Available online at www.sciencedirect.com
ScienceDirect
Journal of Experimental Social Psychology 44 (2008) 370–377
Journal of Experimental Social Psychology
www.elsevier.com/locate/jesp

When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation ☆

Bertram Gawronski [a,*], Roland Deutsch [b], Sawsan Mbirkou [a], Beate Seibt [c], Fritz Strack [b]

[a] *Department of Psychology, University of Western Ontario, Social Science Centre, London, Ont., Canada N6A 5C2*
[b] *University of Würzburg, Germany*
[c] *Utrecht University, The Netherlands*

Received 7 February 2006; revised 21 November 2006
Available online 30 December 2006
Communicated by John Skowronski

Abstract

Previous research has shown that extended training in non-stereotypic responding (i.e., negating stereotypes and affirming counterstereotypes) can reduce automatic stereotype activation. In the present research, we claim that the effects of non-stereotypic association training on automatic stereotype activation are primarily driven by the affirmation of counterstereotypes rather than by the negation of stereotypes. In two experiments, participants received extensive training in either (a) negating stereotype-congruent information or (b) affirming stereotype-incongruent information, and then completed a measure of automatic stereotyping (Experiment 1) or automatic evaluation (Experiment 2). Consistent with our predictions, only training in the affirmation of counterstereotypes led to a reduction in the activation of stereotypes and negative evaluations. In contrast, extended training in the negation of stereotypes enhanced rather than reduced the activation of stereotypes and negative evaluations. Implications for prejudice and stereotype control are discussed.

Keywords: Automaticity; Cognitive control; Ironic effects; Negation; Prejudice; Stereotyping

With the discovery that stereotypes can be activated automatically (Devine, 1989), researchers became naturally interested in means to reduce automatic stereotype activation. Although people may be generally able to prevent unwanted influences of stereotypes on overt behavior, such correctional efforts depend on a number of resource-demanding cognitive processes, which can be easily undermined (Muraven & Baumeister, 2000; Strack & Deutsch, 2004; Wegner, 1994). Hence, a more effective means to change unwanted stereotyping is to combat the automatic activation of stereotypes in the first place rather than to deliberately control their influence on behavior once they are activated (Bodenhausen & Macrae, 1998). Up to now, a number of studies have shown that automatic stereotype activation is not an invincible "cognitive monster" (cf. Bargh, 1999). Rather, it seems that automatic stereotype activation can be successfully overcome or at least significantly reduced (for a review, see Blair, 2002).

One particularly interesting means to reduce automatic stereotype activation is non-stereotypic association training. In a series of studies, Kawakami, Dovidio, Moll, Hermsen, and Russin (2000) found that extended practice in non-stereotypic responding is capable of reducing the subsequent activation of stereotypes. In one of their studies, participants were presented with pictures of Black and White individuals and traits that were related either to the stereotype of Blacks or to the stereotype of Whites.

☆ The present research was supported by Canada Research Chairs (CRC) Grant 202555, Social Sciences and Humanities Research Council of Canada (SSHRC) Grant 410-2005-1339, and Academic Development Fund (ADF) Grant 05-303 to Bertram Gawronski; and German Science Foundation (DFG) Grant DE 1150/2-1 to Roland Deutsch. Portions of this article were presented at the 7th Annual Meeting of the Society of Personality and Social Psychology, Palm Springs, CA (January 2006). We are grateful to thank Xenia Avvakumova, William Dunlop, Sara Hart, and Arjun Sharma for their help in collecting the data.
* Corresponding author.
E-mail address: bgawrons@uwo.ca (B. Gawronski).

0022-1031/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2006.12.004

Participants' task was to respond with a *NO* key each time they saw a stereotype-congruent person–trait combination (e.g., a Black face with a stereotypically Black trait word) and to respond with a *YES* key each time they saw a stereotype-incongruent person–trait combination (e.g., a Black face with a stereotypically White trait word). After the training, participants completed a primed Stroop task (Kawakami, Dion, & Dovidio, 1999) designed to assess automatic stereotype activation. Results showed a significant effect of the training, such that automatic stereotype activation was considerably lower after than before the training. This reduction was still present 24 h after the training, providing clear evidence for long-term effects of the training task.

