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
Available online 03 July 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