ATTITUDES AND SOCIAL COGNITION

Putting the Brakes on Prejudice:
On the Development and Operation of Cues for Control

Margo J. Monteith, Leslie Ashburn-Nardo, Corrine I. Voils, and Alexander M. Czopp
University of Kentucky

A model concerning the establishment and operation of cues for control was developed and tested to understand how control can be exerted over (automatic) prejudiced responses. Cues for control are stimuli that are associated with prejudiced responses and the aversive consequences of those responses (e.g., guilt). In Experiments 1 and 2, 3 events critical to the establishment of cues occurred: behavioral inhibition, the experience of guilt, and retrospective reflection. In Experiment 3, the presentation of already-established cues for control did, as expected, produce behavioral inhibition. In Experiment 4, participants were provided with an experience in which cues could be established. Later presentation of those cues in a different task resulted in behavioral inhibition and less racially biased responses.

You are driving home from work, following the same route that you have followed every day for the past 10 years. Because frequently and repeatedly activated constructs become highly accessible and are likely to prevail on judgment and behavior (e.g., Bargh, Bond, Lombardi, & Tota, 1986; Higgins & King, 1981; Higgins, Rhotes, & Jones, 1977; Srull & Wyer, 1979), you have no trouble executing the responses that allow you to reach your doorstep with very little thought. Indeed, your responses to various environmental stimuli, such as a convenience store sign that marks a place to turn right, may be so well learned that you need not devote conscious thought to their execution. Having the ability to develop patterns of responses that are guided by automatic processes certainly is efficient and adaptive. Nevertheless, the advantages afforded by automatic mental processing come with potentially significant drawbacks. When people no longer want their responses to be driven by easily activated constructs, efficient mental processing ceases to be adaptive and can even be highly maladaptive. For example, suppose a new intersection requiring you to stop is introduced to your well-learned driving route. To avoid negative consequences (e.g., collision with another car), you must deautomatize old patterns of responding and replace them with newer, more acceptable ways of responding. As difficult as breaking old habits may be, deautomatization is a common occurrence in a world that is far from static. The present research seeks to increase our understanding of how automatic processes can be controlled and changed in the context of stereotyping. In doing so, this research revisits Allport’s (1954) classic question of how people can put the brakes on their prejudices.

Automatic Stereotyping and Its Controllability

Researchers are in widespread agreement that stereotypes and implicit (unconscious) prejudices can be automatically activated when one encounters a group member and identifies the person’s skin color, facial features, gender characteristics, and the like (e.g., Bargh, 1989, 1994; Brewer, 1988; Devine, 1989; Fiske & Neuberg, 1990; Greenwald, McGhee, & Schwartz, 1998; Macrae, Bodenhausen, & Milne, 1995; Pratto & Bargh, 1991). Once activated, stereotypes are often applied without awareness or conscious intent. This may result in discriminatory responses, such as Blacks being judged as hostile (Devine, 1989; Sagar & Schofield, 1980) and treated with hostility (Bargh, Chen, & Burrows, 1996) and women and men being perceived as dependent and aggressive, respectively (Banaji, Hardin, & Rothman, 1993; Dunning & Sherman, 1997). Individuals’ own reports of their struggles to avoid the use of stereotypes provide further verification of the ease with which stereotypes are activated and applied (Devine, Monteith, Zuwerink, & Elliot, 1991; Monteith, 1996a, 1996b; Monteith, Devine, & Zuwerink, 1993; Monteith & Voils, 1998; Voils, Ashburn-Nardo, & Monteith, 2002; Zuwerink, Monteith, Devine, & Cook, 1996).
A number of stereotype control strategies have been identified (for a review, see Devine & Monteith, 1999). People may try to suppress or banish stereotypic thoughts from their mind (Macrae, Bodenhausen, & Milne, 1998; Macrae, Bodenhausen, Milne, & Jetten, 1994; Monteith, Sherman, & Devine, 1998; Monteith, Spicer, & Tooman, 1998), actively seek out information beyond what is provided at the category level or through stereotype activation (Brewer, 1988; Brewer & Feinstein, 1999; Fiske, Lin, & Neuberg, 1999; Fiske & Neuberg, 1990), override a stereotypic response with a clearly identifiable egalitarian response (Devine, 1989; Monteith, 1993), or correct a response in the direction opposite to the presumed influence of the stereotype (Wegener & Petty, 1997). Although sometimes useful, these strategies often fail to result in the successful control of stereotyping (see Bargh, 1999; Devine & Monteith, 1999). Thus, the questions of whether and how people can learn to gain greater control over the automatic processes that give rise to stereotyping effects become critically important.

Toward Understanding the Mechanisms for Taking the Automatic Out of Prejudiced Responding

Clinical research (Beck, 1976; Beck, Rush, Shaw, & Emery, 1979; Kanfer & Gaillick, 1986) has long focused on understanding people’s ability to control automatically activated constructs (e.g., the automatic activation of negative thoughts among depressed individuals). Social–psychological conceptualizations of self-regulation (e.g., Carver & Scheier’s, 1981, 1990, cybernetic model; Higgins’s, 1997, regulatory focus model) likewise indicate that people bring their behavior in line with their standards, perhaps even if it is automatic processes that underlie the behavior. However, this theory and research do not explain precisely how this can occur, particularly in the context of controlling automatic processes. Exactly how does one regulate the self out of its rut?

Monteith (1993) presented and tested a theoretical account of how people might learn to avoid prejudiced responses. The main idea behind this account is that people learn to control prejudiced responses through self-regulatory outcomes that follow from awareness of failures to control stereotyping. Specifically, consistent with Higgins’s (1987) self-discrepancy theory, one consequence of realizing that one has responded with greater prejudice than one’s personal standards suggest is that negative self-directed affect (hereafter referred to as negself; e.g., guilt) is experienced (Devine et al., 1991; Monteith, 1993; Monteith, 1996b; Monteith et al., 1993). Consistent with Pyszczynski and Greenberg’s (1986, 1987) research, Monteith (1993) argued and found that another consequence of awareness of discrepant responses is that individuals show heightened self-focus, which theoretically should facilitate subsequent efforts at self-regulation. Gray’s (1982) theory of motivation and learning led to the additional prediction that people who had detected a prejudice-related discrepancy should become preoccupied with it and direct their attention toward understanding why it occurred, which was confirmed in Monteith’s (1993) research. The theoretical analysis offered by Monteith was that these self-regulatory outcomes of violating one’s nonprejudiced standards should help individuals to exert control over potentially prejudiced responses in subsequent situations. This reasoning was supported (Monteith, 1993, Experiment 2) in that participants induced to experience a discrepancy in an initial experimental task were, in the context of a subsequent and seemingly unrelated experimental task, more likely to inhibit stereotypic responses.

The Present Analysis: The Importance of Cues for Control

Monteith’s (1993) research showed that people’s initial failures to respond consistently with their personal standards can ultimately have positive effects. However, the research did not provide a complete theoretical account or empirical evidence to explain precisely how automatic processes might be interrupted and eventually changed. Our intentions are to present a theoretical analysis of self-regulation that involves mechanisms called cues for control and to empirically test for the development and operation of cues for control.

Consistent with Monteith’s (1993) self-regulation model, we argue that automatic stereotype activation and use can result in a response that is discrepant from how one believes one ought to respond and that awareness of this discrepant response leads to negself. At this point, we draw from Gray’s (1981, 1982, 1987; Gray & McNaughton, 1996) neuropsychological model of motivation and learning to explain how cues for control can be developed and how their operation can help people to identify situations in which ongoing behavior needs to be interrupted so that unwanted responses can be inhibited and replaced.

Gray (e.g., 1982) has identified motivational systems that have been linked to the operation of different brain structures. Of relevance here is the behavioral inhibition system (BIS), the seat of self-regulation. Gray’s model has been most frequently applied to problems that result when the motivation and learning system (and its associated brain structures) fail to function properly so as to achieve optimal outcomes through self-regulation, such as in cases of psychopathy and disinhibition (Fowles, 1988; Newman, Wallace, Schmitt, & Arnett, 1997; Patterson & Newman, 1993), alcoholism (Pihl, Peterson, & Finn, 1990), affective disorders (Fowles, 1988), and conduct disorder (Walker et al., 1991). In contrast, we apply the model to understand the steps involved in successful self-regulation of prejudiced responses.

The BIS is responsive to various stimuli, and the two of present interest are (a) mismatches between expected and actual stimuli and (b) stimuli that have come to be associated with unwanted responses and the aversive consequences of those responses. More specifically, the BIS initially functions as a comparator, checking actual and expected stimuli. If there is a mismatch, the BIS takes control over behavior and functions in what Gray (1982) called control mode. If we apply this to stereotyping and prejudice (see left part of Figure 1), the BIS

---

1 Monteith’s (1993) analysis included some discussion of cues for control, but not as detailed as the discussion found here, and the empirical work she reported did not focus on the development and operation of cues for control. Following Gray (1982), Monteith (1993) referred to these cues as cues for punishment. We have decided to use the label cues for control instead because we believe that it is more descriptive of the actual function of the cues. That is, the cues ultimately operate to facilitate control over unwanted responses (or responses that could result in negative consequences).
Consider a concrete example. Suppose Pat is grocery shopping and looking for an item. Pat sees a person who happens to be Black down the aisle and automatically assumes that the person works there. (Blacks are often assumed to be “the help”; Feagin, 1991). Pat asks for help finding the item, and the Black fellow shopper explains that she is shopping, not working there. Pat realizes his mistaken assumption, stops for an instant (i.e., behavioral inhibition occurs), and feels bad. Pat notes the situation and engages in retrospective reflection, so that he comes to associate this discrepant response with feeling bad and discrepancy-relevant stimuli. This should result in the development of cues for control. For example, if the grocery store, the item for which Pat was searching, or seeing a Black person in a store becomes linked with the discrepant response and feeling bad, then these stimuli should serve as cues for control in subsequent situations in which prejudiced responses are possible. Throughout the remainder of Pat’s shopping trip, he will of course be highly motivated to avoid prejudiced responses. But how will the cues for control that were developed through this experience help in the control of prejudiced responses in the more distant future? In subsequent situations, the presence of cues for control (e.g., a Black person in a store) should interrupt ongoing behavior (i.e., produce behavioral inhibition) and instigate prospective reflection, which involves thinking twice before acting. Pat should generate responses more carefully, making sure that they will not be discrepant with his standards. For example, in the future Pat should not make the same mistaken assumption about a Black person being “the help” if previously developed cues for control are present in the situation. Presumably, all other things being equal, the greater the similarity is between the environmental stimuli that were tagged in the initial situation and the environmental stimuli present in the subsequent situation, the more effective cues are likely to be.

We conceptualize the establishment of cues for control as essential for gaining control of the automatic processes involved in stereotyping. With the establishment of cues for control, ongoing behavior in which a prejudiced response is possible can be interrupted (i.e., behavioral inhibition should occur), and, through the operation of the BIS, conscious processing resources can be recruited. Equipped with these resources, one can gate stereotypes before they influence responses, and one can generate alternative, more personally acceptable responses.

How the Present Analysis Differs From Other Theory and Research

We have described a neuropsychological account of how people can learn to interrupt and change responses that could otherwise result from automatic processes through the development and operation of cues for control. Although our analysis overlaps some with other self-regulation and stereotype control work, there are critical differences.

