Journal of Experimental Psychology: General
2003, Vol. 132, No. 3, 455–469

0096-3445/03/$12.00 DOI: 10.1037/0096-3445.132.3.455

Contextual Variations in Implicit Evaluation

Jason P. Mitchell
Harvard University

Brian A. Nosek
University of Virginia

Mahzarin R. Banaji
Harvard University

In the present research, the authors examined contextual variations in automatic attitudes. Using 2 measures of automatic attitudes, the authors demonstrated that evaluative responses differ qualitatively as perceivers focus on different aspects of a target's social group membership (e.g., race or gender). Contextual variations in automatic attitudes were obtained when the manipulation involved overt categorization (Experiments 1–3) as well as more subtle contextual cues, such as category distinctiveness (Experiments 4–5). Furthermore, participants were shown to be unable to predict such contextual influences on automatic attitudes (Experiment 3). Taken together, these experiments support the idea of automatic attitudes being continuous, online constructions that are inherently flexible and contextually appropriate, despite being outside conscious control.

From describing the mechanics of light as waves to the lock-and-key nature of enzyme specificity, metaphors have served to create and communicate ideas about complex systems. To make sense of the interactions between brain, mind, and environment, psychologists have routinely used this tool of language and imagination to develop new ways of representing such interactions. Commenting on the dominant metaphor for construing the nature of mental representation, Smith (1996) observed that until the 1980s, mental representations were cast as things, capable of being stored or retrieved, as one might locate a can of beans in the pantry. More recently, in an effort to change that way of thinking, psychologists have instead cast mental representations as distributed patterns of activation (McClelland & Rumelhart, 1985).

From these two metaphors of mental representation—as things or as idealized neural networks—several differences in the assumed character of mental representations follow, including their structure as well as the nature of learning and retrieval. Traditional models have assumed that "learning involves the explicit construction of new representations," that "representations are passive and inert," and that "use of representations inherently involves two separate stages: activation or retrieval from storage, followed by use" (Smith & DeCoster, 1998, p. 21). Instead, distributed models (McClelland & Rumelhart, 1998; McClelland, Rumelhart, & Hinton, 1988) make the assumption that "representations are not static entities that are 'stored' inertly until retrieved by a search process and used. Instead, a single mechanism, the flow of activation along connections between units, accounts for both storage and processing of information" (Smith & DeCoster, 1998, p. 22). As Smith (1996) advised earlier,

> it is better to think of a representation as being *re-created* or *evoked* than as being *searched for*. . . . The re-creation will often be imperfect and subject to influence from the person's other knowledge . . . but this characteristic is typical of actual human memory performance. (p. 896; italics in original)

In viewing mental representations as dynamically reconstructed, not statically retrieved, distributed models have highlighted previously unexplored theoretical questions.

Our focus here is a particular mental representation, *attitude* or *evaluation*, and we attend specifically to those attitudes that appear to operate relatively outside conscious control. When attitudes are considered not as evaluative things that are retrieved but rather as patterns that are reconstructed within the parameters of a particular context, their dynamic and variable nature becomes highlighted. When the variability in attitude expression is shown to be a function of characteristics such as frames of reference or a particular orientation shaped by past and recent experience, it does more than contribute a new empirical finding: It provides a picture of the fundamental features that make up the very nature of such attitudes (Blair, 2002).

Jason P. Mitchell and Mahzarin R. Banaji, Department of Psychology, Harvard University; Brian A. Nosek, Department of Psychology, University of Virginia.

This research was supported by National Science Foundation and National Research Service Award predoctoral fellowships to Jason P. Mitchell, Grant MH-57672 from the National Institute of Mental Health, and Grants SBR-9422241 and SBR-9709924 from the National Science Foundation to Mahzarin R. Banaji. Portions of this research were presented at the 1999 annual meeting of the Midwestern Psychological Association in Chicago, Illinois, and at the 1998 annual meeting of the American Psychological Society in Washington, DC. We thank Rainer Banse, Richard Hackman, and Eliot Smith for comments on a previous version of this article; Daniel Schacter for generous use of the lab space used to complete Experiments 4 and 5; and Ethan Haymovitz for help with data collection.

Correspondence concerning this article should be addressed to Jason P. Mitchell or Mahzarin R. Banaji, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, Massachusetts 02138. E-mail: jmitchel@wjh.harvard.edu or banaji@wjh.harvard.edu

A view of attitudes as contextually variable is not altogether easy to assimilate, because attitudes are assumed to be inherently stable in both lay and scientific thinking. In his article on conviction, Abelson (1988) included longevity ("How long have you held your views?") as a component of strongly held convictions, with items such as "I can't imagine ever changing my mind" loading well on a conviction factor that captures emotional commitment. Consistent with this notion, Gross and Ellsworth (2003) reported that a majority of Americans indicate that an attitude of strong conviction, specifically, their attitude toward the death penalty, has not changed over time. They pointed out that "except for radical conversion experiences, people are rarely aware that their attitudes have changed: they report their current attitudes as attitudes they have held as long as they can remember" (p. 12).

An alternative view of attitudes being sensitive to context has been pointed out as well. An early experiment on group perception—and a remarkable interpretation offered therein—shows the power of context in shaping the expressed attitude. In 1940, Solomon Asch reported over a dozen experiments demonstrating "the ways in which judgments are affected by knowledge of and beliefs about the standards of groups and individuals" (p. 433). Among the best known of those experiments is one in which Asch (1940) presented respondents with one of two rank-ordered lists of 10 professions, ostensibly so ranked by 500 of their peers on qualities such as social usefulness, idealism, and intelligence. The simple manipulation consisted of placing the profession of politician either at the top or at the bottom of the peer-ranked list. This variation produced a large effect on participants' own ratings: Those who believed politicians to be ranked first by their peers also ranked them higher in their own assessments compared with those who believed politicians to be ranked last by their peers.