Even though this finding has been replicated in multiple studies, the specific mechanism that leads to a reduction in automatic stereotype activation is still not clear. Kawakami et al. (2000) discussed three possible mechanisms that may account for their findings. First, resembling the operation of auto-motives (Bargh, 1997), the training task may have created an automatic goal to respond in an unbiased, non-stereotypic manner. Second, repeatedly responding *NO* to stereotype-congruent information may have weakened the strength of stereotypic associations in memory. Third, participants may have acquired new, counterstereotypic associations in the course of repeatedly responding *YES* to stereotype-incongruent information.

The present research is primarily concerned with the cognitive mechanisms implied by the latter two accounts. Drawing on recent research by Deutsch, Gawronski, and Strack (2006), we argue that the effects of Kawakami et al.'s (2000) training task are primarily driven by the affirmation of counterstereotypes rather than by the negation of stereotypes. Deutsch et al. (2006) hypothesized that the general procedure of negating the meaning of a stimulus is a propositional, rule-based process that cannot be automatized (see Gilbert, 1991; Strack & Deutsch, 2004). In a series of studies, participants trained to give summary evaluations of positive or negative words that were either affirmed (e.g., *a friend*; *a cockroach*) or negated (e.g., *no friend*; *no cockroach*). Participants' task was to indicate the valence of the compound terms as quickly as possible. Even though participants' responses became faster over the course of the training task, the difference between responses to affirmed and negated words remained constant with responses to negated words being approximately 100 ms slower than responses to affirmed words. This difference was reduced only under conditions that facilitated the storage of the outcome of a particular negation in memory (e.g., a positive evaluation of the stimulus *no cockroach*). However, this effect did not generalize to other negated stimuli. Moreover, negations did not reverse automatic evaluative responses to a particular stimulus, unless the negation was included in the associative representation of that stimulus. These results suggest that the negation training did not increase participants' general ability to negate the meaning of evaluative stimuli (*procedural learning*; see Anderson, 1993). Instead, negation training seems to be effective only when it changes the underlying associative representation of well-learned stimuli, such that the overall meaning of the negated stimulus is stored as an independent unit in associative memory (*instance learning*; see Logan, 1988).

These results have important implications for training effects on automatic stereotype activation. Specifically, Deutsch et al.'s (2006) findings suggest that the negation component in Kawakami et al.'s (2000) training task (i.e., repeatedly saying *NO* to stereotype-congruent information) is rather unlikely to enhance participants' general efficacy in inhibiting automatically activated stereotypes through procedural learning. The only way in which the negation component could reduce automatic stereotyping is by changing the underlying associative representation of the stereotyped group through instance learning. However, such instance learning effects resulting from negation training should occur only when a negated stimulus can be associated with the *outcome* of applying the negation (see Kaup, 2001; Mayo, Schul, & Burnstein, 2004). For example, in Deutsch et al.'s (2006) studies, the negation training required participants to process the reversed meaning of the negated stimulus (e.g., to infer that *no cockroach* is positive). Under such learning conditions, the outcome of the negation (i.e., positive) is repeatedly paired with the negated stimulus (i.e., *no cockroach*), which may ultimately result in an association between the two elements. This process seems quite different from the cognitive operations in Kawakami et al.'s (2000) training task, which simply required participants to respond *NO* to stereotype-congruent information; a processing of the outcome of the negation (i.e., reversing the stereotype into the counterstereotype) was not necessary. Therefore, a change in the associative representation, such that the meaning of the negated stereotype (i.e., the counterstereotype) is stored as an independent unit in memory, seems rather unlikely. To the contrary, enhanced attention to stereotype-congruent information without replacement by the counterstereotype—such as implied by a simple *NO* response—may enhance rather than reduce the activation of stereotypic associations in memory (Wegner, 1994). Thus, if the negated meaning is not activated in memory, negating stereotypes may actually lead to ironic or rebound effects (e.g., Macrae, Bodenhausen, Milne, & Jetten, 1994) rather than to a reduction in automatic stereotype activation. Hence, a more effective means to reduce automatic stereotype activation seems to be the affirmation of counterstereotypes, which inherently implies an activation of counterstereotypical associations in memory. Drawing on these considerations, we argue that only extended training in the affirmation of counterstereotypes, but not training in the negation of stereotypes, should reduce automatic stereotype activation.