First, various psychological models have explained that people can engage in self-regulation to bring their responses in line with their behavior (e.g., Carver & Scheier, 1981, 1990; Duval & Wicklund, 1972; Higgins, 1997). However, these models do not outline the mechanisms by which people do this, which is the focus of our analysis. Second, our analysis has some similarities with Devine’s (1989) dissociation model in that both suggest that low-prejudice people can override prejudiced responses resulting from
automatic processing. The dissociation model set the stage for the present research, but our analysis takes this line of thinking further by explaining precisely how people might be able to learn to control prejudiced responses through self-regulation and, in particular, through the development and operation of cues for control. Also, Devine discussed control over prejudiced responses as being possible only if one has plenty of controlled processing resources already available, whereas the present analysis maintains that cues for control can serve as signals that trigger the recruitment of cognitive resources.

Third, like the present research, Kawakami, Dovidio, Moll, Hermens, and Russin (2000) focused on how people might be able to learn to avoid automatic stereotyping effects. These researchers used a training paradigm in which participants practiced negating stereotypes across hundreds of trials (i.e., participants were to respond “no” when they saw a photograph of a stereotyped group member accompanied by a stereotypic descriptor and “yes” if the photograph was accompanied by a nonstereotypic descriptor). Following this extended practice, participants showed reduced levels of stereotype activation. Although this research provides an important empirical demonstration that stereotype activation can be reduced, the practical implications of the training activity for the everyday control of stereotyping in real-world contexts seem less clear. Our analysis suggests that whether people are able to practice the negation of stereotypes in real-world situations may depend on the establishment of cues for control and the self-regulatory processes described herein.

Fourth, Moskowitz, Gollwitzer, Wasel, and Schaaf’s (1999) work also shares some overlap with the present work, although it, too, is clearly distinct. These researchers proposed that people who hold chronically accessible egalitarian goals can control stereotype activation through the preconscious activation (Bargh & Gollwitzer, 1994) of their goals. Moskowitz et al. first developed a procedure for assessing the extent to which participants held chronically accessible egalitarian goals. Specifically, participants who were first forced to make stereotypic judgments and then engaged in compensatory behavior (i.e., demonstrated very egalitarian responses) were defined as having chronically accessible goals (Study 1). Such individuals were able to avoid automatic stereotype activation (Study 3). Moreover, this appeared to be accomplished through preconscious control of stereotype activation and did not simply result as a byproduct of atrophied stereotypic associations (Study 4). Although this research is similar to our analysis in suggesting that some people can control prejudiced responses that otherwise would result from automatic stereotypic processing, the analyses of how this can be accomplished are very different. Moskowitz et al. (1999) focused on the role of an individual-differences variable (i.e., goal chronicity) in an examination of whether preconsciously activated egalitarian goals can prevent stereotypes from being activated. In contrast, our analysis focuses on the role of discrepancy experiences in instigating a host of self-regulatory outcomes that result in the development of cues for control that can signal the need for exerting control over potentially prejudiced responses.

Summary of Empirical Goals

The four experiments reported herein were designed to examine whether there is evidence for the development and operation of cues for control in the ways suggested by the model shown in Figure 1. We focus on the establishment of cues for control among individuals who hold low-prejudice attitudes and thus are internally motivated to control their prejudiced responses. Of course, high-prejudice people also might self-regulate their responses in similar ways, but only if there are external motivations to do so (e.g., the threat of punishment from others for exhibiting prejudice; Plant & Devine, 1998). Given our current focus, only low-prejudice individuals were recruited for several of the experiments, although people with more prejudiced personal standards were recruited for comparison purposes in one experiment.

Experiments 1 and 2 test whether events that theoretically are necessary for the development of cues for control do indeed occur. These events include, first, behavioral inhibition following the detection of prejudiced responses, which indicates the triggering of inhibitory activity. Second, an aversive outcome in the form of negself must be experienced. Third, retrospective reflection should occur, which facilitates the building of associations among prejudiced responses, aversive outcomes, and stimuli that predict the occurrence of the prejudiced responses. Finally, our analysis also implies that the magnitude of the aversive outcome should be significantly related to the extent of behavioral inhibition and retrospective reflection.

The remaining experiments focus on the actual operation of established cues for control. Experiment 3 tests whether, given the presence of stimuli that can be expected to serve as cues for control, participants show evidence of behavioral inhibition. That is, this experiment tests whether the presence of already established cues results in a brief interruption of ongoing behavior. In Experiment 4, we provided participants with the opportunity to establish cues for control and then examined whether the subsequent presence of the cues resulted in behavioral inhibition. Theoretically, this interruption is important because it allows individuals to engage in prospective reflection so that alternative, less prejudiced responses can be generated. Thus, an additional goal of Experiment 4 is to empirically examine whether less prejudiced responses actually are provided in the presence of cues.

Experiments 1 and 2

Experiments 1 and 2 involve similar procedures but include some different measures. To examine whether events that theoretically should occur if cues for control are being developed actually do occur, Experiments 1 and 2 used a false physiological feedback procedure based loosely on that of Dutton and Lake (1973). White participants were led to believe that the experiment concerned people’s ability to control and reduce negative physiological arousal. Participants were assigned either to a racial feedback condition or to a nonracial feedback condition. Those in the racial feedback condition received feedback that they were unable to control their negative physiological reactions in relation to pictures that included Blacks. Thus, these participants believed they had prejudiced responses to the pictures. Participants in the nonracial feedback condition believed they were unable to control negative physiological arousal in relation to certain nonracial pictures that did not include Blacks. We used this false physiological feedback paradigm because we wanted to ensure that participants did believe they had engaged in prejudiced responses (i.e., we wanted to achieve experimental control). Readers should note, however, that
pilot testing indicated that low-prejudice individuals thought it was probable that they would have responses to the pictures that were more negative than they believed were appropriate.2

Experiment 1 included manipulation checks to ensure that participants interpreted the feedback as intended and believed it. More important, in both studies, a variety of measures were included to examine whether events required for the development of cues occurred. The extent to which participants paused (in milliseconds) when presented with the feedback they received in relation to the pictures was taken as an index of behavioral inhibition. Using time to process feedback that is indicative of a discrepant response (or mismatch) is a standard way of assessing the behavioral inhibition that should occur if cues for control are being established. For example, Newman and his colleagues (e.g., Arnett, Smith, & Newman, 1997; Newman et al., 1997; Nichols & Newman, 1986; Patterson, Kosson, & Newman, 1987; Wallace, Schmitt, Vitale, & Newman, 2000) have operationalized BIS activity in such a way in numerous investigations of self-regulatory deficits among psychopaths versus nonpsychopathic controls. As in past research (e.g., Devine et al., 1991; Monteith, 1993), the measure of aversive consequences that we used was how participants were feeling after receiving their physiological feedback. Finally, we measured retrospective reflection (Experiment 2) using a thought-listing task to determine the extent to which participants were preoccupied with their inability to control their negative arousal.

We expected that, relative to participants in the nonracial feedback condition, those in the racial feedback condition would show greater evidence of behavioral inhibition following high arousal feedback, would report greater levels of negself, and would show greater evidence of retrospective reflection. Finally, the extent of negself participants experienced was expected to be related to the extent of behavioral inhibition and retrospective reflection, which is consistent with the idea that cues for control were being established.

Method

Participants

Forty (16 male, 24 female) White participants were recruited from a larger pool of possible participants for Experiment 1. Eligible participants were identified on the basis of the racial attitudes they reported during a mass testing session at the beginning of the semester. Specifically, eligible participants had scores on the Attitudes Toward Blacks Scale (ATB; Brigham, 1993) of 51 or less and thus fell in the bottom third of the possible distribution of ATB scores, which is 20–140. Eligible participants also were those who self-reported being at least somewhat prone to discrepant responses in relation to Blacks during the mass testing, because these are the individuals for whom the operation of cues for control is relevant (e.g., one would not investigate how depressed individuals learn to control depressive thoughts among nondepressed individuals).3 In particular, participants had discrepancy scores of at least 2 on the 10-item Should–Would Discrepancy Questionnaire used by Devine et al. (1991).

In Experiment 2, 38 (10 male, 28 female) White participants were again recruited from a larger pool of possible participants. This time the Modern Racism Scale (MRS; McConahay, Hardee, & Batts, 1981) was included in the mass testing session to measure prejudiced attitudes. Eligible participants had MRS scores of 13 or less and thus fell in the bottom third of the possible MRS distribution, which is 128 to 28. Also, eligible participants had discrepancy scores of at least 10 on the 32-item Should–Would Discrepancy Questionnaire developed by Monteith and Voils (1998).

Eligible participants were identified on calling lists, called at random, and asked to participate in return for research credit applied to their introduction to psychology course grade.

Design

Participants viewed two types of pictures along with filler pictures. The racial pictures included Blacks, and the nonracial pictures did not. Depending on the condition to which participants were randomly assigned, they received either high or low negative arousal feedback in connection with these two types of pictures. Participants in the racial feedback condition received feedback suggesting they were having high negative arousal in relation to the racial pictures but low arousal in relation to the nonracial pictures. Participants in the nonracial feedback condition received feedback suggesting they were having high negative arousal in relation to the nonracial pictures but low arousal in relation to the racial pictures. An advantage of this design is that it held constant across feedback conditions the number of instances in which participants apparently could not reduce their negative arousal. Following practice trials, participants viewed several rounds of pictures, and they attempted to decrease any negative arousal they had across the rounds. There were three rounds in Experiment 1. Thus, Experiment 1’s design was a 2 (gender) × 2 (picture type: racial vs. nonracial) × 2 (feedback condition: racial feedback vs. nonracial feedback) × 3 (round: first, second, and third) mixed model, with feedback condition and gender treated as between-subjects factors and the other factors treated as repeated measures. The same design was used in Experiment 2, except that there were too few men to consider gender as a factor, and five rounds of viewing the pictures were used rather than three.

Materials and Procedure

Participants completed the experiment individually. They were greeted by an experimenter who, after obtaining informed consent, explained that the experiment was part of a nationwide project designed to gather data on the types of physiological arousal that occur when people are negatively aroused. In particular, the experimenter explained that the experiment was designed to determine how successful people are at controlling negative physiological reactions.

Initial activities. The experimenter explained the need to collect some baseline physiological arousal data, and she applied some electrode paste to the second phalanges of the first and third fingers of the participant’s nondominant hand and attached electrodes to those fingers. These electrodes were wired to actual physiological equipment that interfaced with the computer. The experimenter started a computer program that ostensibly collected the physiological data, and participants were able to view their (bogus) physiological activity on the computer monitor for a moment. Then the experimenter turned off the monitor and asked participants to relax for a few minutes and not to move their nondominant hand. The experimenter left the room and returned after 3 min to close the supposed program for collecting the baseline physiological data. This procedure was followed simply to enhance the cover story.

The experimenter went on to explain that, because the experiment was part of a nationwide project, a standardized, tape-recorded set of instructions was being used. She started the tape. A professional-sounding male voice explained the importance of controlling negative physiological arousal with statements such as the following:

2 Details of this pilot research are available from us on request.

3 Lest readers be concerned that this strategy serves to identify a select group of low-prejudice individuals, they should keep in mind that the vast majority of low-prejudice persons report that they are prone to discrepancies (see Monteith & Voils, 1998).
Studies such as these are important because the results can be applied to pain management, stress reduction, therapy for the depressed and, overall, increasing mental health. If people can learn to control their negative arousal, they will be able to live happier, healthier lives.

This portion of the tape also explained how the equipment measures and conveys physiological arousal with statements such as the following:

Recent technological developments have produced relatively easy methods for measuring and recording physiological arousal by using physiological equipment that interfaces with the personal computer. Your level of negative arousal will be transmitted through electrodes attached to your fingers that feed information to the computer, and a computer program will then convert your arousal level into a bar graph that you can view.