It is important to note that Asch (1940) did not interpret the finding as reflecting mere conformity with peer opinion or as revealing a shift in attitude toward politicians as a group because of peer opinion. Rather, the explanation was psychologically far more interesting (Lord & Lepper, 1999). Participants in the "politician-first" condition, Asch noted, had temporarily represented the category by imagining its more admirable exemplars, whereas those in the "politician-last" condition had represented the same category by imagining exemplars from the bottom of that barrel. According to Asch (1940), "the group standards have not worked *directly* . . . by virtue of their suggestiveness or prestige, but their action is confined to the definition of the object of the judgment. The standards changed the stimulus-situation" (p. 457; italics in original).

More recently, Schwarz and colleagues have conducted extensive research to show the sensitivity of explicitly stated attitudes to contextual variables. They demonstrated that self-reported attitudes are influenced by a variety of factors, including interpretation of question meaning, constraints on responding (e.g., whether a "no opinion" option is available), question order, comprehension of questions, references to social norms, and even question formatting (see Schwarz, Groves, & Schuman, 1998, for a review).

From such work, social psychologists know that contextual factors can systematically shift self-reported attitudes and beliefs. But it has been suggested that the pliability of attitudes in laboratory studies ought to be taken as just that: attitude change that occurs merely in laboratory studies because such studies use attitude objects that have little bearing on the strong attitudes that are held with conviction outside the lab. According to Abelson (1988), this difference in conviction leads laboratory studies to reveal attitude change but field studies and everyday experience to suggest that attitudes are long lasting and unchanging.

Moreover, the belief that attitudes are stable is especially conspicuous when considering unconsciously held or implicit attitudes. The assumption is that unconscious attitudes, that is, those that lie outside conscious awareness or control, are invariant. By the very fact that they are dissociated from consciousness, unconscious representations have been thought to be less malleable, less sensitive to intervention, and less likely to change as a result of contextual variation (see Banaji, in press). Such assumptions may be quite reasonable, because the very idea of nonconscious mental representation signals an imperviousness to change, at least through the efforts of conscious will. Indeed, influential models of behavioral control have suggested that as a general class of phenomena, automatic behaviors (including automatic evaluations) are inevitably elicited in the presence of appropriate triggering stimuli and can only be reshaped by secondary control processes that require attentional resources (Norman & Shallice, 1986; Wegner & Bargh, 1998).

Contextual Effects on Automatic Attitudes

In a recent review, Blair (2002) integrated a number of studies that suggest that features present in the evaluation context can shape even automatic attitudes. For example, one study using the Implicit Association Test (IAT; see below for a description of the technique) demonstrated that exposure to positive African American exemplars resulted in participants producing evaluations of that group that were not as negative as those produced in a control condition (Dasgupta & Greenwald, 2001), contrary to expectations that automatic attitudes are unbending and invariant. Likewise, the very presence of an African American experimenter can influence participants to produce evaluations that are more positive toward that group than the presence of a European American experimenter (Lowery, Hardin, & Sinclair, 2001). Such findings have raised the possibility that even those attitudes that operate relatively outside of conscious control can fluctuate in evaluation as a function of the context in which they are elicited.

In demonstrating that automatic attitudes toward African Americans were less negative after exposure to positive exemplars, the research reviewed above (Dasgupta & Greenwald, 2001; Lowery et al., 2001) is consistent with two models of "change" in automatic attitudes (see Blair, 2002, for an extended discussion of these different models). One possibility suggests that participants in these experiments had a stable, negative attitude toward African Americans but that exposure to positive exemplars temporarily shifted such attitudes in the positive direction. In this stable-but-malleable view, an encounter with positive members of a disliked group can immediately produce less negative automatic attitudes, but the impact of such exemplars will decay over time. Indeed, although participants' race attitudes were initially more moderate after the participants were exposed to positive African American exemplars, they were characteristically negative when measured after a 24-hr delay (Dasgupta & Greenwald, 2001). To the extent that participants can be thought to have stable-but-malleable race attitudes, this later negativity simply reflects the inevitable atten-

uation of the positive exemplar manipulation over time and the return to one's stable, baseline attitude toward African Americans.

In contrast, another view suggests the more radical notion that no such things as stable, precompiled attitudes exist in the first place and that what appears to be attitude change is, in fact, attitude construction. Consistent with theories of mental representations being inherently constructed rather than retrieved, such a possibility suggests that automatic attitudes are built from the bottom up each time they are elicited. Such constructed attitudes necessarily arise as part of a wider situational context and incorporate information present in the environment. For example, when automatic attitudes toward African Americans are assessed after participants are exposed to positive African American individuals, their expressed attitudes may incorporate some of the positivity associated with those exemplars of the group. To push the metaphor of attitudes as constructions further, one might imagine that the very material out of which the attitude is constructed can contain an admixture of positivity or negativity picked up from the environment. If one's attitude toward the same object is later measured in a different context, it will incorporate different informational material, potentially resulting in the construction of an attitude qualitatively distinct from the one observed in an earlier context. In this view, perceivers' negativity toward African Americans 24 hr after encountering positive exemplars does not represent a return to some kind of evaluative baseline but rather the elicitation of a distinct attitude within a different context.