In order to test the differential effects of negating stereotypes and affirming counterstereotypes, we conducted two experiments. In both studies, participants were presented with stereotype-congruent and stereotype-incongruent information. Half of the participants were requested to respond *NO* to stereotype-congruent information and to

show no response to stereotype-incongruent information. The remaining half was requested to respond *YES* to stereotype-incongruent information and to show no response to stereotype-congruent information. Immediately after the training task, participants completed a measure of automatic stereotype activation (Experiment 1) or automatic evaluation (Experiment 2). Experiment 1 tested training effects on automatic gender stereotyping; Experiment 2 tested training effects on automatic evaluations of Blacks in comparison to Whites.

Experiment 1

Method

Participants and design

Eighty-two psychology undergraduates at the University of Western Ontario (58 female; 24 male) participated in a study on "attitudes and attention" in return for course credit. The experiment consisted of a 2 (time of measurement: before vs. after training task) ×2 (training task: affirmation vs. negation) mixed-model design with the first variable representing a within-subjects factor and the second a between-subjects factor. Due to a computer malfunction, data from one participant were only partially recorded and thus were excluded from analyses.

Training task

Upon arrival, participants were greeted by the experimenter and then seated in a cubicle in front of a personal computer. In the affirmation training condition, participants received the following instructions on the computer screen:

> "The following task is concerned with the cultural stereotype of men and women. As you probably know, men are often considered as strong whereas women are often considered as weak. This, however, is a cultural stereotype that may or may not be true. In the following task, you will be presented with male and female names. In addition, you will be presented with words relating to strongness and weakness that will appear on the screen shortly after the names. Your task is to respond "YES!" each time you see a combination that is INCONSISTENT with the cultural stereotype of men and women. Specifically, you are asked to respond "YES!" with the space bar each time you see a FEMALE name and a word relating to "STRONGNESS" or a MALE name and a word relating to "WEAKNESS." Please attend particularly to combinations that are INCONSISTENT with the cultural stereotype of men and women! For combinations that are consistent with the cultural stereotype of men and women, you do not have to do anything. Again, please respond "YES!" with the space bar each time you see a combination that is INCONSISTENT with the cultural stereotype of men and women. Please try to respond as quickly as possible!"

Instructions for the negation training condition were identical, the only exception being that participants were asked to respond *NO* to stereotype-congruent combinations and not to respond to stereotype-incongruent combinations. Participants in both conditions were then presented with a total of 200 name–trait pairings. These pairings included 50 combinations of each (a) a female name with a strength-related trait word, (b) a male name with a strength-related trait word, (c) a female name with a weakness-related trait word, and (d) a male name with a weakness-related trait word (see Appendix A). Pairings of male versus female names with strength-related versus weakness-related trait words were randomly created by the computer. For each trial, a name was presented in capital letters at the top of the screen. After 500 ms, a trait word in lower-case letters appeared underneath the name. When participants correctly pressed the space bar in response to a stereotype-congruent combination in the negation condition or to a stereotype-incongruent combination in the affirmation condition, the stimuli disappeared and the next trial started. If participants incorrectly pressed the space bar in response to a stereotype-incongruent combination in the negation condition or to a stereotype-congruent combination in the affirmation condition, the stimuli were replaced by the message "ERROR!" which appeared for 1500 ms in the center of the screen. If participants did not respond to a given combination, the stimuli disappeared after 2500 ms and the next trial started. The inter-trial interval for all responses was 1000 ms. Both training tasks consisted of five blocks of 40 trials each, resulting in a total of 200 training trials. After each block, participants were asked to take a moment to relax, and to press the space bar whenever they felt ready to continue with the task.

Automatic stereotyping

As a measure of automatic stereotype activation, we used a sequential priming paradigm designed to assess automatic associations between the two gender categories and strength versus weakness (see Banaji & Hardin, 1996). Each trial started with a fixation cross ("+") which was presented for 500 ms in the center of the screen. Immediately afterwards, a male or female name from the training task (see Appendix A) was presented as a prime stimulus for 200 ms. The prime stimulus was then replaced by a strength- or weakness-related target word from the training task (see Appendix A), which remained on the screen until participants had responded. Participants were instructed to press the left-hand key ("A") as quickly as possible when they saw a weakness-related word and the right-hand key ("5" of the number pad) when they saw a strength-related word. Prime-target pairs were randomly created by the computer. The task included 40 trials for each of the four prime-target combinations (i.e., male-strong, male-weak, female-strong, female-weak), resulting in a total of 160 trials. Order of trials was randomized individually for each participant. Incorrect responses were indicated with the word "ERROR!" appearing for 1000 ms in the center of the

screen. The inter-trial interval for both correct and incorrect responses was 1000 ms. Participants completed the same sequential priming task twice, once immediately before the training and once immediately after the training.