The tape recording then explained that participants should learn how to try to control their negative physiological arousal (e.g., deep breathing) by reading an article from Psychology Today (Lang, 1970). The experimenter stopped the tape at this point, provided participants with the article (along with a summary of it), and encouraged participants to study the summary.

After participants had studied the summary, the tape was started again. This portion of the tape first emphasized that the program was only designed to measure negative arousal and that positive emotions or general arousal could not be measured with the program. This procedure was followed so that participants would interpret any high arousal feedback they received as negative arousal. The mechanics of the experimental procedure (see details below) were also explained during this portion of the tape.

Pictures. The pictures were obtained from secondary resources. Two types of pictures were used during the practice trials only. These consisted of five pictures that were neutral in content (e.g., leaves) and five pictures that naturally would result in varying levels of negative arousal (e.g., bloody hands, a ferocious police dog). The first round of viewing, which followed the practice trials, included four types of pictures (all different from those used during practice trials): six neutral, five arousing, three nonracial, and three racial. The nonracial pictures included content that was intended to be ambiguous with respect to whether it would be associated with high levels of arousal. That is, the images were not violent or gruesome but conceivably could evoke negative arousal (e.g., an object that was featureless except that one could tell it was a head-and-shoulder sculpture that had some nails stuck in it). The racial pictures were not violent or alarming in any sense; nonetheless, if one were having a negative emotional reaction (reflecting the current level of arousal), they would see two bar graphs—one labeled LAST TIME (indicating how much arousal was experienced on the previous viewing), and one labeled THIS TIME (reflecting the current level of arousal).

Participants then began the second round of viewing, which included the arousing filler pictures and the racial and nonracial pictures. The third round of viewing was completed after participants were once again encouraged to try to control any negative arousal they might have. For participants in Experiment 1, the third round was the final round. Five rounds of viewing were included in Experiment 2 to determine whether behavioral inhibition would continue to occur across the additional rounds. To discourage an increase in suspicion that might result from adding the two rounds, we used only two racial and two nonracial pictures across the five rounds of viewing and added one filler arousal picture.

The feedback schedule for these various rounds is summarized in Table 1. Participants in both feedback conditions supposedly were able to lower their negative arousal in connection with the arousing filler pictures across the rounds. Participants in the racial feedback condition never were able to lower their arousal in connection with the racial pictures and never experienced high arousal in relation to the nonracial pictures. Participants in the nonracial feedback condition received a reverse pattern of feedback for the racial and nonracial pictures across all rounds.

Postviewing measures. The experimenter explained that the computer program would now be used to ask participants a few questions. The experimenter started a program that served to collect data concerning participants’ current feelings. Participants indicated the extent to which they were experiencing 29 different emotions, making each rating by typing a number between 1 (does not apply at all) and 7 (applies very much). The affect items are noted in the context of the results. The
Experimenters waited in the adjoining room while participants completed the items.

Next, participants in Experiment 1 completed manipulation check items, which assessed participants’ own perceptions of their level of physiological arousal in relation to the pictures. Specifically, the questionnaire instructions stated that

Research has shown that sometimes people perceive themselves to be experiencing the same level of arousal that registers on the physiological apparatus. However, on occasion, people think that they are experiencing a level of arousal that is inconsistent with the physiological measures.

The instructions then indicated that participants should report their own perceptions of how much arousal they experienced in relation to the pictures. The two ratings of greatest interest were for “scenes involving Blacks (e.g., interracial scenes)” and “disturbing scenes (strange scenes, though not violent).” Ratings were made on scales ranging from 1 (not at all arousing) to 7 (very arousing). Participants then made ratings to indicate their own perceptions of how successful they were at controlling their arousal in relation to the different types of pictures on scales ranging from 1 (not at all successful) to 7 (very successful). A final item was included to determine whether participants in the two feedback conditions perceived equal levels of success (or failure) throughout the slide-viewing procedure. This item asked participants to rate “how successful you think you were at controlling your physiological arousal in general” on a scale from 1 (not at all successful) to 7 (very successful).

Rather than completing manipulation checks, participants in Experiment 2 were asked to complete a thought-listing task about their reactions to the experiment. Participants were asked to record all thoughts they had concerning the experiment, writing one thought per box on a form. They were encouraged to write whatever occurred to them, including feelings about and reactions to the tasks they completed during the experiment. Interest centered on whether participants who received high arousal feedback in relation to the racial pictures would show evidence of being concerned about their inability to control their arousal (i.e., show evidence of retrospective reflection) whereas those who received high arousal feedback in relation to the nonracial pictures would not.

At the conclusion of the experiment, participants provided a written account of what they thought the experiment was about. This information, along with information obtained during debriefing, was used to assess suspicion. No participants from Experiment 1 had to be excluded because of suspicion. Four participants in Experiment 2 (beyond the number noted in the Participants section) had to be excluded (3 from the racial feedback condition, and 1 from the nonracial feedback condition).

### Results and Discussion

#### Manipulation Checks: Experiment 1

A 2 (feedback condition: racial vs. nonracial) × 2 (gender) between-subjects analysis of variance (ANOVA) performed on the rating of how arousing participants perceived the racial pictures to be revealed a main effect for gender, $F(1, 36) = 16.46, p = .001, r = .36$; a main effect for feedback condition, $F(1, 36) = 36.69, p = .001, r = .65$; and a Gender × Feedback Condition interaction, $F(1, 36) = 10.03, p = .003$. The nature of the interaction indicated that, although both men and women in the racial feedback condition reported that they perceived more arousal than did participants in the nonracial feedback condition, this difference was larger for female participants (respective racial and nonracial $M_s = 4.50$ and 1.31, $r = .81$) than for male participants (respective racial and nonracial $M_s = 2.00$ and 1.00, $r = .53$).$^4$

The ANOVA performed on the rating for how arousing participants perceived the nonracial pictures to be yielded a main effect for gender, $F(1, 36) = 16.88, p = .001, r = .53$, such that women ($M = 4.37$) reported greater arousal than did men ($M = 2.56$). There was also a significant main effect for feedback condition, $F(1, 36) = 10.27, p = .003, r = .43$, such that participants in the nonracial feedback condition perceived greater arousal ($M = 4.33$) than did participants in the racial feedback condition ($M = 2.89$). The interaction between gender and feedback condition was non-significant ($F < 1.00$).

The ANOVA performed on participants’ perceived success at controlling arousal in relation to the racial pictures revealed a significant main effect for feedback condition only, $F(1, 36) = 26.89, p = .001, r = .65$, such that greater success was perceived in the nonracial feedback condition ($M = 6.43$) than in the racial feedback condition ($M = 3.21$). Likewise, the only significant effect for the perceived success ratings for the nonracial

---

$^4$ Three of the 8 male participants in the racial feedback condition gave seemingly nonsensical ratings across the arousal and success manipulation checks, which may indicate that a misinterpretation of the scale for the arousal ratings caused them to mistakenly report relatively low levels of arousal for the racial pictures. They reported very low arousal for the racial pictures but also rated themselves as highly unsuccessful at controlling their arousal for these pictures. Thus, the interaction may have resulted because of participants’ errors when making their ratings.
pictures was a main effect for feedback condition, $F(1, 36) = 15.94, p = .001, r = .55$, with greater success reported in the racial feedback condition ($M = 5.90$) than in the nonracial feedback condition ($M = 4.14$). These findings indicate that participants perceived the feedback as intended.

Despite the different levels of arousal and success experienced in relation to certain pictures as a function of feedback condition, the ANOVA performed on reports of how successful participants perceived themselves to be at controlling their arousal throughout the entire experiment did not reveal any significant effects, $F < 1.20, ps > .28$. Thus, as intended, participants in the two feedback conditions perceived overall equal levels of success at controlling their negative physiological arousal.

**Affect: Experiments 1 and 2**

As in past research (e.g., Devine et al., 1991; Monteith & Voils, 1998), we formed the following five affect indices by averaging ratings for relevant items: negself (i.e., “angry at myself,” “guilty,” “annoyed with myself,” “disappointed with myself,” “disgusted with myself,” “regretful,” “ashamed,” and “self-critical”), discomfort (i.e., “fearful,” “uneasy,” “embarrassed,” “bothered,” “anxious,” “tense,” “threatened,” “uncomfortable,” and “frustrated”), positive (i.e., “friendly,” “happy,” “energetic,” “optimistic,” and “good”), down (i.e., “depressed,” “sad,” “helpless,” and “low”), and negative other-directed affect (hereafter referred to as negother; i.e., “angry at others,” “irritated with others,” and “disgusted with others”). The range of reliability coefficients (alphas) across the various affect indexes and the two experiments was .81 to .95.

Table 2

<table>
<thead>
<tr>
<th>Affect index</th>
<th>Racial</th>
<th>Nonracial</th>
<th>$r$</th>
<th>$p$</th>
<th>Racial</th>
<th>Nonracial</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negself</td>
<td>2.97</td>
<td>1.60</td>
<td>.49</td>
<td>.001</td>
<td>3.21</td>
<td>1.85</td>
<td>.41</td>
<td>.01</td>
</tr>
<tr>
<td>Discomfort</td>
<td>3.02</td>
<td>2.20</td>
<td>.38</td>
<td>.02</td>
<td>3.20</td>
<td>2.23</td>
<td>.39</td>
<td>.02</td>
</tr>
<tr>
<td>Positive</td>
<td>4.11</td>
<td>4.85</td>
<td>.29</td>
<td>.08</td>
<td>3.64</td>
<td>4.17</td>
<td>.19</td>
<td>ns</td>
</tr>
<tr>
<td>Down</td>
<td>2.42</td>
<td>1.80</td>
<td>.28</td>
<td>.08</td>
<td>2.26</td>
<td>1.54</td>
<td>.30</td>
<td>.07</td>
</tr>
<tr>
<td>Negother</td>
<td>2.03</td>
<td>2.20</td>
<td>.06</td>
<td>ns</td>
<td>1.92</td>
<td>1.54</td>
<td>.18</td>
<td>ns</td>
</tr>
</tbody>
</table>

Note. Experiment 1, $N = 40$; Experiment 2, $N = 38$. Negself = negative self-directed affect; Negother = negative other-directed affect.

Recall that the time participants took to press the space bar following the arousal feedback they viewed for each picture was measured in milliseconds to provide an indication of behavioral inhibition in relation to the feedback. Outlying reaction times were defined as those falling three standard deviations from the mean, and these data points were eliminated. Outliers were 0.8% of the data in Experiment 1 and 3.5% of the data in Experiment 2. Reaction time data from 1 participant in Experiment 2 were eliminated entirely (before the overall outlier rate was calculated) because of consistent outlying times. Indices representing the average reaction times for the arousing filler, racial, and nonracial pictures were formed. Analyses were performed on log-transformed data but are summarized here in their original millisecond form.

For both experiments, the mixed model ANOVA performed on the reaction time data treated feedback condition (racial vs. nonracial) as a between-subjects factor and picture type (racial, nonracial, or arousal) and round (Rounds 1–3 for Experiment 1 and Rounds 1–5 for Experiment 2) as within-subject factors. Also, gender was included as a factor in Experiment 1. The results for both experiments revealed a variety of significant lower order effects, all of which were qualified by a higher order Feedback Condition × Picture Type × Round interaction, $F_{Exp 1}(1, 36) = 4.73, p < .04$, and $F_{Exp 2}(1, 34) = 23.96, p < .001$. The patterns of these interactions are shown in Figures 2 and 3.