The existing literature on automatic attitude change does not suggest whether automatic attitudes are better viewed as stable-but-malleable representations or as contextually bound, online constructions (Blair, 2002). These two views can be supported, respectively, by theories of automaticity (Wegner & Bargh, 1998) on the one hand and theories of the reconstructive nature of mental representations on the other (Smith, 1996). The current research attempts to arbitrate between these two possibilities. In the experiments reported here, we attempt to examine such contextual effects on automatic attitudes directly with an eye toward resolving the theoretical ambiguity surrounding the processes that underlie such attitude fluctuations. In five experiments, we examined the possibility that changes in context can provoke rapid, reversible shifts in automatic attitudes within an individual perceiver. To the extent that automatic attitudes are constructed anew each time they are elicited, abrupt fluctuations in the valence of an attitude should be observed across changes in the context in which the attitudes are evoked. In contrast, more stable evaluative representations would be expected to resist such rapid alternations and instead prove relatively intransigent in the face of quickly changing contexts.

IAT

The IAT (Greenwald, McGhee, & Schwartz, 1998) is capable of measuring differences in association between target concepts (e.g., *Black* or *White*) and evaluative attributes (e.g., *good* or *bad*). The IAT operates on the principle that it should be easier to make the same behavioral response to concepts that are associated than to concepts that are not associated. Like the evaluative priming task (Bargh, Chaiken, Govender, & Pratto, 1992; Fazio, Jackson, Dunton, & Williams, 1995; Fazio, Sanbonmatsu, Powell, & Kardes, 1986), the IAT involves the following assumptions: (a) that strength of evaluative association can be measured, (b) that the extent to which concepts share evaluative meaning (independent of semantic meaning) is revealed in the ease with which they can be mentally paired, (c) that one way to measure the strength of evaluative association is to measure the speed of concept-plus-evaluation pairs, and (d) that the strength of evaluative association as measured under conditions of speeded responding is a measure of automatic attitude (Banaji, 2001).

More specifically, the IAT relies on a response latency indicator obtained in the process of pairing an attitude concept (e.g., a social group such as *old–young*) with an evaluative attribute (e.g., *good–bad*) or specific attributes that may not be purely evaluative (e.g., *self–other*, *home–career*, *science–arts*). In computerized versions of the task, the pairing is achieved by assigning a keyboard key (e.g., a left key) to be pressed in response to items from the two linked categories, such as *old* + *bad*, while another key (e.g., the right key) is used for the other pair, in this example, *young* + *good*. The differential speed required to complete these two types of pairings, that is, the relative ease of pairing *old* + *good* and *old* + *bad* in the context of *young* is interpreted as a measure of the strength of implicit evaluation (i.e., attitude). The IAT effect is a difference score reflecting a relative attitude that shows both the direction (positive vs. negative) of implicit attitude as well as the magnitude of the attitude. Besides traditional tests of significance, this measure has typically been reported with an additional test of effect size, many instances of which have demonstrated that the IAT effect is a large one (see Greenwald & Nosek, 2001; Nosek, Banaji, & Greenwald, 2002).

The Current Research

In Experiment 1, we demonstrate that the automatic attitudes toward well-known Black athletes and White politicians can vary as a function of categorization by race or occupation. Contrary to previous suggestions that the IAT effect is wholly produced by the label that identifies the category, Experiment 2 provides evidence that the IAT is sensitive both to these kinds of shifts in categorization as well as to the identity of the exemplars composing a group. Having provided initial evidence of rapid automatic attitude change, we go on to examine whether perceivers are able to anticipate the effects of contextual changes in producing attitudinal shifts. In Experiment 3, we replicate the contextual effects on automatic attitudes and further demonstrate that rapid automatic attitude shifts occur even under conditions where participants do not predict such change. In this way, Experiment 3 underscores the implicit operation of shifts in automatic attitudes by suggesting one reason why perceivers may come to believe that their attitudes are not susceptible to variation, thus creating the illusion of attitude stability. Finally, in Experiments 4 and 5, we examine more dramatic changes in automatic attitude while removing a possible confound in the procedures of the first three experiments. Using a variation of the IAT for measuring automatic evaluations, the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), we reveal even sharper dissociations in evaluation than previously demonstrated.

Multiply categorizable social targets, that is, those that belong to two or more groups simultaneously, provide an opportunity to observe such qualitative shifts in evaluation. Many researchers have been led by the evidence to suggest that evaluation and

judgment are inevitable consequences of categorizing a person as a member of a social group (Brewer, 1988; Devine, 1989; Fiske & Neuberg, 1990). The question posed in these five experiments concerns the variation in automatic attitudes that stems from variation in contexts that may highlight membership in one or another category. At the conscious level, a social target is clearly capable of activating multiple, mutually contradictory evaluations, depending on the features of the target that form the basis of judgment. For example, one may lament the fact that Charlton Heston is president of the National Rifle Association but admit that he was great in *Planet of the Apes*. Bottom-up, attitudes-as-constructed views predict that automatic attitudes will vary in step with manipulations that highlight one or another feature of a target.

We apply two indirect measures of evaluation, the IAT and the GNAT (detailed in the introduction to Experiment 4). Across five experiments, we use these procedures to converge on an understanding of automatic attitudes under differing contextual conditions.

Experiment 1: Context-Driven Shifts in Automatic Attitudes

Overview

In Experiment 1, we examined whether a particular exemplar can elicit qualitatively divergent automatic evaluations as a function of contexts that highlight different superordinate categories into which the exemplars fit. Participants in Experiment 1 completed two tasks that manipulated the categorization frame used to classify exemplars. In the occupation categorization task, liked Black athletes and disliked White politicians were categorized on the basis of occupation, using the category labels *athlete* and *politician*. In the race categorization task, the same targets were categorized on the basis of race, using the category labels *Black* and *White*.

