. Next, we meta-analyzed the partial correlations between Negself and stereotype application (controlling for IMS and EMS in both experiments). We found a reliable effect of Negself on stereotype application, $r = -0.25$, 95% CI [-0.41, -0.09].

In sum, these results are consistent with our expectation that counterstereotyping would reduce stereotype activation but not stereotype application. Neither the meta-analyzed direct effect of counterstereotyping nor the meta-analyzed indirect effect of counterstereotyping through IAT scores significantly predicted stereotype application. In contrast, although the meta-analyzed direct effect of

[2] As with Experiment 2, we also conducted more elaborate moderated mediation analyses to test whether the interaction between IMS and discrepancy scores on stereotype application was mediated by Negself. Using Hayes's (2013) PROCESS macro (Model 8) with 5,000 bootstraps we first predicted the continuous ratings of the racist jokes with discrepancy scores entered as the independent variable, IMS entered as the moderator, Negself as the mediator, and EMS and the continuous ratings of the neutral jokes as covariates. Results did not support mediation for participants high on IMS, 95% CI [-0.34, 0.05] or low on IMS 95% CI [-0.21, 0.04]. Next, we used the same model but predicted the number of Boo!s given to the racial jokes (controlling for number of Boo!s given to the neutral jokes instead of continuous rating). At high levels of IMS, the indirect effect of discrepancies on rejection of racist jokes was significant, 95% CI [0.004, 0.34]. At low levels of IMS, a significant mediating effect was not observed, 95% CI [-0.05, 0.23].

discrepancy salience was not significant for stereotype activation or application, the hypothesized link between negative self-directed affect and stereotype application was significant for people in the discrepancy salience condition.

6. General discussion

Contemporary intergroup bias all too often infiltrates our perceptions, impressions, judgments and actions without our conscious awareness (e.g., Greenwald & Banaji, 1995; Monteith, Woodcock, & Gulker, 2013). Increased awareness of the damaging outcomes of implicit biases spurred the question, can training effectively erase the cognitive footprints of bias? The hope, of course, is that this cognitive retraining can thwart stereotype application. Three experiments reported herein demonstrated that training participants to associate Blacks with motivated rather than unmotivated powerfully altered automatically activated associations; stereotyping on an implicit measure was reduced substantially by practicing counterstereotypes. However, even though participants could more easily associate Blacks with traits such as motivated, achiever, educated, and intelligent following cognitive retraining, this change did not translate into reduced application of stereotypes of Blacks such as criminal, prisoner, homeless, and addict in Experiments 1 and 2. Furthermore, although counterstereotype training was associated with greater rejection of racist jokes in Experiment 3, compared to the no training condition, we did not find that rejection was mediated by the reduction in implicit stereotyping. Note that we also tested whether the effects of counterstereotype training stereotype application might depend on participants' internal or external motivations to respond without bias, and did not find this to be the case in any of the experiments. Altogether, these findings do not support the idea that counterstereotyping can reduce stereotype application by changing implicit biases, although it of course remains possible that an unexamined moderator may exist that would suggest that counterstereotyping is effective for some people.

In contrast, activating the motivation to self-regulate, as indexed by heightened negative self-directed affect, was associated with reduced stereotype application. Specifically, priming people's awareness of their propensity to respond in biased ways toward Blacks that conflicted with their less prejudiced standards for responding stimulated feelings of negative self-directed affect, which predicted participants' subsequent ability to avoid subsequent stereotype application.

6.1. *Limited benefits of counterstereotyping*

Our findings are consistent with Dasgupta and Rivera's (2008) results showing that reduced implicit bias following counterstereotypic exemplar training did not lead to a reduction of biased responding. However, as these researchers pointed out, their measure of biased responding involved a very conscious and deliberate task in which activated stereotypes may have played a minimal role. In contrast, the stereotype application task used in Experiments 1 and 2 was subtle, so much so that even an unequivocal warning not to use stereotypes (Experiment 1), as well as practicing counterstereotyping, did not reduce stereotype application. In Experiment 3, participants who practiced counterstereotypes were more likely to reject racial jokes as unfunny compared to participants in the no training condition; however, this effect was not mediated by reduced bias on the implicit stereotyping measure. Although one should always be cautious when interpreting null effects, our meta-analytic results also did not provide support for the hypothesis that counterestereotype training practicing counterstereotyping reduced stereotype application directly, or that it did so indirectly through a reduction of bias on an implicit stereotyping measure.

Importantly, Kawakami et al. (2005) and Kawakami, Dovidio, et al. (2007) argued that counterstereotyping will reduce stereotype application only with a distracting task between the training task and subsequent stereotyping, which we included in all of our experiments, and yet we did not find any evidence that counterstereotyping drove down stereotype application. Given that Kawakami et al. (2005) and Kawakami, Dovidio, et al. (2007) did not test the mediating role of implicit bias reduction following stereotyping on subsequent stereotype application, we conclude that extant evidence does not support the notion that cognitive retraining through counterstereotyping will reduce stereotype application.