**Behavioral Inhibition Following Arousal Feedback: Experiments 1 and 2**

In addition, there were trends toward racial feedback participants feeling more down (in both experiments) and less positive (in Experiment 1). These findings provide strong evidence that participants’ perceptions of negative responses in relation to Blacks had negative affective consequences. The only significant effect associated with gender in the analyses of Experiment 1 data was that women reported overall higher levels of discomfort ($M = 2.87$) than did men ($M = 2.24$), $F(1, 36) = 4.12, p = .05, r = .27$. 

Despite the different levels of arousal and success experienced in relation to the racial slides, in the form of negative and positive physiological arousal, receiving high feedback in relation to the racial slides would cause a unique type of affective reaction, in the form of negative and positive other-directed affect (hereafter referred to as negother; i.e., “angry at others,” “irritated with others,” and “disgusted with others”). The range of reliability coefficients (alphas) across the various affect indexes and the two experiments was .81 to .95.
in relation to the other two types of pictures (ps < .001), but there was no significant difference between reaction times for the nonracial and arousal pictures. Thus, when participants in the nonracial feedback condition were initially given high arousal feedback in relation to the nonracial pictures, it was not associated with greater behavioral inhibition than when they were presented with the same level of high arousal feedback in relation to the arousal pictures. In contrast, when Experiment 1 participants in the racial feedback condition received their first feedback in Round 1, they showed greater evidence of behavioral inhibition following feedback for the racial pictures than following feedback for either the nonracial or the arousal pictures (ps < .001), and the means for these latter two conditions were not significantly different. Thus, in the racial feedback condition, the same level of high arousal feedback was associated with greater behavioral inhibition for the racial pictures than for the arousal pictures. A similar pattern emerged in Experiment 2. None of the comparisons as a function of picture type was significant for the nonracial pictures among participants in the nonracial feedback condition in Round 1 (in fact, this was also the case for Round 2). In contrast, behavioral inhibition was especially pronounced in the racial feedback condition for the racial pictures (with reaction times for these pictures being significantly different than times for the arousal and nonracial pictures, ps < .02). These findings indicate that behavioral inhibition was immediately apparent when high arousal feedback was provided in connection with the racial pictures only.

We turn now to reaction times for the final round. The patterns in both experiments indicate that participants showed higher levels of behavioral inhibition in relation to pictures for which they were apparently unable to reduce their arousal. That is, participants in the nonracial feedback condition showed greater evidence of pausing for the nonracial pictures than for the other pictures (ps < .01), and participants in the racial feedback condition showed greater evidence of pausing for the racial pictures than for the other pictures (ps < .05). Thus, behavioral inhibition was equally evident in both feedback conditions whenever participants consistently were unable to control their negative arousal.

Altogether, these findings suggest that behavioral inhibition was an immediate and enduring response to feedback that participants were showing high negative arousal in relation to pictures of Blacks, but it occurred for race-unrelated pictures only after participants were repeatedly unable to reduce their level of negative arousal. It is sensible that the pattern emerged for participants in the nonracial condition in that they undoubtedly expected to be able to reduce their negative arousal in relation to the nonracial pictures (after all, they were able to bring it down for the arousal pictures), and the inconsistency between this expectancy and the actual state of affairs would be expected to result in BIS activity.

Given that behavioral inhibition and aversive consequences in the form of negative affect are both thought to be involved in the establishing of cues for control, we expected reaction times for the racial pictures and negative affect to be significantly correlated. As shown in Table 3, reaction times for the racial pictures were significantly related to negself and discomfort feelings in both experiments. The worse participants felt about their performance during the picture viewing task, the longer they paused for feedback in connection with the racial pictures. Trends toward significant relations were also obtained for other affect indices. In contrast, none of the correlations between reaction times to feedback for the nonracial pictures and affect approached signifi-

Figure 2. Reaction times (RTs) to racial and nonracial pictures as a function of round and feedback condition, Experiment 1.

Figure 3. Reaction times (RTs) to racial and nonracial pictures as a function of round and feedback condition, Experiment 2.
Table 3

<table>
<thead>
<tr>
<th>Affect index</th>
<th>Racial</th>
<th>Nonracial</th>
<th>Racial</th>
<th>Nonracial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negself</td>
<td>.44**</td>
<td>-.13</td>
<td>.42**</td>
<td>-.18</td>
</tr>
<tr>
<td>Discomfort</td>
<td>.46**</td>
<td>.01</td>
<td>.34*</td>
<td>-.14</td>
</tr>
<tr>
<td>Positive</td>
<td>-.33*</td>
<td>.04</td>
<td>-.25</td>
<td>-.11</td>
</tr>
<tr>
<td>Down</td>
<td>.17</td>
<td>-.10</td>
<td>.31†</td>
<td>-.04</td>
</tr>
<tr>
<td>Negother</td>
<td>-.19</td>
<td>-.14</td>
<td>-.02</td>
<td>-.21</td>
</tr>
</tbody>
</table>

Note. Experiment 1, N = 40; Experiment 2, N = 37. Negself = negative self-directed affect; Negother = negative other-directed affect.
† p = .064. * p < .05. ** p < .01.

cance. These findings are consistent with the idea that cues for control were being developed when participants believed they were having negative reactions to slides that included Blacks.

Reactions to Experiment: Experiment 2

The thought-listing protocols for participants’ reactions to the experiment were content analyzed. First, the total number of thoughts was tallied. On average, 3.84 thoughts were recorded, and the number of thoughts recorded did not differ as a function of participants’ feedback condition. Second, the thoughts were content analyzed according to a five-category scheme that encompassed all thoughts generated. Two coders analyzed all of the thoughts. Their percentage of agreement was 94%, and disagreements were resolved through discussion. Although five categories were used in the content analysis, only one was of theoretical interest. Indeed, this was the only category that was associated with any significant effects. This category included any reference to an inability to control arousal accompanied by concern over this inability (e.g., “I found it disturbing that my highest negative responses were to the pictures involving African-Americans” and “I feel upset because I couldn’t get my negative responses down in two pictures”).

We calculated retrospective reflection scores by dividing the number of times each participant’s thoughts fell in this category by the total number of thoughts each listed, and we used t tests to analyze the proportion of times this category was used as a function of whether participants were in the racial or nonracial feedback condition. As expected, participants in the racial feedback condition were significantly more likely to be concerned with their inability to control their arousal, t(35) = 2.04, p < .05, r = .33. On average, one out of every four thoughts generated by participants in the racial feedback condition reflected such concerns (M = 0.22), but virtually none of participants’ thoughts in the nonracial feedback condition did so (M = 0.01). These findings provide further evidence of processes that should be at work when cues for control are being established. Specifically, participants in the racial feedback condition engaged in extensive retrospective reflection, which occurs when associations are being built among discrepant responses, their aversive consequences, and the stimuli that predict those consequences (see Patterson & Newman, 1993). Further supporting this reasoning, the correlation between participants’ retrospective reflection score and their reported neg-

self was significant, r(37) = .61, p < .001. The correlation between participants’ retrospective reflection scores and behavioral inhibition for the racial pictures was marginally significant, r(37) = .28, p < .08.

In sum, Experiments 1 and 2 establish that events that theoretically are critical to the development of cues for control do occur. When participants were led to believe they had prejudiced responses, they showed evidence of three outcomes that occur when cues for control are being established. Behavioral inhibition occurred, and participants experienced an aversive event (i.e., negself) and engaged in retrospective reflection. These events facilitate the building of associations among the unwanted prejudiced response, negative affect, and stimuli that predict the occurrence of an unwanted prejudiced response.

Experiment 3

Whereas the focus of Experiments 1 and 2 is on events that should occur if cues for control are being established, the focus of Experiments 3 and 4 is on the operation of established cues for control.

Experiment 3 examines a key consequence of having established cues for control that, according to the present theoretical account, is essential for gaining control over automatized patterns of responding. We investigated whether the presence of an already-established cue for control in a stereotype-relevant situation triggers BIS activity and, more specifically, results in behavioral inhibition. Theoretically, this interruption of ongoing behavior then provides the opportunity for prospective reflection (see right portion of Figure 1), so that one has the opportunity to generate an alternative to the response that would have occurred had automatic processes continued uninterrupted.

The stimuli we used as cues for control in this experiment were photographs of Blacks paired with stereotypic descriptions. For example, presenting a photograph of a Black person accompanied by the description “This person spends a lot of time on the streets” involves pairing the Black person with a stereotypic situation (i.e., that the person is a bum). Because most low-prejudice people report experiencing guilt in relation to stereotypic responses to Blacks (e.g., Devine et al., 1991; Monteith & Voils, 1998), such a pairing should constitute an already-established cue for control.

---

5 A more stringent test would be to perform the correlational analyses within the racial feedback conditions to determine whether greater behavioral inhibition was associated with greater negself. The drawback of such an analysis is that participants in the racial feedback condition were expected to score (and indeed did score) relatively high on both of these variables, which resulted in restricted ranges. Nonetheless, correlational analyses within the racial feedback condition are in the expected direction in both experiments, r(18) = .42, p < .07, and r(16) = .23, p = .37, for Experiments 1 and 2, respectively. In contrast, it is clear that cues were not being developed in the nonracial feedback condition, for which the correlation between pausing and negself did not approach significance, r_{exp 1} = .05, and r_{exp 2} = .05, ns (one outlying affect data point was removed for Experiment 2).

6 The other categories were (a) questioning the accuracy of the feedback, (b) referring to inability to control arousal, but in a neutral way, (c) positive thought relating to ability to control arousal or about the experiment, and (d) miscellaneous.
Behavioral inhibition in the form of prospective reflection was examined in the context of a dual-task activity. One task purportedly examined visual memory. For this task, participants indicated whether photographs of people that were presented individually on a computer screen were the same photographs as they had been shown earlier in the procedure or were the mirror image of the original photographs. The second task purportedly examined cognitive inferences. Participants were presented with person descriptions one at a time on a computer screen and provided the first response that came to mind in relation to each person description. For example, one person description was “This person has to do a lot of reading,” and a frequent response was “college student.” Trials for the two activities occurred sequentially, such that a photograph appeared on the computer screen, participants indicated whether it was the original or not using specified keys on the keyboard, the person description appeared on the screen, participants provided a verbal response as quickly as possible, then another photograph appeared, and so on. The dual nature of this task, coupled with the fact that determining whether the photographs were originals or mirror images was not a simple visual activity, made for an overall cognitively taxing situation.

Sometimes the photographs that appeared were of Black men, and sometimes they were of White men. Also, some of the person descriptions were purposely constructed so that the most natural responses would be stereotype consistent. For example, “drug addict” is a natural response to “This person uses needles for recreation.” We expected that, when a photograph of a Black person was followed by a person description for which a stereotype-consistent response would naturally come to mind, the Black person would constitute a cue for control and result in behavioral inhibition. In contrast, when the same person description was preceded by a photograph of a White person, behavioral inhibition was not expected to occur. These predictions were for participants who were preselected so that they held low-prejudice standards for how they believe they should respond to Blacks. For comparison purposes, we also included high-prejudice participants in this experiment. Because individuals whose standards are relatively prejudiced are likely to have experienced little guilt in relation to their prejudiced responses in the past (Monteith, 1993), the Black photographs paired with stereotypic situations should not result in behavioral inhibition among these participants.