If automatic attitudes are indeed constructed from the bottom up, transient changes in the salience of different exemplar features should alter the automatic evaluation of those exemplars. When a multiply categorizable target belongs to two groups that are typically associated with discrepant evaluations, cues that highlight membership in one or the other of those groups determine which evaluation is expressed. For instance, Michael Jordan is simultaneously a famous athlete (positive evaluation) and a Black man (negative evaluation), and the automatic evaluation he elicits may depend on whether he is encountered on a basketball court or elsewhere. As a result, we expected automatic evaluations of consciously liked Black athletes and consciously disliked White politicians to differ as a function of whether targets were categorized according to race or occupation.

Method

Participants

A total of 91 volunteers at Yale University either were paid $10 or received partial credit in an introductory psychology course in exchange for participation. Results from 2 participants were excluded from analysis because of a computer malfunction that erased some of the critical data, and 7 participants were excluded because of an excess number of fast responses (i.e., they responded to more than 10% of trials in under 300 ms) on the implicit measures (see Greenwald, Nosek, & Banaji, 2003), leaving a total of 82 participants for the analysis.

Stimuli

For the race and occupation categories, 3 liked Black athletes and 3 disliked White politicians were selected for each participant. In separate blocks, participants rated 19 well-known male athletes, 13 of whom were Black, and 19 male politicians, 14 of whom were White. Participants considered each person and rated him using a 9-point Likert scale anchored by 1 = *dislike strongly*, 5 = *neither like nor dislike*, and 9 = *like strongly*. Participants were instructed to circle the word *unfamiliar* if they did not recognize the name. The 3 Black athletes that a participant rated highest (most liked) and the 3 White politicians that participant rated lowest (least liked) were used as stimuli in subsequent IAT tasks. As such, each participant received an individually tailored list of Black athletes and White politicians on the tasks to measure automatic attitude. Also, items representing the evaluative categories good (e.g., *caress*) and bad (e.g., *agony*) were taken from Greenwald et al. (1998), who normed these words for use in the original demonstration of the IAT.

In the experiments reported in this article, the strength of evaluation of race and occupation concepts should be observed through a task in which stimuli representing these concepts are paired with evaluative terms. A strong association of White with good and Black with bad should lead to the more rapid classification of these items when they are paired with one another (by being assigned to the same computer key for responses) compared with the opposite pairing of White with bad and Black with good. In over 500,000 instances of the race IAT completed via the Internet (http://buster.cs.yale.edu/implicit/), we have observed a strong overall association of Black with bad and White with good (Nosek et al., 2002).

Apparatus and Program

Presentation of experimental stimuli was controlled by IBM (80486 processor) desktop computers running Inquisit software (Version 1.00; Draine, 1997). Participants were instructed to give responses indicating the correct answer was on the left with their left forefinger (using the *A* key) and responses indicating the correct answer was on the right with their right forefinger (using the *5* key on the numeric keypad).

Stimuli were presented sequentially at the center of a computer screen. Response time was recorded from the onset of a target to its correct classification. Correct responses terminated a trial and initiated the subsequent trial following a 150-ms intertrial interval. Categorization labels were positioned to the left and right of the target stimuli to remind participants of the key with which targets were to be classified. If a target was incorrectly classified, a red *X* appeared below the target stimulus, indicating an error, and the program paused until the participant responded correctly.

Procedure

Participants first rated athletes and politicians for conscious expressions of liking. After the rating task, they engaged in race and occupation categorization IATs, in counterbalanced order. In the occupation categorization task, Black athletes and White politicians were classified using the labels *athlete* and *politician*. In the race categorization task, those same targets were categorized as *Black* or *White*. The task followed the basic procedure outlined by Greenwald et al. (1998; also available for demonstration at http://implicit.harvard.edu/), with 40 response trials in each of the critical conditions.

Design

In Experiment 1, we used a 2 (block: Black or athlete + good, Black or athlete + bad) × 2 (categorization task: race, occupation) within-subject

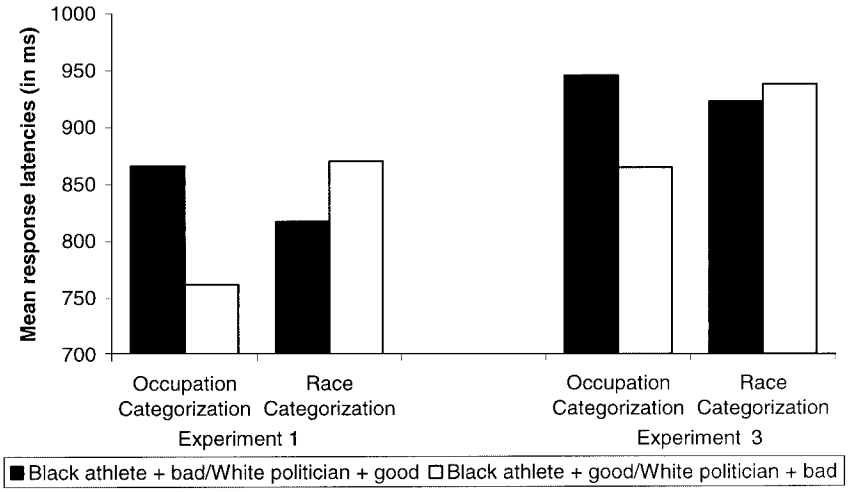

Figure 1. Response latencies in occupation and race categorization tasks by evaluative pairing in Experiment 1 (left panel) and Experiment 3 (right panel).

design. Although not of theoretical interest, Name Rating Order (politician names first, athlete names first), Task Order (race categorization task first, occupation categorization task first), and Block Order (Black or athlete + bad first, Black or athlete + good first) were included as between-subject counterbalancing factors. None of the counterbalancing factors interacted with the primary comparison between race and occupation tasks, all Fs < 1.33, all ps $> .25$; accordingly, all results are reported collapsed across these factors.