Why does counterstereotyping reduce stereotyping on an implicit measure, but this reduced implicit bias does not lead, in turn, to reduced stereotype application? We posit that stereotypic associations are represented in complex, multidimensional networks (Casper et al., 2010; Gawronski & Bodenhausen, 2006; Kunda & Thagard, 1996; Wittenbrink et al., 2001), and that counterstereotype training affects the strength of associations with certain nonstereotypic traits. However, it leaves intact many negative associations that are often developed early in life (Dunham, Chen, & Banaji, 2013) and culturally reinforced across the lifetime. Future research is needed to determine whether this explanation or other possible explanations best apply to our findings. For instance, counterstereotyping may affect the accessibility of certain traits only very temporarily. Indeed, recent evidence suggests that implicit bias reduction resulting from counterstereotyping (and other short-term implicit bias interventions) does not even last after several hours to several days (Lai et al., 2016).

We do not wish to suggest that counterstereotyping can play no role in the reduction of implicit bias and consequently stereotype application. Certain applications of counterstereotyping in long-term interventions may be more effective than short-term cognitive retraining. For instance, protracted exposure to women leaders (e.g., faculty) has been linked with reduced automatic gender stereotyping (Dasgupta & Asgari, 2004; Stout, Dasgupta, Hunsinger, & McManus, 2011). Prolonged positive contact with outgroup members, which in part enables people to form and reinforce non-stereotypic associations, likewise reduces implicit biases (Rudman, Ashmore, & Gary, 2001; Turner, Hewstone, & Voci, 2007). Perhaps these kinds of real-world contact experiences can weaken a broad web of stereotypic associations, which may in turn result in reduced stereotype application.

Furthermore, other forms of cognitive retraining may well reduce implicit bias and, in turn, stereotypic responses. For instance, Kawakami, Phills, Steele, and Dovidio (2007) trained participants to associate approach behavior with Blacks and avoidance behavior with Whites. This manipulation both reduced bias on an evaluative IAT and improved nonverbal behaviors in an interracial interaction. However, Kawakami, Phills, et al. (2007) did not investigate whether the effect of approach training on nonverbal behaviors was mediated through reduced implicit bias, leaving open the possibility of a direct rather than indirect effect.

6.2. *Motivated self-regulation*

The motivated self-regulation strategy operates by making people aware of their biased responses that stand in conflict with their personal beliefs. The resulting feelings of guilt and disappointment with the self then lead to a cascade of consequences that help people monitor for and regulate potentially biased responses in the future. Consistent with prior research (Amodio et al., 2007; Czopp et al., 2006; Monteith et al., 2002, 2010), the findings reported herein supported the critical relationship between negative self-directed affect and reduced stereotype application. It should be noted, however, that negative self-directed affect was measured rather than manipulated, preventing strong causal conclusions.

Moreover, the present work extends prior research in important ways. First, previous research has tested the self-regulation of prejudice theory by experimentally manipulating feedback to give participants the impression that they had engaged in stereotypically biased responses (Amodio et al., 2008; Monteith, 1993; Monteith et al., 2002). In contrast, the present research indicates that raising people's awareness of their everyday proneness to discrepant responses through completion of the Should-Would Discrepancy Questionnaire can elicit negative self-directed affect, which in turn is associated with less biased responding. This strategy of raising awareness of discrepancies may be useful in interventions designed to combat the negative consequences of implicit bias, and it carries the benefits of being inexpensive, noninvasive, and easily deployable.

Second, previous research with the Should-Would Discrepancy Questionnaire has focused on college samples, with the exception of one study that recruited participants from a local airport and laundromat (Voils, Ashburn-Nardo, & Monteith, 2002). This is problematic, because college often provides a liberalizing experience that may encourage people to question whether their prejudice-related biases are inconsistent with egalitarian precepts. The current findings provide the first evidence that prejudice-related discrepancies and the associated negative self-directed affect are experienced in much more diverse samples, and furthermore that discrepancy-associated affect is associated with less biased responding in such samples.

However, the efficacy of the motivated self-regulation strategy depends on having sufficient motivation to self-regulate in the first place. Simply completing the Should-Would Discrepancy Questionnaire was not, overall, followed by a reduction of stereotype application, as evidenced in our meta-analytic results. Rather, reduced stereotype application was restricted to participants who experienced more guilt following increased discrepancy awareness, and these individuals typically are internally motivated to respond in non-biased ways. Understanding how people can be encouraged to be internally motivated to respond without bias is an understudied topic; precisely how this motivation develops remains unclear, aside from recent findings that feeling accepted by outgroups can play a role (Kunstman, Plant, Zielaskowski, & LaCrosse, 2013). This issue clearly deserves future attention.

7. Conclusion

The idea that people should be "retrained" to combat implicit biases and their negative consequences has been increasingly advocated. Indeed, presidential candidate Hillary Clinton called for retraining to address discrimination based on implicit bias during the first 2016 presidential debate (Hunter, 2016). Just what should this (re)training entail? The present research indicates that concentrated cognitive retraining to affect group-based associations in the mind does not, in turn, reduce stereotype application. In contrast, interventions that promote people's awareness of discrepancies between their biased responses and their personal standards for responding can elicit negative self-directed affect, and this affect was related to subsequent self-regulation and avoidance of biased responses. If used in concert with other empirically support strategies in a multi-prong approach to interventions that address both stereotype activation and stereotype application (Devine, Forscher, Austin, & Cox, 2012), we may be in the best position to produce meaningful change.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2017.06.003.