**Method**

**Participants**

Seventy-two White participants (34 male and 38 female) from the introductory psychology participant pool were preselected for participation. The 16-item Should scale from Monteith and Voils’s (1998) Should--Would Discrepancy Questionnaire had been included in the semester’s mass testing and was used to assess the extent to which participants’ personal standards for responding to Blacks were prejudiced. (The correlation between Should scores and standard measures of prejudice such as the ATB routinely are greater than .80.) The possible range of Should scores is 16 (low-prejudice standards) to 112 (high-prejudice standards). Participants were preselected to obtain as reasonable a representation across this range as possible (obtained range = 16–74; M = 30.10, SD = 12.37) while not confounding gender with the extent to which their personal standards were prejudiced (the correlation between Should scores and gender in the obtained sample was .05, n.s.). In the other experiments, participants were also preselected such that they were at least somewhat prone to discrepant responses, having scores of 7 or greater on the Monteith and Voils (1998) Should--Would Discrepancy Questionnaire. Potential participants were contacted by phone by a caller who was naive to the preselection criteria. Participants were given 1 hr of research credit in exchange for their participation.

**Materials**

**Photographs.** The photographs for the visual memory task were of four different men, two of whom were White, and two of whom were Black. The photographs showed the subjects’ faces and necks only. They were selected on the basis of their similar composition and of the similar expressions of their subjects. Each photograph was digitized, cropped to eliminate background details, and resized to a 2 × 1 in. (5.08 × 2.54 cm) image that was centered on a blue background. We created mirror images of the photographs by reversing the image, which resulted in eight images altogether.

**Person descriptions.** Twenty-four person description sentences were used for the inferential task. Each sentence described an activity that a person who belongs to a particular category of people would be likely to do (e.g., “This person works with numbers” could activate the default response of “mathematician”). Six sentences were critical sentences, such as “This person can be found on the streets” (stereotype-consistent response) and, if not first to come to mind, a stereotype-consistent response was generated within the 10-s period by all but a small minority of participants. The 6 critical sentences were as follows:

1. (“This person can be found on the streets”) (stereotype-consistent response: “bum,” “homeless person”); first response to come to mind for 73% of the participants was stereotype consistent; stereotype-consistent response was generated during the 10-s period among 92% of the participants).

2. (“This person uses needles for recreation”) (stereotype-consistent response: “drug addict”; first response, 75%; generated sometime, 83%).

3. (“This person depends on money from the government”) (stereotype-consistent response: “person on welfare”; first response, 70%; generated sometime, 76%).

*Personal standard data for 2 of the participants were not available, apparently because of a clerical error in recording the identification numbers for these participants, which typically allows us to match data from the experimental session with data from the mass testing. These participants’ data were necessarily excluded from all analyses involving personal standards.*
4. “This person can be found behind bars” (stereotype-consistent response: “prisoner,” “criminal”; first response, 82%; generated sometime, 94%).

5. “This person deals with a lot of drugs” (stereotype-consistent response: “drug dealer”; first response, 67%; generated sometime, 92%).

6. “This person is good at getting into locked doors” (stereotype-consistent response: “burglar”; first response, 67%; generated sometime, 83%).

Programs. Two computer programs were created (using MEL Professional 2.0 software; Schneider, 1995) for presenting the 24 person descriptions, each of which was preceded by a photograph of a White or a Black person. The two programs differed according to which critical sentences were paired with White versus Black photos. Specifically, in one program, Items 1–3 (above; hereafter called Item Set 1) were preceded by a photograph of a White person, and Items 4–6 (above; hereafter called Item Set 2) were preceded by a photograph of a Black person. In the other program, Item Set 1 was preceded with a photo of a Black person, and Item Set 2 was preceded with a photo of a White person. By creating the two programs, we were able to counterbalance which items were paired with photos of which race across participants. This enabled us to ensure that, if certain participants did show evidence of behavioral inhibition when generating responses for critical sentences paired with a photo of a Black person, this occurred because a Black person was being paired with a stereotypic context (i.e., in the presence of a cue for control) and not simply because the critical sentence itself resulted in unusually long response times.

Design

The design was a 2 (personal standards: low prejudice vs. high prejudice) × 2 (critical item set: 1 vs. 2) × 2 (pairing scheme: White photographs paired with Item Set 1 and Black photographs with Item Set 2 vs. Black photos paired with Item Set 1 and White photos paired with Item Set 2) mixed model. Critical Item Set 1 versus 2 was treated as a repeated measure, and the other factors were treated as between-subjects factors. Approximately an equal number of participants were randomly assigned to the pairing scheme conditions. Sex was initially included in analyses but was not associated with significant effects and therefore was dropped from analyses.

Procedure

Participants completed the experiment individually. After obtaining informed consent, the experimenter explained that cognitive categorization and inferential processes were of interest, so participants would complete a number of different cognitive tasks. In actuality, the first three tasks were fillers, included to reduce suspicion about the final task. These filler tasks included a computerized name categorization task, an anagram task, and a reading comprehension task. Altogether, they took approximately 20 min to complete.

The main experimental task was described as a dual-processing task. The experimenter explained that the purpose of this task was to assess how well people can simultaneously perform two cognitive activities—a visual memory task and an inferential task. The experimenter then started one of the two computer programs. Each program first displayed a series of four photographs individually. Participants were given an opportunity to study each photograph until they felt confident that they could later recognize it. The program then presented an example of the inference task. Participants read a sentence (“This person cuts a lot of hair”) and were asked to state aloud the first response that came to mind (e.g., “hairstylist”). So that participants could expect that responses that were more negatively valenced were also acceptable, the experimenter also provided the example “This person gets in a lot of fights” and the response “gang member.” The experimenter then emphasized that participants were to press the space bar immediately after providing their response.

The experimenter then explained that these two tasks would be alternated in a dual-task activity. For the visual memory portion of task, participants would first see a photograph that would be either one of the original four photographs shown earlier or its mirror image. They were to indicate as quickly as possible whether the photograph was the same as or different from the photographs they saw at the beginning of the program by pressing the Z key (labeled S for same) or the M key (labeled D for different). When participants made their response, a sentence that described a person then appeared on the screen. Participants were instructed to say aloud the first response that came to mind for the person description as quickly as possible. Responses were recorded with a tape recorder and were later transcribed by the experimenter. The experimenter emphasized that participants should press the space bar at the same time as they provided each verbal response. Unbeknownst to participants, the computer recorded the amount of time that expired between the presentation of each person description sentence and when participants pressed the space bar. This time to generate a response (i.e., indication of pausing) served as the measure of behavioral inhibition.

After explaining these details and answering any questions participants had, the experimenter reiterated that participants should complete the task as quickly as possible, stating the first response that came to mind for each sentence. Then, as the experimenter started the tape recorder, he or she mentioned, “Oh, by the way, you can assume that each person description refers to the person in the preceding photograph.” The experimenter then left the room and waited in an adjoining room.

For all participants, the first 5 person descriptions were filler sentences included to familiarize participants with the task. The remaining person descriptions were presented in a fixed random order. Across the two critical item sets, half of the sentences were randomly paired with photographs of Blacks, and half were paired with photographs of Whites. Further, each of the four photographs (i.e., of the two Black men and the two White men) was presented an approximately equal number of times both in its original form and in the mirror image form. The same strategy was used to pair photographs and sentences for the 18 filler items. One of the filler descriptions was clearly negative, and several others could elicit negative responses. Thus, the stereotype relevancy of the items was not confined with negativity.

When participants completed the dual-task activity, they were probed for suspicion, debriefed, and provided with their experimental credit.8

Results and Discussion

Visual Memory Task

Participants’ responses to the photographs (i.e., “same” vs. “different”) and the time they took to provide their responses had been automatically recorded by the computer. Thus, these data could be examined to ensure that participants were paying attention to the task and that it seemed to be cognitively taxing. The overall accuracy rate of participants’ responses was high (M = 0.82, SD = 0.16), suggesting that participants were attempting to complete the task accurately. Nevertheless, the accuracy rate certainly was not perfect, pointing to the difficulty of the task. The reaction time data were first examined to identify and eliminate outlying times (defined as three standard deviations from the mean

8 On the basis of their written and/or verbal comments, 7 participants appeared to be suspicious about the procedure, suspecting that it may have something to do with stereotypes. Analyses performed without these participants did not yield altered results. Data from 2 participants were excluded altogether because these participants apparently did not understand directions (e.g., one said “same” or “different” aloud for each trial rather than providing a verbal response to each person description).
of response times for a given trial). There were relatively few outliers (1.9%), suggesting that participants were attempting to respond to each trial in a reasonable time frame. Also, the fact that participants took, on average, about 2 s ($M = 2,091.00$ ms, $SD = 477.00$ ms) to decide whether the photographs were the same or the mirror images suggested that the task was rather taxing.

**Inferential Task**

**Behavioral inhibition.** The time participants took to respond to the person description sentences was first examined to identify and eliminate outliers, which were defined as values more than three standard deviations from the mean of response times for a given trial. For the filler sentences, 2.2% of the data constituted outliers. A similar rate of 1.5% was observed for the critical item sets. Response times for the person descriptions were log transformed, and analyses were performed on the transformed values. However, results are reported in terms of milliseconds to ease interpretation.

We formed several indices by averaging response times for the relevant person description items. We formed an Item Set 1 index by averaging response times for the three critical items in this set. We formed an Item Set 2 index in a like manner, using the three critical items in this set. Finally, the remaining response times for the filler person descriptions were averaged to form an index.

Our central predictions concerned how long participants would take to respond to the critical items depending on whether those items were preceded by White or Black photographs and on the nature of participants’ personal standards. We consider first participants with low-prejudice personal standards. Because these participants should show evidence of behavioral inhibition when a Black person is encountered in a stereotypic context (i.e., in the presence of an already-established cue for control), they were expected to respond more slowly to critical items when these items were paired with Black faces than when the same items were paired with White faces. In contrast, evidence of behavioral inhibition was not expected among participants with high-prejudice personal standards, who should not have such preexisting cues.

These predictions were tested with a $2 \times 2 \times 2$ mixed model ANOVA. Item set (Critical Item Set 1 vs. 2) was treated as a repeated measures variable, and pairing scheme (Item Set 1 preceded by Black photos and Item Set 2 preceded by White photos vs. Item Set 1 preceded by White photos and Item Set 2 preceded by Black photos) was treated as a between-subjects factor. The personal standards measure was dichotomized such that participants with Should scores of 26 or less were defined as having low-prejudice personal standards and participants with Should scores greater than 26 were defined as having high-prejudice personal standards. This strategy was used because we believe it provides the clearest illustration of the pattern of results (but for a regression analysis approach, see Footnote 9). The average response time for the filler items was used as a covariate.

The ANOVA revealed a significant effect for the covariate, $F(1, 65) = 79.29, p = .001$. (The effect of the covariate was significant in all analyses in which it was included, but because it is of little theoretical importance, it is not mentioned further.) Of greater interest, the analysis also revealed a significant main effect for personal standards, $F(1, 65) = 6.04, p < .02$, and a significant Item Set $\times$ Pairing Scheme interaction, $F(1, 65) = 12.55, p < .001$. These effects were further qualified by the predicted three-way interaction among personal standards, item set, and pairing scheme, $F(1, 65) = 5.08, p < .03$. As can be seen in Figure 4, participants with low-prejudice standards responded differently to Item Sets 1 and 2 depending on whether those items were paired with Black or White faces. Specifically, these participants were slower whenever the items were paired with Black faces, thus showing evidence of behavioral inhibition in the presence of a cue for control. The two-way interaction among participants with low-prejudice standards was significant, $F(1, 28) = 21.99, p =
.001. Also, within each item set and among participants with low-prejudice standards, the effect of pairing scheme was significant, $F$s > 6.25, $p$s < .02, $b$s > .33. Although high-prejudice participants showed a tendency toward the same pattern, it was much less pronounced, and the interaction between item set and pairing scheme was not significant, $F < 1.00.9$

Responses to critical items. The present theoretical account of the mechanisms that enable individuals to change automatic patterns of responding maintains that the brief stop that behavioral inhibition places on ongoing behavior enables individuals to inhibit prejudiced responses and to generate alternative, more personally acceptable responses instead. Thus, we examined the possibility that alternatives to stereotype-consistent responses were generated for the critical items in the present research, particularly among participants with low-prejudice standards and when the critical items were preceded by photographs of Black persons.