Results and Discussion

Data Preparation

Data were prepared following the IAT scoring algorithm recommended by Greenwald et al. (2003). In brief, (a) trial response latencies less than 400 ms or greater than 10,000 ms were eliminated (of 31,640 total trials, 9, or .028%, were discarded), (b) participants whose response times were less than 300 ms on more than 10% of the trials were excluded, (c) all response latencies were included in the analysis, and (d) a difference score (IAT D) was calculated between the two critical blocks of trials (e.g., [athlete + bad and politician + good] – [athlete + good and politician + bad]) and divided by the standard deviation of the latencies across both blocks.[1] The resulting score reflected the IAT D effect: Positive values indicated an automatic preference for White politicians over Black athletes in both the race-salient and occupation-salient tasks (see Greenwald et al., 2003, for additional details about the scoring algorithm). For ease of interpretation, the figures present mean response latencies for each of the critical blocks before calculation of the IAT D effect.

Automatic Attitude Dissociation as a Function of Race and Occupation

We hypothesized that automatic attitudes elicited by Black athletes and White politicians would differ as a function of the way in which targets were categorized. The left panel of Figure 1 presents mean response latencies as a function of IAT block and categorization task. When the categorization task emphasized occupation (athletes and politicians), Black athletes were preferred to White politicians, IAT D $= -0.29$, $SD = 0.39$, $t(81) = -6.7$, $p < .0001$, $d = -0.74$. This is not surprising, because the exemplars consisted of 3 liked Black athletes and 3 disliked White politicians, as rated by each participant. However, when the categorization task emphasized race (Black and White), White politicians were preferred to Black athletes, IAT D $= 0.13$, $SD = 0.43$, $t(81) = 2.8$, $p = .006$, $d = 0.31$. This finding is substantially more surprising, because negativity toward the category Black relative to White was observed despite the self-reported positivity of Black exemplars and negativity of White exemplars. A t test comparing the automatic attitudes elicited between the occupation-salient and race-salient conditions confirmed that manipulating the categorization frame did indeed influence the elicited automatic attitude, $t(81) = 7.18$, $p < .0001$, $d = 0.80$.

Experiment 1 demonstrated that social objects evoked different automatic attitudes as a function of the context in which they were encountered. When highly regarded Black athletes such as Michael Jordan were categorized by occupation, positive automatic attitudes were elicited, in line with consciously reported attitudes of liking. However, when the exemplars were categorized by race, the elicited attitude was qualitatively different from the one observed under occupation categorization.

One possible interpretation of these results would suggest that the IAT merely measures attitudes toward the category labels and that the exemplars composing the categories do not contribute to the elicited automatic attitude. That is, by manipulating the labels used to categorize targets on the IAT, we might not have measured automatic attitudes toward multiply categorizable targets but rather attitudes toward two different attitude objects, that is, athletes and politicians in one task and racial groups in another. As an initial way of addressing this concern (we return to the issue fully in

[1] Because of a procedural variation in the practice blocks, only the 40 trials in the two critical blocks were used for analysis in Experiment 1. In Experiments 2 and 3, we used trials from both practice and critical blocks, following the recommendation of Greenwald et al. (2003).

Experiments 4 and 5), we conducted a follow-up data collection that demonstrated that the IAT is, in part, sensitive to the identity of the exemplars used in measuring automatic attitudes.

Experiment 2: The Role of Exemplars in Automatic Attitudes

In Experiment 2, the categorization task consistently measured attitudes toward a single social dimension (race) while varying the particular exemplars representing the groups Black and White. Participants performed two identical race categorization tasks, which differed only in the stimulus set used: One race categorization task included liked Black and disliked White targets, whereas the other included disliked Black and liked White targets. The main issue of interest is the question of contextual shift, but this experiment also allows a test of a recurring issue in research on implicit attitudes that uses the IAT: Is the effect solely a function of the category labels, as some believe (De Houwer, 2001; Fazio & Olson, 2003), or do the exemplars that represent the category contribute to the attitude that is elicited? If positive exemplars representing the category Black American produce a more positive attitude toward this category than what is usually obtained, the data would challenge the assertion that the IAT attitude effect is solely driven by the category labels.

Method

Participants

A total of 58 participants received partial credit in an introductory psychology course at Yale University in exchange for participation. Data from 4 participants were excluded from analysis because of an excessive number of fast responses.

Stimuli

To create groups of race exemplars that differed in conscious liking, we asked each participant to indicate 3 liked and 3 disliked people from each of two lists of entertainers (musicians and actors), athletes, and politicians or leaders. One list consisted of 45 names of Black Americans and the other consisted of 57 names of White Americans. Participants were also encouraged to generate other names if the lists did not suffice. In addition, Experiment 2 used the same evaluative words as in Experiment 1.

Procedure

Participants were first asked to select liked and disliked Black targets and liked and disliked White targets (3 in each category) from the two lists of names. After this, participants completed two race categorization IAT measures. These two measures were identical to the race categorization task in Experiment 1 and to each other, except for the exemplars presented in each: One IAT included the 3 liked Black and 3 disliked White self-selected names (the liked-Black task), whereas the other IAT included the 3 disliked Black and 3 liked White self-selected names (the disliked-Black task). In both tasks, participants categorized names using the same category labels, Black and White.

Design

In Experiment 2, we used a 2 (block: Black + good, Black + bad) × 2 (exemplar set: liked Blacks and disliked Whites, disliked Blacks and liked Whites) design. Both factors were manipulated within-subject. In addition, Exemplar Set Order (disliked Blacks and liked Whites first, liked Blacks and disliked Whites first) and Block Order (Black + bad first, Black + good first) were included as counterbalancing factors. Neither counterbalancing factor interacted with the primary comparison of interest, Fs < 1.84, ps $> .18$; results are reported collapsed across these counterbalancing factors.