References

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science*, *18*, 524–530.

Amodio, D. M., Master, S. L., Yee, C. M., & Taylor, S. E. (2008). Neurocognitive components of the behavioral inhibition and activation systems: Implications for theories of self-regulation. *Psychophysiology*, *45*, 11–19.

Badger, E. (2016, October 5). *We're all a little biased, even if we don't know it*. The New York Times. Retrieved from http://www.nytimes.com/2016/10/07/upshot/were-all-a-little-biased-even-if-we-dont-know-it.html?_r=0.

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

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology*, *41*, 131–136.

Corell, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*, 359–378.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge Press.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*, 784–803.

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

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

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

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, *60*, 817–830.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.

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

Ford, T. E., & Ferguson, M. A. (2004). Social consequences of disparagement humor: A prejudiced norm theory. *Personality and Social Psychology Review*, *8*, 79–94.

Ford, T. E., Richardson, K., & Petit, W. E. (2015). Disparagement humor and prejudice: Contemporary theory and research. *Humor*, *28*, 171–186.

Forscher, P. S., Lai, C. K., Forscher, P. S., Axt, J., Ebersole, C. R., Herman, M., & Nosek, B. A. (2016). *A meta-analysis of change in implicit bias*. (Manuscript under review).

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

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.

Gray, J. A. (1987). Perspectives on anxiety and impulsivity: A commentary. *Journal of Research in Psychology*, *21*, 493–509.

Gray, J. A., & McNaughton, N. J. (2000). *The neuropsychology of anxiety*. Oxford: Oxford Medical Publications.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

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*, *75*, 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology*, *23*, 253–283.

Hunter, M. (2016, September 27). Clinton: 'I think implicit bias is a problem for everyone.'. Retrieved from http://www.cnsnews.com/news/article/melanie-hunter/clinton-i-think-imlpicit-bias-problem-everyone.

Kaleem, J. (2016). *Why the Department of Justice wants to force its 28,000 employees to confront unconscious racial biases*. Los Angeles Times. Retrieved from http://www.latimes.com/nation/la-na-doj-implicit-bias-20160627-snap-story.html.

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.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*, 68–75.

Kawakami, K., Dovidio, J. F., & Van Kamp, S. (2007). The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group Processes & Intergroup Relations*, *10*, 139–156.

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*, 957–971.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*, 522–544.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–308.

Kunstman, J. W., Plant, E., Zielaskowski, K., & LaCrosse, J. (2013). Feeling in with the outgroup: Outgroup acceptance and the internalization of the motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *105*, 443–457.

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

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., ... Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016.

Lane, S. P., & Hennes, E. P. (2017). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships* (in press).

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.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*, 1029–1050.

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, *64*, 198–210.

Monteith, M. J., Lybarger, J. E., & Woodcock, A. (2009). Schooling the cognitive monster: The role of motivation in the regulation and control of prejudice. *Social and Personality Psychology Compass*, *3*, 211–226.

Monteith, M. J., Mark, A. Y., & Ashburn-Nardo, L. (2010). The self-regulation of prejudice: Toward understanding its lived character. *Group Processes & Intergroup Relations*, *13*, 183–200.

Monteith, M. J., Spicer, C. V., & Tooman, G. D. (1998). Consequences of stereotype suppression: Stereotypes on and not on the rebound. *Journal of Experimental Social Psychology*, *34*, 355–377.

Monteith, M. J., & Voils, C. I. (1998). Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of Personality and Social Psychology*, *75*, 901–916.

Monteith, M. J., Woodcock, A., & Gulker, J. E. (2013). Automaticity and control in stereotyping and prejudice: The revolutionary role of social cognition across three decades. In D. Carlston (Ed.), *Oxford handbook of social cognition* (pp. 74–94). New York: Oxford University Press.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *1*, 1–53.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*, 384–396.

Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., & Markova, T. (2010). Aversive racism and medical interactions with Black patients: A field study. *Journal of Experimental Social Pscyhology*, *46*, 436–440.

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

Plant, E. A., & Devine, P. G. (2001). Response to other-imposed pro-Black pressure: Acceptance or backlash? *Journal of Experimental Social Psychology*, *37*, 486–501.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115.

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

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

Turner, R. N., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The medicating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology*, *93*, 369–388.

Voils, C. I., Ashburn-Nardo, L., & Monteith, M. J. (2002). Evidence of prejudice-related conflict and associated affect beyond the college setting. *Group Processes & Intergroup Relations*, *5*, 19–33.

Wen, Z., & Fan, X. (2015). Monotonicity of effect sizes: Questioning kappa-squared as mediation effect size measure. *Psychological Methods*, *20*, 193–203.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*, 815–827.

Woodcock, A., & Monteith, M. J. (2013). Forging links with the self to combat implicit bias. *Group Processes & Intergroup Relations*, *16*, 44–461.