The tape recordings of participants’ responses to the critical items were dummy coded according to whether the responses were stereotypic of Blacks (coded as 1) or not (coded as 0). We generated overall scores for each item set by adding the relevant values for items within each item set. These scores were then examined in a $2 \times 2 \times 2$ ANOVA with item set, pairing scheme, and personal standards as the factors, with item set as the repeated measures variable. No significant effects were obtained (nor were effects significant in parallel regression analyses). Analyses of responses for each individual person description item also did not yield any significant effects.10

Although finding that stereotype-consistent responses were replaced would have nicely complemented the behavioral inhibition data, there were four characteristics of the research that made such a finding very unlikely. First, participants were explicitly instructed to provide the first response that came to mind. Second, in an effort to examine the likelihood of behavioral inhibition, we created a situation in which stereotype-consistent responses were the most likely responses to the critical item sets, regardless of whether the sentences were paired with White or Black persons. This was clearly established in our pilot work. By being relatively certain that a stereotype-consistent response would come to mind, we could determine whether behavioral inhibition occurred in the presence of a cue for control. However, because the stereotype-consistent responses were perfectly appropriate responses to the critical items (i.e., such responses did correspond highly to the person descriptions), the responses could be expected even if behavioral inhibition had been triggered by the situation. Third, even if participants attempted to generate alternatives to stereotype-consistent responses, they may not have been able to think of alternatives because of the manner in which we had intentionally constructed the sentences. Fourth, the measure of stereotype-consistent responses had a restricted range because each critical item could yield only one of two values (i.e., 0 for responses not related to the stereotype or 1 for stereotype-consistent responses), and critical items were paired with Black photographs in only three instances (a procedure that helped to reduce the likelihood of participant suspicion). A dependent measure with a possible range of only 0–3 is not very sensitive. Thus, a number of characteristics of the research may have prevented us from observing the replacement of stereotypic responses. Given the importance of the argument that replacement ultimately should occur when cues for control are present, Experiment 4 was designed to test this part of the self-regulation model.

Experiment 4

The goals of Experiment 4 were to actually establish cues for control, to examine subsequent prospective reflection in the presence of those cues, and to determine whether people who have established cues for control are less likely to generate responses that could be construed as racially biased when those cues are present. The task used for establishing cues for control was the racial Implicit Association Test (IAT; Greenwald et al., 1998). This is a dual-categorization task that assesses the extent to which automatically activated evaluations in relation to Blacks are less pleasant than in relation to Whites. More specifically, the task measures the strength of association between pleasant and unpleasant words (e.g., sunshine vs. sink) and, historically, White versus Black names (e.g., Adam vs. Tyrone). Nearly all Whites show an implicit bias that favors Whites over Blacks (Greenwald et al., 1998), and many people can feel their racial biases in relation to the IAT (Monteith, Voils, & Ashburn-Nardo, 2001). That is, they realize that they have a more difficult time pairing pleasant words with Blacks and unpleasant words with Whites than the reverse. For this reason, taking the IAT might serve as a cue-establishing experience among low-prejudice individuals. However, having low-prejudice standards appears to be a necessary but not sufficient condition for experiencing negself in relation to racial IAT performance (Monteith et al., 2001). This is because some low-prejudice individuals attribute their IAT performance to factors other than race or racial bias (e.g., being unfamiliar with Black names). Only individuals who attribute performance to racial bias should establish cues for control as a consequence of completing the IAT, as only these individuals experience a discrepancy and the resulting punishment. Thus, negself following IAT performance was taken as the index of the extent to which cues had been established.

9 Another data analysis approach can be used that preserves the continuous nature of the prejudice measure. First, pausing time for Item Set 1 can be standardized within its distribution, and the same can be done for Item Set 2. (Standardization of these variables controls for the overall tendency to take longer to respond to items in Item Set 1.) Second, the difference between time to respond for each item set can be computed such that (regardless of pairing scheme), higher numbers reflect slower times for items paired with Black photos than for items paired with White photos. When this was done and participants’ personal standards were used to predict their difference scores, a significant effect emerged, $r(67) = -.27$, $p < .03$. The less prejudiced participants’ personal standards were, the longer they paused for items paired with Black relative to White photos. Although it also provides support for our hypothesis, we believe that this approach does not illustrate the pattern of findings as clearly as the approach reported in the text.

10 Although the finding is not associated with significant effects, all participants did not uniformly provide stereotype-consistent responses to the person descriptions when they were paired with Blacks. Additional analyses of the reaction time data as a function of whether responses were stereotype consistent did not change the pattern of results (i.e., participants with low-prejudice standards took relatively longer to respond when the stereotypic descriptions were paired with Black photos, regardless of whether they ultimately provided stereotype-consistent responses).
Later in the study, participants were presented with a series of words, presented individually on the computer screen, and indicated as quickly as possible whether they liked or disliked each word. Some of these words were historically Black names used in the IAT. We expected participants who felt worse about their IAT performance to show evidence of behavioral inhibition in relation to these words, such that their behavior would be briefly interrupted. Further, because indicating that one does not like historically Black names could be construed as racially biased, participants should be more likely to provide “like” responses to the critical trials including Black names than should participants who had experienced less nsegself.

It is important to note that most participants had relatively low-prejudice standards in this experiment because the methodology was such that it should facilitate the development and operation of cues for control only among persons with such standards. Although recruiting participants who vary more in their personal standards (such as in Experiment 3 herein and in Monteith, 1993) helps to underscore the importance of individual differences, providing another demonstration of this was not a goal in the present study.

Method

Participants

Forty-eight White (23 male, 25 female) introductory psychology students were preselected on the basis of their responses to the Should portion of the Monteith and Voils’ (1998) Should–Would Discrepancy Questionnaire, which had been included in a mass testing session at the beginning of the semester. We allowed Should scores to vary (M = 29.13, SD = 13.54), although the distribution was positively skewed such that most participants held relatively low-prejudice standards. Also, participants had scores of at least 4 on the Monteith and Voils’ (1998) Discrepancy Questionnaire. Participants either received research credit as part of their course requirement or were paid $10.

Procedure

Participants completed the experiment individually. The experimenter informed them that the study concerned categorization and that they would complete several tasks related to how well people can categorize items. The experimenter explained that he or she would provide instructions prior to each task and then leave the room so that the participant could do each task privately. A number of the tasks were presented by computer, with programming accomplished with MediaLab (Jarvis, 1998) and Inquisit (Millisecond Software, 2000).

Participants first completed a filler task that required them to identify 80 names as either male or female. Participants then completed a racial IAT (FIAT for Windows 2.3; Farnham, 1998) that required them to categorize items as either Black or White names or pleasant or unpleasant words. These names or words randomly appeared one at a time on the computer screen, and participants had to categorize them into their appropriate groups by pressing either the left (A key) or the right (S key on the number keypad) key. Instructions explaining which words would be categorized on the left and right preceded all blocks. All trials also included category labels indicating which categories corresponded to their appropriate sides and response keys (e.g., pleasant on the left side, and unpleasant on the right side). The names and words were drawn from Greenwald et al.’s (1998) stimulus lists. Eighteen different Black and White names were used. Because we wanted the ease of pairing pleasant versus unpleasant words with each type name to be especially salient, we used only six of the most pleasant (happy, paradise, pleasure, love, peace, cheer) and unpleasant (filth, vomit, death, evil, ugly, rotten) words from Greenwald et al.’s stimulus lists.

The IAT consisted of seven blocks of trials presented in a fixed order across participants:12

1. Categorize White names on left, Black names on right.
2. Categorize pleasant words on left, unpleasant words on right.
3. Categorize White names or pleasant words on left, Black names or unpleasant words on right (practice).
4. Repetition of Block 3 (test).
5. Categorize Black names on left, White names on right.
6. Categorize Black names or pleasant words on left, White names or unpleasant words on right (practice).
7. Repetition of Block 6 (test).

Reaction times were recorded on the test blocks, allowing a comparison of performance on congruent trials (i.e., Block 4) with performance on incongruent trials (i.e., Block 7). The test blocks consisted of 40 trials, whereas the other blocks consisted of 20 trials. The intertrial interval was 150 ms. A red X appeared in the middle of the screen when an incorrect response was made (e.g., if a participant categorized an item on the left side when it should have been categorized on the right side) and remained there until a correct response was made. A green O appeared for correct responses.

Participants notified the experimenter when they had completed the IAT, and the experimenter retrieved each participant’s reaction times for the congruent and incongruent trials (using the “report” command in the FIAT program). The experimenter then pointed out the discrepancy between performance on congruent and incongruent trials. Indeed, all participants did perform faster on congruent trials than on incongruent trials. The experimenter further emphasized this discrepancy by having participants mark their reaction times on a chart that allowed for a clear and visible comparison of their average performances on the two types of trials. The experimenter suggested that, even measured in milliseconds, such discrepancy performance could have important implications for behavior. The experimenter then exited the room, leaving participants to read a similar written explanation of biased IAT performance that was presented on the computer.

The next screen noted that researchers are often interested in how people feel after certain procedures and asked participants to record their affect using a 32-item affect scale similar to the scales used in Experiments 1 and 2. Each affect item was rated using a scale ranging from 1 (does not apply at all) to 7 (applies very much). Participants then completed a filler task that required them to read a list of 125 words and place each into one of seven categories (food, instruments, clothes, sports, furniture, appliances, or school subjects).

Participants then completed two final tasks. The experimenter informed participants that they would see a list of words presented individually and would have to categorize the words as quickly as possible by pressing one of two keys. First, they categorized items as either living or nonliving by pressing the z key if the item was living (e.g., person’s name, animal) or the 2 key on the number keypad if the item was nonliving (e.g., chair, book). There were two blocks including 30 trials each, yielding a total of 60 trials. The items presented contained words from the previous filler task, new filler words, and names from the IAT. Specifically, each block included 5 Black names and 5 White names, plus 20 filler words presented in a random order that was fixed across participants. Thus, across the two

12 When the possible range of should scores is divided into three levels, 6 participants scored above the lowest possible level.

12 We did not manipulate the order in which critical blocks of IAT trials were completed because our goal was to use the IAT as a means to an end (rather than the goal being to examine IAT effects per se).
blocks, participants saw 20 names and 40 filler words. The intertrial interval was 300 ms.

The living–nonliving task was included simply to familiarize participants with this type of activity and was not expected to yield activity of the BIS. For example, responding whether Tyrone is living or not living should not result in a pause in ongoing behavior nor have any effect on responses, because this is not a context in which a prejudiced response could occur. However, the next task was expected to reveal the operation of cues for control. For this task, participants categorized the same words (presented in reverse order) in terms of whether they liked or disliked each item. Participants were instructed to perform the task as quickly as possible, using the z key for like and the 2 key for dislike. The presentation of Black names in the context of indicating one’s likes and dislikes should constitute the presentation of cues for control among participants who felt guilty in relation to their IAT performance. That is, because these participants had biased IAT performance (constituting a discrepant response) and experienced an aversive outcome (guilt), they were expected to build associations among stimuli related to their prejudiced responses (e.g., the Black names), the aversive outcome, and their prejudiced responses. Thus participants who felt guilty in relation to their IAT performance should pause when these names are presented and be more likely to indicate that they like the names.