Results and Discussion

Of primary interest was the question of whether different automatic race attitudes would be elicited in response to differing exemplars of social groups. If so, we should observe a less strongly negative automatic attitude toward liked Black exemplars relative to disliked White exemplars, marked by comparable reaction times within Black + good and Black + bad blocks. On the other hand, if automatic race attitudes are not sensitive to exemplars of social groups, we should observe equally negative race attitudes toward both sets of exemplars.

As in Experiment 1, data were prepared following the IAT scoring algorithm recommended by Greenwald et al. (2003). Of critical interest, the IAT D effects were compared for the race categorization task as a function of whether the stimulus set consisted of disliked Black and liked White or liked Black and disliked White individuals. As expected, participants showed a strong preference for White compared with Black targets when those categories were represented by disliked Black and liked White targets, mean IAT D = 0.44, $SD = 0.27$, $t(53) = 12.22$, $p < .0001$, $d = 1.68$. In contrast, when the categories were represented by liked Black and disliked White targets, participants showed a nonsignificant preference for the White targets, mean IAT D = 0.08, $SD = 0.37$, $t(53) = 1.7$, $p = .10$, $d = 0.23$. Critically, automatic race evaluations were significantly stronger when the Black exemplars were disliked rather than liked and the White exemplars were liked rather than disliked, $t(53) = 5.57$, $p < .0001$, $d = 0.77$, showing that the IAT is sensitive to the specific exemplars used to represent social groups.

Together with the main results of Experiment 1, these data suggest that automatic evaluations indexed by the IAT can be influenced by both (a) the exemplars composing a social group as well as (b) the categories into which those exemplars are classified. However, it is clear that because of this apparent interaction between exemplar composition and category labels, the IAT may not be able to provide an unambiguous measure of contextual effects on automatic attitude. To redress this limitation, we revisit the issue of contextual effects on automatic attitudes with a more flexible measure of automatic associations in Experiments 4 and 5.

However, prior to addressing these methodological concerns, we first turn to a point of theoretical interest. Having demonstrated that contextual cues can influence automatic attitudes, we ask whether perceivers have any explicit insight into the bottom-up nature of attitude construction. That is, can perceivers anticipate that targets (e.g., Michael Jordan) will elicit a highly positive automatic attitude in one context and a neutral or negative automatic attitude in another? Or, in contrast, will perceivers instead predict that their evaluations will remain stable across contexts?

In addition, one potential concern arising from these demonstrations is that the processes giving rise to the evaluative shifts observed in the first two experiments are not entirely automatic but rather reflect the operation of more controlled, explicit mecha-

nisms. Although the IAT appears to be resistant to self-presentational artifact (Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim & Greenwald, 1998), such rapidly shifting attitudes have not been examined in previous research. In an attempt to make sense of the fluctuating task demands of our first two experiments (e.g., first categorizing a target by race, then by occupation), participants may have deployed top-down processes that they would not typically bring to bear on more traditional IAT tasks. We address these potential concerns in Experiment 3 by examining whether participants can consciously predict contextual shifts in automatic evaluations.

Experiment 3: Explicit Predictions of Automatic Attitudes

Experiment 3 consisted, in part, of a replication of Experiment 1, given that it represents the first such study to demonstrate a sharp, rapid attitude dissociation as a function of the attended category. In addition, we examined whether perceivers have any conscious understanding that changes in evaluative context can alter the automatic evaluations they will express toward a target. This question may provide insight into whether the effects of changing evaluative contexts is sensed and understood consciously by participants. As in Experiment 1, participants alternately categorized liked Black athletes and disliked White politicians by occupation and race. Prior to each IAT block, however, participants predicted the speed with which they believed they would be able to complete the task. Because they were unlikely to be able to verbalize the speed with which they could complete the IAT, participants were instructed to simulate IAT responding by pressing two response keys at the same speed at which they anticipated being able to respond to items on the critical task. In this way, Experiment 3 measured each participant's automatic attitude toward targets as a function of categorization task (replicating Experiment 1) as well as their predicted attitude. If their predicted attitude was consistent with the one measured by the IAT, that is, if perceivers were able to demonstrate a relative shift in their attitude as a function of the category, we would learn that an understanding of the effects of evaluative context was accessible to conscious thought. However, if participants were not good at predicting the changes in their behavioral attitude, the results would show such shifts were inaccessible and suggest a mechanism by which a sense of attitude stability may be maintained in the face of variations in behavior.

Method

Participants

A total of 32 volunteers at Harvard University were paid $5 each in exchange for participation. Six participants did not follow instructions for the simulation tasks and were removed from the analysis, leaving 26 participants for analysis.

Procedure

Using the same stimulus materials from Experiment 1, participants first rated athletes and politicians for conscious expression of liking. After the rating task, they engaged in three IAT categorization tasks. The first IAT measured automatic evaluations toward Coke and Pepsi colas and served as a practice phase to introduce the IAT and simulation tasks. The remaining two IATs were identical to the race and occupation tasks in Experiment 1 except for the addition of a simulation block. In the Coke–Pepsi task, participants first practiced categorizing good and bad evaluative words and then practiced categorizing Coke-related and Pepsi-related pictures. After these two practice blocks, participants were shown one of two dual-categorization configurations (e.g., Coke + good and Pepsi + bad) and were instructed that they were to press one key in response to Coke-related pictures and good words and another key in response to Pepsi-related pictures and bad words.