Results and discussion

Prior to analyses, outliers were excluded by discarding responses lower than 300 ms (1.0% at time 1; 1.6% at time 2) and higher than 1000 ms (9.0% at time 1; 7.7% at time 2). Error trials were excluded from analyses (1.1% at time 1; 1.3% at time 2). Mean response latencies for the different conditions are presented in Table 1. A 2 (prime) × 2 (target) × 2 (training) × 2 (time) mixed-model ANOVA revealed a significant main effect of time, $F(1,79) = 13.11$, $p = .001$, $\eta^2 = .142$, a significant main effect of target, $F(1,79) = 17.62$, $p < .001$, $\eta^2 = .182$, a significant two-way interaction of prime and target, $F(1,79) = 22.27$, $p < .001$, $\eta^2 = .220$, and, most importantly, a significant four-way interaction of prime, target, training, and time, $F(1,79) = 8.39$, $p = .005$, $\eta^2 = .096$.

To specify this interaction in terms of the present hypotheses, we calculated difference scores reflecting automatic gender stereotyping. These indices were calculated by first subtracting the mean response latency to strength-related target words after male primes from the mean response latency to strength-related target words after female primes (i.e., higher scores indicate stronger activation of strength for male as compared to female primes), and by subtracting the mean response latency to weakness-related target words after female primes from the mean response latency to weakness-related target words after male primes (i.e., higher scores indicate stronger activation of weakness for female as compared to male primes). Weakness scores were then added to strength scores, resulting in an index of automatic gender stereotyping with higher scores indicating higher levels of stereotype activation.

Fig. 1 presents the mean values of this index as a function of training task and time of measurement. Consistent with our predictions, affirmation training led to a marginally significant reduction in automatic gender stereotyping, $F(1,39) = 3.46$, $p = .07$, $\eta^2 = .081$. In contrast, negation training significantly increased automatic gender stereotyping, $F(1,40) = 4.97$, $p = .03$, $\eta^2 = .111$. Moreover, automatic gender stereotyping did not differ as a function of the training conditions before the training, $F(1,79) = 2.32$, $p = .13$, $\eta^2 = .029$, but showed a significant difference after the training, $F(1,79) = 6.37$, $p = .01$, $\eta^2 = .075$. These results indicate that only training in the affirmation of counterstereotypes reduced automatic stereotype activation. In contrast, training in the negation of stereotypes enhanced rather than reduced automatic stereotype activation.

Table 1
Mean response latencies in milliseconds as a function of prime (Male vs. Female), Target (Weak vs. Strong), Training task (Affirmation of Stereotypes vs. Negation of Counterstereotypes), and Time of measurement (Before Training vs. After Training), Experiment 1

	Before training		After training	
	Male prime	Female prime	Male prime	Female prime
Affirmation training				
Weak target	624	617	599	599
Strong target	604	620	585	592
Negation training				
Weak target	635	632	628	610
Strong target	617	622	605	617

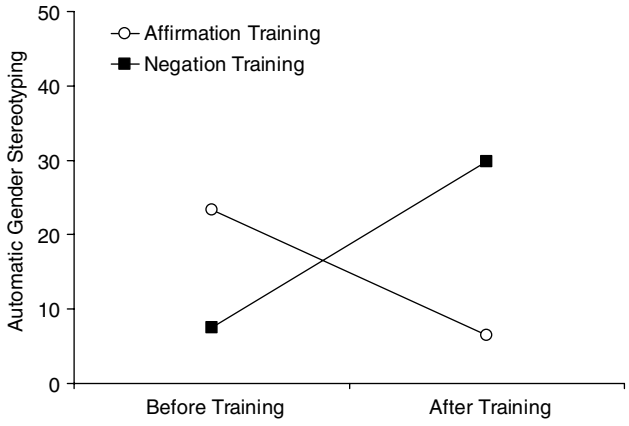

Fig. 1. Mean scores of automatic gender stereotyping as a function of training task (affirmation of counterstereotypes vs. negation of stereotypes) and time of measurement (before vs. after training), Experiment 1.