Finally, the experimenter probed participants for suspicion by asking them to describe their interpretation of the purpose of the experiment. Specifically, it was important to identify any participants who perceived an intended link between the IAT and the like–dislike task. None of the 48 participants expressed any suspicion regarding the true nature of the experiment. Participants were then debriefed and provided with research credit or payment.

Results and Discussion

IAT Scores

Following Greenwald et al. (1998), we recoded response latencies less than 300 ms to 300 ms and latencies greater than 3,000 ms to 3,000 ms. After eliminating the first two trials (because they tend to include considerable error), we formed congruent and incongruent indices by averaging the appropriate trials (i.e., trials in Block 4 for the congruent index, and trials in Block 7 for the incongruent index). All analyses were performed on log-transformed latency data, although results are reported in milliseconds to ease interpretation.

A paired t test treating congruent and incongruent trials as repeated measures revealed the usual robust IAT effect, in which participants have an easier time responding on congruent ($M = 732.60$ ms) than on incongruent ($M = 1,045.00$ ms) trials, $t(47) = 15.06, p = .001, r = .61$. The congruent index was then subtracted from the incongruent index to yield an overall IAT score, with higher numbers indicating a greater racial bias. Correlational analyses indicated that IAT scores were unrelated to participants’ gender and Should scores ($rs = -.14$ and .17, respectively, $ps > .26$).

Affective Consequences of IAT Performance

Affect indices were formed in the same manner as in Experiments 1 and 2 (i.e., negself, discomfort, positive, down, and negother; alphas ranged from .79 to .90) and predicted in regression analyses using participants’ centered IAT scores, Should scores, and the interaction between these variables. Main effects were entered simultaneously, followed by the interaction. On the basis of previous findings (Monteith et al., 2001), we expected that participants who showed a greater racial bias against Blacks on the IAT would feel greater negself in relation to their performance.

Analyses of the affect indices revealed a significant effect in relation to negself, such that participants with greater IAT biases reported greater negself, $F(1, 45) = 4.10, p < .05, \beta = .29$. The effect of Should scores on negself was not significant, nor was the interaction between IAT and Should scores, $Fs < 1.00$. Because only people who experience negself in relation to their IAT performance theoretically would be expected to have established cues for control, the next steps in the analyses were to examine whether feelings of negself predicted subsequent self-regulatory activity in the form of prospective reflection and the inhibition of racially biased responses.

Relation Between Negself and Pausing

Reaction times for trials on which Black or White names were presented were examined to eliminate outliers, which were defined as scores that were greater than three standard deviations from the mean (outlier rate = 1.13%). All analyses were performed on the basis of log-transformed reaction times.

We first tested the prediction that participants who were experiencing greater negself feelings would experience a brief interruption of ongoing behavior when the Black names were presented in the context of the like–dislike task. Initial examination of these data revealed an unanticipated effect of a third variable that was causing the true effect of negself on reaction times to be masked. Specifically, across all types of words presented during the like–dislike task, participants took substantially longer to press the 2 key on the number keypad (which was used for “dislike” responses) than to press the z key (which was used for “like” responses), $t(46) = 3.10, p = .003$. Response times therefore were being affected by the extent to which the names were serving as a cue for control, as indexed by negself feelings, and by whether participants were providing a “like” or a “dislike” response. To control for this confound, we formed separate reaction time indices for trials on which participants used the 2 key and the z key, standardized each index within its own distribution, and then computed the average of the two indices to reflect overall pausing time for the Black names. (In cases in which participants used the 2 or z key for all responses, the overall pausing time was based only on the one index.) This procedure essentially puts reaction times in relation to both keys on the same scale, thus eliminating the confounding factor. So that reaction times for White names could be used as a covariate in analyses of the pausing data, we computed overall pausing time for the White names in the same way.

We performed a regression analysis using centered negself and Should scores to predict the (standardized) pausing times for the Black names, with pausing times for the White names entered as a covariate. This analysis revealed the anticipated significant relation between negself and pausing for the Black names, $F(1, 43) = 7.52, p < .01, \beta = .34$. The more participants experienced negself feelings in relation to their IAT performance, the more they paused in relation to the Black names. This suggests that the Black
names were serving as cues for control. The covariate effect was also significant, $F(1, 43) = 17.27, p = .001$.\(^{13}\)

We also examined the reaction time data from the living–nonliving task. Recall that the same words were presented in this task as in the like–dislike task but that we did not expect the Black names to serve as cues for control in the context of the living–nonliving task. That is, the presence of the names was expected to interrupt ongoing behavior only in the context of a prejudice-relevant situation. Initial examination of these data indicated that, as in the like–dislike task, participants took significantly longer to press the 2 key on the number keypad (which was used for “nonliving” responses) than the 2 key (which was used for “living” responses), $t(47) = 9.89, p = .001$. We controlled for this confound in the same way as we did for the like–dislike data. The regression analysis revealed that the effect of negself was not significant, $F < 1.00$, and the effect of the covariate was significant, $F(1, 44) = 68.72, p = .001$. The findings thus indicated that the Black names served as cues for control only in the context of a task in which racially biased responses could be generated.

Relation Between Negself and Liking Responses

This experiment allowed us to test whether the presence of stimuli that serve as cues for control helps people to provide racially unbiased responses. The specific expectation was that participants who experienced greater negself in relation to their IAT performance would be more likely to say that they liked the Black names. To test this, we summed participants’ responses to the 10 Black names that were presented in the like–dislike task (scoring each “like” response as 1, which yielded an index that ranged from 0 to 10). This index was predicted in a regression analysis using participants’ negself and Should scores (both of which were centered). This analysis revealed the expected effect of negself on reports of liking the Black names, such that participants who experienced more negself as a consequence of their IAT performance were more likely to report that they liked the Black names, $F(1, 44) = 4.30, p < .05, \beta = .275$. There also was an unexpected significant effect for Should scores, such that participants with more prejudeiced Should scores were less likely to say they liked the Black names, $F(1, 44) = 9.41, p = .004, \beta = -.41$. Thus, although the range of Should scores was restricted, there apparently was enough variability to significantly affect liking responses.\(^{14}\)

Mediation Analysis

Our argument that negself results in pausing in relation to the Black names, which in turn encourages participants to provide “like” responses to these names, can be tested further with a mediational analysis. We have already established two criteria for mediation (see Baron & Kenny, 1986), which are that negself is significantly related to both pausing and liking. Three additional criteria must be met to establish mediation: (a) Pausing and liking must be related, (b) the effect of negself on liking must be significantly reduced when negself and pausing are used simultaneously to predict liking, and (c) the effect of pausing must be significant when negself and pausing are used simultaneously to predict liking. The first of these criteria received some support. Specifically, the partial correlation between pausing for Black names (with pausing for White names controlled) and liking was marginally significant, $r(43) = .27, p = .07$. The second criterion was met in that, when negself and pausing were used simultaneously to predict liking, the effect of negself was reduced, $F(1, 42) = 2.23, p = .14, \beta = .20$. However, the third criterion was not met. That is, the effect of pausing on liking was not significant when both negself and pausing were used to predict liking, $F < 1.00$. Also, a test of the significance of the difference in betas (i.e., $\beta = .27$ when negself is used to predict liking vs. $\beta = .20$ when negself and pausing are both used to predict liking) was not significant according to Wills, DuHamel, and Vaccaro’s (1995) procedure for performing the test. In sum, although negself feelings did predict both pausing (behavioral inhibition) and liking of Black names (nonprejudiced responding), we did not find evidence that pausing served as a mediator of the latter effect.

General Discussion

The successful self-regulation of behavior is a requirement for effective everyday functioning. Fortunately, people apparently can regulate themselves out of a variety of habitual behaviors, from learning to execute motor patterns differently (e.g., when driving to work or typing) to learning to think and feel differently (e.g., when overcoming depressive tendencies or quitting smoking). We assume that people can also learn how to regulate themselves out of the stereotyping habit, though this is no easy task (cf. Devine, 1989). The present research provides a better understanding of just how people can learn to put the brakes on their prejudice and control the influence of processes that otherwise could result in racially biased behavior.

This research focuses on processes of inhibition involved in self-regulation (e.g., Gray, 1987; Patterson & Newman, 1993) and, more important, on the establishment and operation of cues for control. Experiments 1 and 2 manipulated participants’ perceptions of whether they had engaged in prejudiced responses, and they revealed that low-prejudice individuals showed evidence of behavioral inhibition following prejudiced responses. That is, participants showed a brief interruption in ongoing behavior when presented with feedback that they had negative reactions to pictures of Blacks. This type of momentary interruption is a hallmark of BIS activity (e.g., Gray, 1987; Newman et al., 1997; Nichols & Newman, 1986; Patterson et al., 1987). Other participants who received feedback that they had negative reactions to pictures that did not

13 Readers may wish to note that, when we performed a parallel analysis using participants’ IAT scores to predict pausing for the Black names, only the covariate was significant, $F(1, 43) = 13.58, p = .001$. Also, when pausing in relation to the Black names was predicted with IAT scores as well as negself, Should scores, and White pausing times (i.e., the covariate), the effect of negself remained significant, $F(1, 42) = 6.68, p < .02$, and the effect of IAT scores was not significant, $F < 1.00$.

14 As with the pausing results, readers may wish to note that, when we performed a parallel analysis using participants’ IAT scores rather than negself to predict liking of the Black names, only Should scores were associated with a significant effect, $F(1, 44) = 10.30, p < .01$. Also, when liking was predicted with IAT scores as well as negself and Should scores, the effect of negself was marginally significant, $F(1, 43) = 3.63, p = .063$, and the effect of IAT scores was not significant, $F < 1.00$. 

include Blacks showed some evidence of behavioral inhibition, but this reaction was not immediate and enduring (i.e., evidence of inhibition emerged only after repeated failure to reduce negative arousal). We also found that participants experienced aversive consequences in the form of guilty feelings only in relation to their negative reactions to Blacks, and they showed evidence of retrospective reflection (i.e., a preoccupation with their responses) only when they presumably were unable to control their negative arousal in relation to Blacks. Finally, we found that feelings of guilt were significantly related to the behavioral inhibition and retrospective reflection data. Altogether, these results demonstrate that events that one would expect to happen if participants were developing cues for control in relation to their prejudiced responses did occur. The inhibitory and reflective activity observed and the experience of aversive outcomes (i.e., negative affect) facilitates the development of associations between prejudiced responses, negative consequences of such responses, and stimuli that predict the occurrence of prejudiced responses.

Experiments 3 and 4 examine the operation and effects of established cues for control. Our expectation was that, in subsequent situations that afford prejudiced responses, the presence of cues for control should trigger inhibitory processes. Support for this idea was found in Experiment 3. When low-prejudice individuals were presented with Blacks in a stereotypic context (i.e., when already-established cues for control were presented), they experienced a brief interruption of ongoing behavior. Experiment 4 gave participants the opportunity to establish cues for control based on their actual behavior (i.e., not based on a false feedback procedure, as in Experiments 1 and 2). Participants who initially felt guilty about their IAT performance later paused briefly when Black names from the IAT were presented in a context that afforded prejudiced responses, which is consistent with the idea that the Black names had come to constitute cues for control. Moreover, participants who felt guilty about their IAT performance also were less likely to generate racially biased responses. Because the participants had established cues for control, prospective reflection could occur, and racial biases were less likely to be manifested.