Prior to performing the actual categorization task, however, participants were asked to demonstrate the speed at which they believed they could classify stimuli within that configuration. For 20 trials, participants pressed either of the two response keys at a rate they predicted they would be able to respond during the actual categorization task. A counter on the screen began at 20 and decreased by 1 on each key press, and the reaction time between each key press was recorded. The first two trials from each block were eliminated as buffer trials. To capture participants' explicit predictions of task performance before they unintentionally formed nonexplicit response sets yet still retain enough data to assess a reliable effect, we used Trials 3–15 from each block to calculate predicted response latencies. After the simulation block, participants performed the actual categorization task for 20 trials, paused, and then completed 40 more trials of the actual categorization task. Following the procedures of Greenwald et al. (2003), we used all trials from these blocks to calculate the actual response latencies for each block. After completion of the critical block for one configuration, the key mapping for Pepsi- and Coke-related pictures was reversed. Participants practiced categorizing Pepsi- and Coke-related pictures with the new key mappings and were then shown the remaining dual-categorization configuration. Participants once again simulated their reaction times within this new categorization block, performed the actual categorization task for 20 trials, paused, and then performed the task again for an additional 40 trials.

After this practice IAT, participants completed race (Black or White) and occupation (athlete or politician) categorization IATs, in counterbalanced order. The details of these IATs were exactly as those described for the Coke–Pepsi IAT, except that stimuli consisted of the individually selected liked Black athlete and disliked White politician names.

Design

In Experiment 3, we used a 2 (block: Black or athlete + good, Black or athlete + bad) × 2 (categorization task: race, occupation) × 2 (response phase: simulated, measured) design. Although not of theoretical interest, Name Rating Order (politician names first, athlete names first), Task Order (race categorization task first, athlete categorization task first), and Block Order (Black or athlete + good first, Black or athlete + bad first) were included as between-subject counterbalancing factors. None of the counterbalancing factors interacted with any of the primary comparisons of interest, Fs < 2.77, ps $> .11$; results are reported collapsed across these counterbalancing factors.

Results and Discussion

Data for 3 evaluative IATs (Coke–Pepsi, Black–White, athlete–politician) and 3 simulation tasks for those IATs were analyzed using the same procedures described in Experiment 1. Of 16,001 total trials, 7 (0.044%) were discarded as outliers. The right panel of Figure 1 presents mean response latencies as a function of IAT block and categorization task in the critical blocks of Experiment 3. As before, positive IAT D scores reflect positive evaluations of White politicians relative to Black athletes in both the race-salient and occupation-salient conditions. As in the previous experiment, actual automatic evaluations of the liked Black athletes and disliked White politicians were dependent on the salience

of race or occupation. Participants preferred Black athletes to White politicians when occupation was salient ($d = -0.41$, $p < .05$) but showed a nonsignificant preference for White politicians when race was salient ($d = 0.10$, $p = .63$). A comparison of the effects obtained in the race and occupation tasks revealed a significant shift in automatic evaluations as a function of the categorization task, $t(25) = 2.63$, $p = .015$, $d = 0.54$.

It is critical to note, however, that participants failed to predict this evaluative shift in the simulation phase, $t(25) = 0.93$, $p = .36$, $d = 0.19$. Instead, participant simulations showed a directional preference for Black athletes over White politicians whether occupation ($d = -0.41$, $p = .03$) or race ($d = -0.20$, $p = .32$) was salient. Even though mean level effects suggest the participants had little sensitivity to changes in the salience of race or occupation, it is possible that sensitivity to the salience shift could be observed in individual differences. To test this possibility, we compared simulated performance to actual performance for each of the tasks. Zero-order correlations between simulated and real task performance for all three tasks indicated that participants were unable to anticipate their actual task performance (Coke–Pepsi: $r = .01$, $p = .97$; Black–White: $r = .08$, $p = .68$; athlete–politician: $r = .19$, $p = .35$).

Finally, to examine participants' ability to anticipate shifts in evaluation between the race-salient and occupation-salient conditions more directly, we calculated difference scores between actual shifts in evaluation and between predicted shifts in evaluation in those conditions. The zero-order correlation between real shifts and predicted shifts was nonsignificant and even slightly negative ($r = -.17$, $p = .41$), indicating that participants were unable to predict the real shifts in evaluation of Black athletes and White politicians when the salience of race and occupation were manipulated. In sum, although evaluations shift as a function of changes to category salience, participants are unable to predict the nature of those evaluative shifts.

Experiment 4: Context as Category Distinctiveness

Although we demonstrated in Experiments 1 and 3 that the expression of one's automatic evaluations toward a target can be altered by how the target is categorized, the attitudes-as-constructions view predicts that attitudinal shifts should occur with even more subtle manipulations of the evaluative context. One common aspect of the situation that can produce different contexts for social judgment is the distinctiveness of one feature of a target relative to others. When encountering a group of five women and one man, for example, a perceiver may be likely to construe the singleton man along the dimension of gender but, alternately, use some other construal (e.g., race or age) to individuate the female targets. In Experiments 4 and 5, we examined whether subtle changes in the distinctiveness of target features could alter the evaluative context against which a target is judged. In other words, is manipulating the situation such that a target is the lone African American or female in a group enough to engender a different evaluative context and thus produce different automatic attitudes?

To this end, Experiments 4 and 5 measured automatic attitudes toward two multiply categorizable groups, Black females and White males. These two groups are particularly useful for examining qualitative changes in elicited automatic attitudes, because gender and race are, in both cases, associated with automatic attitudes of opposing valence. As reviewed above, previous research has consistently demonstrated negative automatic attitudes toward African American targets relative to European American targets (e.g., Fazio et al., 1995; Greenwald et al., 1998). In addition, male targets generally elicit more negativity than do female targets, on both indirect (Carpenter & Banaji, 2000; Lemm & Banaji, 1998) and direct (Eagly & Mladinic, 1989) attitude measures. In light of these findings, we predicted that when targets' race was salient, Black females would be evaluated negatively (consistent with negative automatic attitudes toward African Americans generally), whereas White males would be evaluated positively. In contrast, however, when targets' gender was salient, we predicted that Black females would be evaluated positively (consistent with positive automatic attitudes toward females generally), whereas White males would be evaluated negatively.