Experiment 2

The results of Experiment 1 support our assumption that negation of stereotypes and affirmation of counterstereotypes are differentially effective in reducing automatic stereotype activation. In Experiment 2, we tested whether these effects generalize to automatic evaluations of stereotyped groups. Gawronski and Bodenhausen (2006) recently argued that automatic evaluations of a given stimulus depend on the net valence of all (semantic) associations that are activated in response to that stimulus. From this perspective, a change in stereotypic associations should lead to corresponding changes in automatic evaluations, if the former implies a change in the net valence of automatically activated associations. Based on these assumptions, the main goal of Experiment 2 was to test whether the effects of affirmation versus negation training generalize to automatic evaluations, such that repeated affirmation of positive counterstereotypes reduces automatic prejudice whereas repeated negation of negative stereotypes enhances automatic prejudice.

Method

Participants and design

Eighty-three psychology undergraduates at the University of Western Ontario (58 female; 25 male) participated in a study on "attitudes and attention" in return for course

credit. The experiment consisted of a 2 (time of measurement: before vs. after training task) ×2 (training task: affirmation vs. negation) mixed-model design with the first variable representing a within-subjects factor and the second a between-subjects factor. Due to a computer malfunction, data from two participants were only partially recorded, and thus were excluded from analyses.

Training task

The two variants of the training task were identical to Experiment 1 with a few exceptions. Instead of focusing on gender stereotypes, the training tasks in Experiment 2 were concerned with the stereotype of Blacks versus Whites. Participants were first presented with either a Black or a White face that appeared at the top of the screen. After 500 ms, a positive or negative trait word appeared at the bottom of the screen. Trait words were related either to the negative stereotype of Black people or to the positive stereotype of White people (see Appendix A). Participants in the affirmation training condition were asked to respond *YES* with the space bar each time they saw a face–trait combination that was inconsistent with the cultural stereotype of Blacks and Whites. Participants in the negation training condition were asked to respond *NO* with the space bar each time they saw a face–trait combination that was consistent with the cultural stereotype of Blacks and Whites. The general procedure and the number of trials were identical to Experiment 1.

Automatic evaluation

In order to assess automatic evaluations of Blacks and Whites, we employed a subliminal affective priming task (see Fazio, Jackson, Dunton, & Williams, 1995; Wittenbrink, Judd, & Park, 2001) adapted from Gawronski and Bodenhausen (2005). Each trial started with a fixation cross ("+") which was presented for 1000 ms in the center of the screen. Immediately afterwards, the prime word "black" or "white" was presented for 15 ms, which was followed by a masking stimulus ("XXXXX") for 250 ms. The masking stimulus was then replaced by a positive or negative target word which remained on the screen until participants had responded. Participants were instructed to press the left-hand key ("A") as quickly as possible when they saw a positive word and the right-hand key ("5" of the number pad) when they saw a negative word. Each of the 40 target words was presented twice with each of the two prime words, resulting in a total of 160 trials. In order to maximize independence of automatic prejudice and automatic stereotyping at the measurement level, we used positive and negative nouns as target words rather than the trait words from the training task (see Appendix A). Order of trials was randomized individually for each participant. Incorrect responses were indicated with the word "ERROR!" appearing for 1000 ms in the center of the screen. The inter-trial interval for both correct and incorrect responses was 1000 ms. As with Experiment 1, participants completed the same affective priming task twice, once immediately before the training and once immediately after the training.

Results and discussion

Prior to analyses, outliers were excluded by discarding responses lower than 300 ms (0.2% at time 1; 0.2% at time 2) and higher than 1000 ms (5.2% at time 1; 6.3% at time 2). Error trials were excluded from analyses (3.2% at time 1; 3.1% at time 2). Mean response latencies for the different conditions are presented in Table 2. A 2 (prime) × 2 (target) × 2 (training) × 2 (time) mixed-model ANOVA revealed a significant main effect of target, $F(1,79) = 26.36$, $p < .001$, $\eta^2 = .250$, a significant three-way interaction of prime, target, and training, $F(1,79) = 6.49$, $p = .01$, $\eta^2 = .076$, and, most importantly, a significant four-way interaction of prime, target, training, and time, $F(1,79) = 12.49$, $p = .001$, $\eta^2 = .136$.

To specify this interaction in terms of the present hypotheses, we calculated difference scores reflecting automatic preference for Whites over Blacks. These indices were calculated by first subtracting the mean response latency to positive words after White primes from the mean response latency to positive words after Black primes (i.e., higher scores indicate stronger activation of positivity for White as compared to Black), and by subtracting the mean response latency to negative words after Black priming from the mean response latency to negative words after White priming (i.e., higher scores indicate stronger activation of negativity for Black as compared to White). Negativity scores were then added to positivity scores, resulting in an index of automatic preference for Whites over Blacks with higher scores indicating a stronger preference for Whites over Blacks.