These findings go beyond past research that showed that low-prejudice people can sometimes avoid prejudiced responses (Devine, 1989) and that they are especially likely to do so when they have recently caught themselves behaving in prejudiced ways (Monteith, 1993; see also Moskowitz et al., 1999). Specifically, the present research provides support for a process model for understanding how people learn to self-regulate their prejudiced responses through the establishment and operation of cues for control. The development and operation of such cues are crucial for successful self-regulation for at least two reasons. First, the unintentional manner in which stereotypes can be activated and used makes awareness of their possible influence essential for control (cf. Bargh, 1999). When cues for control have been established, they can serve as warning signals that the situation is one in which prejudiced responses may occur and that the brakes need to be applied to ongoing behavior (Allport, 1954). The presence of cues can create awareness of the possibility of biased responding. Second, cognitive resources are often needed to replace stereotypic responses with more egalitarian ones, and cues for control can help to create these resources. That is, the triggering of the BIS in response to cues for control is thought to produce a shift from automatic to controlled processing (Gray, 1987; Newman et al., 1997).

In addition to extending our understanding of how prejudiced responses can be self-regulated, we also believe that our work extends Gray's (1982) model of motivation and learning in an important way. The model typically has been applied to the sort of self-regulation that occurs when tangible negative outcomes occur (e.g., one has to learn to inhibit a particular behavior because it is associated with a loss of money, as in Patterson et al., 1987). The mismatch or discrepancy thus is between engaging in a behavior and having it result in real, negative outcomes. In the present case, the discrepancy is a cognitive one that is based on a conflict between behavior and personal standards, and the negative outcome is purely affective. Our application thus has much more to do with cognition, affect, and—more generally—conscience. Although the present findings provide support for the theoretical analysis that is advanced, an important question is whether people actually establish cues and regulate their responses to these cues in their everyday lives. In the present experiments, procedures were used to make participants' prejudiced responses (or their perceptions of having had such responses) very salient to them. The physiological feedback paradigm in Experiments 1 and 2 repeated salient participants' supposed negative reactions to Blacks. Such salient and repeated feedback is unlikely to occur in the real world. The IAT procedure from Experiment 4 made participants' biases salient to them in a very palpable and basic (i.e., Black—unpleasant, White—pleasant) way. The real world often does not provide such unambiguous information about racial biases. Do people actually develop cues for control under real-world conditions?

We are investigating this question in ongoing research that uses an interview methodology. Participants are asked whether they have had a racial experience in which they did something related to Blacks that they then thought they should not have done, either because of their own reactions or because of others' reactions. If participants can think of a relevant experience, they describe it, and then they are asked whether they had ever thought about the experience again. These interview data can be examined in various ways to determine, among other things, whether cues have been established, and we are finding evidence that they are. Below is a quote from an interview that helps to underscore this real-world evidence.

The participant described the initial experience as follows:

Me and my roommate got approached several times last year in the parking lot by some homeless people and they were Black. After the first 2 or 3 times they asked for money, whenever we would see a group of Black people near there we would automatically assume that we better go the other way because they were going to ask for money. It made me feel bad because I didn’t even give them a chance. Maybe they wanted directions to go somewhere, or wanted to know where to get some coffee or something. We just automatically assumed that they wanted money. That was being judgmental, and I really didn’t feel too good about that.

In response to a question about whether the participant had ever thought about the experience again, the participant said:

This summer my girlfriend and I were looking at the horses downtown, and we were looking at one and this Black guy started walking toward us. Of course I figured he was probably a homeless guy, and
I immediately thought, here comes some homeless guy—he’s going to ask us for money. But from my past experience I had in the parking lot with my roommate, I had to stop and think to myself, “Maybe he’s not homeless, maybe he’s not going to ask me for money. He might not say anything to me.” I stopped and I thought about the past experience and it made me change my decision to something I probably wouldn’t have made.

Consistent with the internal drive to control prejudice that this participant appears to have, his score on a measure of prejudice (the ATB) suggested that he does not consciously hold negative attitudes toward Blacks. As suggested in the introduction, high-prejudice individuals may also develop cues for control, but on the basis of external rather than internal motives to control prejudice (Dunton & Fazio, 1997; Plant & Devine, 1998). The quote below, from a participant who had a high score on the measure of prejudice, illustrates the operation of cues for control owing to external motivation.

The participant described the initial experience as follows:

My roommate’s Black and sometimes when we’re watching shows they kinda like make the Blacks look trashy, you know like on Jerry Springer . . . . I was laughing at it but he wasn’t really and it kind of automatically made me feel like I had done something wrong so I felt bad . . . I didn’t want him to think, “Well he looks like some kind of racist.”

In response to a question about whether the participant had ever thought about the experience again, the participant said:

If something on the TV comes up that’s like shady you know it’s like I think about it . . . you know I think about it to make sure that it doesn’t happen again in case he actually was mad about it. I wouldn’t laugh out loud if I thought maybe it would be offensive to someone else. I’m just a little more careful now.

Future research is needed that provides experimental tests of self-regulation among high-prejudice persons when external motivations encourage them to respond consistently with other people’s low-prejudice standards. Research is also needed that addresses circumstances under which such control of prejudice ultimately has negative versus positive outcomes. For example, Plant and Devine (2001) recently found that high-prejudice individuals who controlled their prejudice later showed evidence of attitudinal and behavioral backlash.

Many other questions related to the operation of cues for control also need to be addressed in future research. First, our theoretical analysis suggests that the effect of negself resulting from an initial discrepancy experience on subsequent prejudiced responding should be mediated by pausing (behavioral inhibition). Because we did not find evidence of mediation (see Experiment 4), future work will need to better examine this issue (e.g., a larger sample size may be needed).

Second, it is important to examine factors that affect the likelihood that environmental stimuli will be immediately recognized as cues for control (i.e., as predictors of possible prejudiced responses). We suspect that this depends on cue strength, which should be determined by the magnitude of aversive consequences and the extent of retrospective reflection (i.e., more reflection results in stronger associations among prejudiced responses, aversive consequences, and stimuli that predict those consequences). We also suspect that repetitions of the pairings of aversive consequences with unwanted responses and associated stimuli are an important factor. In the experiments reported herein in which cues were established, participants had repeated failure experiences that could contribute to the establishing of cues for control. Theoretically, repeated experiences are not necessary for the establishment of cues for control, but repetition no doubt increases the likelihood that cues will be established. In addition to cue strength, availability of cognitive resources may prove important. Some cues may be strong enough to automatically signal the need for careful processing (which is consistent with Gray’s, 1987, and Newman et al.’s, 1997, theorizing). With weaker cues, greater processing resources may be necessary before the cues trigger BIS activation. Presumably, all other things being equal, the greater the similarity is between the environmental stimuli that were tagged in the initial situation and the environmental stimuli present in the subsequent situation, the more likely cues are to be effective.

Third, future research needs to examine issues of generalization. One issue concerns how similar the stimuli tagged in the initial situation need to be to stimuli present in the subsequent situation. For example, we found that the historically Black names from the IAT later served as cues for control when participants were performing the like–dislike task. Would the same effects emerge if one set of Black names were used in the IAT and a different set were used during the like–dislike task? We suspect that, all other things being equal, there is a positive linear relation between the degree of similarity and the power of stimuli to serve as cues.

Another generalization issue concerns the contexts in which stimuli serve as cues. At minimum, the findings from Experiment 4 suggest that stimuli serve as cues only in contexts that afford biased responding. Specifically, participants paused in relation to the Black names when performing the like–dislike task but not when performing the living–nonliving task. The names served as cues only in a context relevant to producing unwanted responses.

Fourth, future research is also needed to examine more long-term consequences associated with the operation of cues for control. The present research measured biased responses (Experiment 4) and showed that their likelihood is reduced in the presence of cues for control. However, can the initial activation of stereotypes or implicit prejudices become less likely through self-regulation? Theoretically, one should be able to automatize the self-regulation process, just as the skill acquisition literature suggests (Shiffrin & Schneider, 1977; Smith, 1990; Smith, Branscombe, & Bormann, 1988). With such constant self-regulation, one would naturally expect the atrophy of associations that encourage the to-be-avoided responses. Although this makes theoretical sense, we are unsure whether it actually happens with much frequency in the case of stereotypes. Many responses that people seek to deautomatize are ones that they execute dozens of times throughout a given day, providing great opportunity for practice. Kawakami et al. (2000) found that stereotypes are less likely to be automatically activated with considerable practice negating them, but everyday circumstances may not typically provide people with the chance for such practice. Also, everyday encounters with stereotypic portrayals (e.g., media influence) may discourage the atrophy of stereotypic associations. Continual self-regulation thus may often be necessary.
Conclusion

Bodenhausen and Macrae (1998) provided an interesting discussion of the evolutionary advantages afforded by the ability to inhibit thoughts and behaviors that disrupt social harmony, such as those that result from stereotyping, and the consequent development of brain structures that enable people to exercise control (see Bjorklund & Harnishfeger, 1995). Just as being able to activate stereotypes automatically can be adaptive because it simplifies the complexities of the social world (Allport, 1954), being able to control and change such automatic processes is highly adaptive when stereotype use conflicts with one’s standards. Our analysis holds that the development and operation of cues for control are critical for such control and change. Cues are essential for being able to—in Allport’s (1954) words—put the brakes on one’s prejudices.

References

Gray, J. A. (1982). The neuropsychology of anxiety: An enquiry into the
function of the septohippocampal system. New York: Oxford University Press.


Received January 9, 2001
Revision received March 15, 2002
Accepted March 27, 2002

New Editors Appointed, 2004–2009

The Publications and Communications Board of the American Psychological Association announces the appointment of five new editors for 6-year terms beginning in 2004. Except where noted, as of January 1, 2003, manuscripts should be directed to the following individuals:

- For *Psychology and Aging* (http://www.apa.org/journals/pag.html), submit manuscripts to Rose T. Zacks, PhD, Department of Psychology, Michigan State University, East Lansing, MI 48824-1117.
- For *Psychological Assessment* (http://www.apa.org/journals/pas.html), submit manuscripts to Milton E. Strauss, PhD, Department of Psychology, Case Western Reserve University, Cleveland, OH 44106-7123.
- For *Journal of Family Psychology* (http://www.apa.org/journals/fam.html), submit manuscripts to Anne Kazak, PhD, ABPP, Oncology Psychosocial Research, The Children’s Hospital of Philadelphia, Room 1486 (Market Street), 34th and Civic Center Boulevard, Philadelphia, PA 19104. For overnight couriers: Room 1486, 3555 Market Street, Philadelphia, PA 19104.
- For *Journal of Personality and Social Psychology: Personality Processes and Individual Differences* section (http://www.apa.org/journals/ppp.html), submit manuscripts to Charles S. Carver, PhD, Department of Psychology, University of Miami, P.O. Box 248185, Coral Gables, FL 33124-2070.

Electronic submission: As of January 1, 2003, authors will be expected to submit manuscripts electronically through the journal’s Manuscript Submission Portal (see the Web site listed above with each journal title). Authors who are unable to do so should correspond with the editor’s office about alternatives.

Manuscript submission patterns make the precise date of completion of the 2003 volumes uncertain. Current editors Leah L. Light, PhD, Stephen N. Haynes, PhD, Ross D. Parke, PhD, Mark E. Bouton, PhD, and Ed Diener, PhD, respectively, will receive and consider manuscripts through December 31, 2002. Should 2003 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2004 volumes.