Fig. 2 presents the mean values of this index as a function of training task and time of measurement. Consistent with our predictions, affirmation training led to a significant reduction in automatic preference for Whites over Blacks, $F(1,40) = 6.56$, $p = .01$, $\eta^2 = .141$. In contrast, negation training significantly increased automatic preference for Whites over Blacks, $F(1,39) = 6.01$, $p = .02$, $\eta^2 = .133$. Moreover, automatic preference for Whites over Blacks did not differ as a function of training before the training, $F(1,79) = 0.56$, $p = .45$, $\eta^2 = .007$, but showed a highly significant difference after the training, $F(1,79) = 24.19$, $p < .001$, $\eta^2 = .234$. These results indicate that training in the affirmation of positive counterstereotypes effectively

Table 2
Mean response latencies in milliseconds as a function of prime (White vs. Black), Target (Positive vs. Negative), Training task (Affirmation of Stereotypes vs. Negation of Counterstereotypes), and Time of measurement (Before Training vs. After Training), Experiment 2

	Before training		After training	
	White prime	Black prime	White prime	Black prime
Affirmation training				
Positive target	573	572	587	576
Negative target	587	585	581	590
Negation training				
Positive target	568	564	572	574
Negative target	582	582	593	583

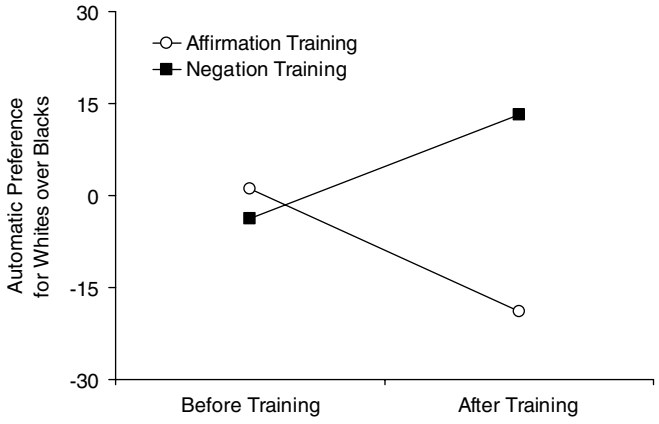

Fig. 2. Mean scores of automatic preference for Whites over Blacks as a function of training task (affirmation of counterstereotypes vs. negation of stereotypes) and time of measurement (before vs. after training), Experiment 2.

reduced automatic negative evaluations. In contrast, training in the negation of negative stereotypes enhanced rather than reduced automatic negative evaluations.

General discussion

According to Kawakami et al. (2000), effects of non-stereotypical association training on automatic stereotype activation could be driven by at least three different mechanisms. First, non-stereotypical association training may create an automatic goal to respond in an unbiased manner. Second, repeatedly responding *NO* to stereotype-congruent information may weaken the strength of stereotypic associations in memory. Third, repeatedly responding *YES* to stereotype-incongruent information may create new, counterstereotypic associations. Addressing the cognitive mechanisms implied by the latter two accounts, the present studies indicate that only training in the affirmation of counterstereotypes effectively reduces automatic stereotype activation. In contrast, extended training in the negation of stereotypes seems to result in ironic effects (Wegner, 1994), such that enhanced attention to stereotypical information—as involved in stereotype negation—strengthens stereotypical associations, thereby leading to an increase in automatic stereotype activation (e.g., Macrae et al., 1994). These effects generalized to automatic evaluations, such that affirmation of positive counterstereotypes reduced automatic negative evaluations whereas negation of negative stereotypes increased automatic negative evaluations.

The latter finding has important implications for controversies regarding the dependency versus independency of automatic stereotyping and automatic prejudice. Amodio and Devine (2006) recently argued that automatic stereotyping and automatic prejudice are generally independent, in that the two have their roots in two distinct memory systems. In support of their view, Amodio and Devine presented three studies showing that (a) automatic stereotyping of African Americans was uncorrelated to automatic evaluations of African Americans, and (b) automatic stereotyping and automatic evaluations uniquely predicted behavioral responses to African Americans in a double dissociation paradigm. However, in evaluating these findings, it seems important to note that Amodio and Devine's stereotyping measure included only two specific dimensions of the stereotype of African Americans (i.e., physical vs. mental). Thus, it remains an open question whether automatic stereotyping would predict automatic evaluations—and behavioral responses predicted by automatic evaluations—if all dimensions of the prevalent stereotype were assessed. An alternative approach to testing the (in)dependency of automatic stereotyping and automatic evaluation—the one which we chose in Experiment 2—is to measure automatic evaluations as a function of experimentally induced changes in automatic stereotyping. If automatic stereotyping and automatic evaluation are independent, changes in automatic stereotyping should leave automatic evaluations unaffected (unless the employed experimental manipulation independently influences both automatic stereotyping and automatic evaluation). This assumption stands in contrast to theorizing by Gawronski and Bodenhausen (2006), who argued that automatic evaluations depend on the particular (semantic) associations that are activated in response to a given stimulus. According to this view, any change in semantic associations—and thus automatic stereotype activation—should lead to corresponding changes in automatic evaluation. Even though we cannot rule out two simultaneous *direct* effects on automatic stereotyping and automatic evaluation, our findings that affirmation and negation training lead to corresponding changes in automatic stereotype activation and automatic evaluation are generally consistent with the claim that the two are directly related. Future research employing mediational designs may help to further clarify the relation between automatic stereotyping and automatic prejudice.

A possible objection in this context is that the obtained changes in automatic evaluation may be driven by evaluative conditioning (for a review, see De Houwer, Thomas, & Baeyens, 2001) rather than by the affirmation versus negation of (counter)stereotypes. In this case, the obtained effects on automatic stereotyping and automatic evaluation could still be independent, as implied by Amodio and Devine's (2006) dual memory account. In response to this objection, it is important to note that the pairings of Black and White faces with positive and negative words were held constant across affirmation and negation conditions, with an equal number of pairings for each of the four trial categories (i.e., Black-positive; Black-negative; White-positive; White-negative). Given that evaluative conditioning effects depend on contiguous pairings of a conditioned stimulus (CS) with either positive or negative unconditioned stimuli (US), evaluative conditioning cannot account for the obtained asymmetry between affirmation and negation conditions. Rather, the present results seem to be driven by participants' attentional focus in responding to the *same* set of stimuli, namely negating stereotypes versus affirming counterstereotypes.

An open question is, however, how the antagonistic impact of affirmation and negation training could still lead to a reduction of automatic stereotype activation in Kawakami et al.'s (2000) studies. One possibility is that procedural differences between the present training paradigm and the one employed by Kawakami et al. (2000) influence the relative effectiveness of negation training in reducing automatic stereotype activation. As outlined above, Kawakami et al.'s (2000) non-stereotypic training task involved both an affirmation and a negation component, such that participants were required to respond *NO* to stereotype-congruent information and to respond *YES* to stereotype-incongruent information. This procedure differs from the one employed in the present studies, in which half of the participants were required to respond *NO* to stereotype-congruent information whereas the remaining half were requested to respond *YES* to stereotype-incongruent information. Thus, it is possible that the combination of affirmation and negation responses in Kawakami et al.'s training task makes participants think of the counterstereotype every time they negate the stereotype. As outlined in the introduction, such activation could enhance counterstereotypical associations in memory, thereby reducing automatic stereotype activation. Note, however, that this explanation implicitly transforms the negation of the stereotype into an affirmation of the counterstereotype, which is consistent with the present conclusion that simply negating a stereotype without simultaneously activating the counterstereotype is ineffective in reducing automatic stereotyping.

The present findings also have important implications for motivated attempts to control stereotypes. Drawing on the proposed distinction between affirmation versus negation foci, different strategies to control unwanted stereotyping may be differentially effective in reducing stereotypical behavior. More precisely, the present findings suggest that thinking about stereotyped groups or individuals in counterstereotypical terms (e.g., "old people are good drivers") is more effective in reducing unwanted stereotyping than attempts to negate an existing stereotype (e.g., "it is not true that old people are bad drivers"). The reason for this difference resides in the different associations that are activated in the course of controlling stereotypes. Whereas the first strategy directly activates counterstereotypical associations, the second strategy activates stereotypical associations that need to be controlled by means of a propositional, rule-based process (Deutsch et al., 2006; Gilbert, 1991). Thus, if people's ability to control these associations is undermined, the first strategy may still lead to non-stereotypical behavior whereas the second strategy may result in ironic stereotyping effects (Wegner, 1994). Drawing on these considerations, it seems advisable to employ stereotype control strategies that imply an affirmation of counterstereotypes rather than strategies that imply a negation of stereotypes.

This difference between affirmation and negation foci is also important for persuasive appeals aimed at reducing prejudice and stereotyping (Gawronski & Bodenhausen, 2006). Grant, Malaviya, and Sternthal (2004) have shown that persuasive arguments can backfire when they include negations. For instance, the persuasive message "diet soft drinks do not promote obesity" may activate the concepts "diet soft drinks" and "obesity," which in turn creates an association between the two. Because processing negations requires a deliberate reversal of the meaning implied by the simultaneous activation of the two concepts, ironic effects are to be expected whenever the message is processed under suboptimal conditions, such as time pressure or insufficient motivation (Deutsch et al., 2006; Gilbert, 1991). Given that recipients of persuasive messages likely adopt the focus in the message as the primary strategy to control unwanted stereotyping, persuasive appeals aimed at reducing prejudice and stereotyping may be well-advised to incorporate an affirmation focus rather than a negation focus. From this perspective, appeals to "just say no" may not be enough—and sometimes even detrimental.

Appendix A. Stimulus material

Male and Female names used in Experiment 1

Female Names: *Angela, Betsy, Peggy, Dianne, Gloria, Janet, Karen, Martha, Rachel, Tanya*

Male Names: *Andrew, Bill, Paul, David, George, Jason, Kevin, Matthew, Richard, Tony*

Strength- and Weakness-related trait words used in Experiment 1

Weakness-Related Words: *dainty*, *delicate*, *weak*, *fragile*, *small*, *tender*, *slight*, *wispy*, *frail*, *feeble*;

Strength-Related Words: *mighty*, *powerful*, *forceful*, *assertive*, *potent*, *tough*, *strong*, *vigorous*, *intense*, *big*

Stereotypical trait words used in the training task in Experiment 2

Trait words related to the negative stereotype of Black people: *poor*, *dishonest*, *complaining*, *violent*, *shiftless*, *superstitious*, *lazy*, *threatening*, *dumb*, *hostile*

Trait words related to the positive stereotype of White people: *intelligent*, *successful*, *ambitious*, *industrious*, *educated*, *responsible*, *wealthy*, *ethical*, *smart*, *friendly*

Positive and negative target words used in the affective priming task in Experiment 2

Positive target words: *paradise*, *summer*, *harmony*, *freedom*, *honesty*, *honor*, *health*, *cheer*, *pleasure*, *heaven*, *friend*, *sunrise*, *love*, *relaxation*, *peace*, *vacation*, *happy*, *lucky*, *miracle*, *gift*

Negative targets words: *evil*, *sickness*, *vomit*, *bomb*, *murder*, *abuse*, *prison*, *death*, *assault*, *cancer*, *rotten*, *accident*, *grief*, *poison*, *stink*, *cockroach*, *virus*, *disaster*, *ugly*, *terror*

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science, 7*, 136–141.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.

Bargh, J. A. (1999). The cognitive monster: the case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York: Guilford Press.

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

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. Wyer Jr. (Ed.), *Advances in social cognition* (Vol. 11, pp. 1–52). Mahwah, NJ: Erlbaum.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853–869.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: negation as reflective operation. *Journal of Personality and Social Psychology, 91*, 385–405.

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

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

Gawronski, B., & Bodenhausen, G. V. (2005). Accessibility effects on implicit social cognition: the role of knowledge activation versus retrieval experiences. *Journal of Personality and Social Psychology, 89*, 672–685.

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.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 107–119.

Grant, S. J., Malaviya, P., & Sternthal, B. (2004). The influence of negation on product evaluations. *Journal of Consumer Research, 31*, 583–591.

Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory and Cognition, 29*, 960–967.

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1999). Implicit stereotyping and prejudice and the primed Stroop task. *Swiss Journal of Psychology, 58*, 241–250.

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 stereotypic activation. *Journal of Personality and Social Psychology, 78*, 871–888.

Logan, G. D. (1988). Toward and instance theory of automatization. *Psychological Review, 95*, 492–527.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*, 808–817.

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

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: does self-control resemble a muscle? *Psychological Bulletin, 126*, 247–259.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220–247.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101*, 34–52.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology, 37*, 244–252.