In Experiments 4 and 5, we tested these predictions under conditions in which one feature of Black females and White males was made salient through a "category distinctiveness" manipulation, whereby target stimuli differed from distractors in either race or gender. In these experiments, we made use of the GNAT (Nosek & Banaji, 2001) to measure automatic attitudes toward Black female and White male targets as a function of the gender and race composition of surrounding distractor individuals.

GNAT

The GNAT is derived from the same logic as the IAT and other response competition tasks: Performance is superior when one is required to make the same response to strongly rather than weakly associated items. The GNAT differs from the IAT in that it measures evaluations toward a single category without necessitating an explicit, contrasting category. Participants are instructed to respond before a prescribed deadline to items that fall into either of two concept-plus-evaluation pairings (using a single key, such as a space bar) and simply to ignore any item that does not fit the two categories. For example, participants might be instructed to respond to items that represent Black males and evaluatively positive items (but to ignore all other types of items, e.g., Black females; Hispanic, Asian, or White males and females; and evaluatively negative items).

Using a response deadline, the GNAT requires participants to respond within a brief window of time that can be varied in length (e.g., 500–700 ms). Rather than using response latencies, the GNAT indexes performance by signal detection theory's estimate of sensitivity, d' (Green & Swets, 1966). Within a concept-plus-evaluation pairing, d' indexes a participant's ability to discriminate targets (the signal) from distractors (noise). For example, when a group is strongly associated with a valence pairing, participants should more easily discriminate targets from distractors, resulting in higher d' scores than when a group is dissociated from or weakly associated with a valence pairing.

Unlike the IAT, the GNAT allows for the measurement of automatic attitudes toward a category of targets without requiring that the contrasting category consist of a homogeneous set of items that can all be classified the same way. Although the GNAT and IAT both measure automatic attitudes relative to some contrasting category, distractor items in the GNAT can be freely manipulated (a) without the need to form a unitary category and (b) without drawing observers' attention to changes in the set of distractor

items. In Experiments 4 and 5, we capitalized on these advantages of the GNAT procedure by manipulating distractor items to make distinctive different features of multiply categorizable targets. Specifically, automatic attitudes toward Black females and White males were measured three times in blocks that made distinctive either gender or race or neither feature of these targets.

By adopting the GNAT method, we did not explicitly manipulate the way that exemplars were categorized in Experiments 4 and 5. Experiments 1 and 3 were limited by the methodological requirements of the IAT that forced participants to switch the way in which they categorized targets, that is, making the context manipulation very explicit. Such an approach left open the possibility that the IAT tasks were simply measuring participants' automatic attitudes toward the category labels of athlete and politician. Although Experiment 2 demonstrated that the exemplar stimuli making up the IAT do influence automatic attitude expression, the GNAT method more directly addresses this concern by allowing context manipulation in the absence of explicit changes to the categorization task.

Method

Participants

A total of 10 White female undergraduates at Harvard University received $5 each in exchange for participation. Participation was restricted to White women because earlier research (e.g., Carpenter & Banaji, 2000) has demonstrated that this population holds strongly positive evaluations of both female and White targets. Consequently, they are a good sample in which to investigate the role of category distinctiveness in creating the evaluative context against which Black females and White males are evaluated.

Apparatus and Stimuli

Stimulus presentation and response latency recording were controlled by an Apple Macintosh G3 running Psyscope (Cohen, MacWhinney, Flatt, & Provost, 1993) software. Experiment 4 used 15 positive words (e.g., *caress*, *paradise*), 15 negative words (e.g., *agony*, *disaster*), 15 stereotypic Black female names (e.g., *Latoya*, *Shaniqua*), 15 stereotypic Black male names (e.g., *Leroy*, *Tyrone*), 15 stereotypic White female names (e.g., *Meredith*, *Peggy*), and 15 stereotypic White male names (e.g., *Brandon*, *Todd*) taken from Greenwald et al. (1998).

Procedure

At the beginning of each GNAT block, two labels appeared on screen, for example, *Black Female* and *good*. Stimulus items were presented sequentially in the center of the screen for 600 ms, and participants were instructed to press the space bar if an item that fell into either category was presented (inclusion trials) but to make no response to any other type of item (distractor trials). The trial was scored as correct if the participant responded to an inclusion trial before the 600-ms deadline (hit) or if they avoided responding to a distractor trial (correct rejection). A buzzer sounded to indicate an error if the participant failed to respond to an inclusion trial before the deadline (miss) or responded to a distractor trial (false alarm). Sensitivity in each block was indexed by using the number of hits and false alarms to calculate d' (Green & Swets, 1966).

Participants first completed a set of six practice blocks. At the start of each practice block, a single category label was presented onscreen: One practice block each was completed for Black female, Black male, White female, and White male names in addition to one practice block each for positive and negative words. Participants were instructed to respond by pressing the space bar whenever an item belonging to the category denoted by the label was presented and to make no response for any other type of item. Each practice block consisted of 26 trials, half of which were inclusion trials that belonged to the category, whereas the other half were exclusion trials randomly distributed among the other five categories.

After the practice blocks, participants completed 20 critical GNAT blocks. At the start of each critical GNAT block, two category labels were presented onscreen: one referring to a social group and the other to a set of valenced words, for example, *male* + *good*. Each GNAT block had a companion block in which the same social group was paired with words of the opposite valence, for example, *male* + *good* and *male* + *bad*. Differences in accuracy (indexed by d') across these paired blocks served as an index of automatic attitudes toward the social group. For example, more accurate responding in the *male* + *bad* block would indicate that male targets elicited a negative automatic attitude.

Superordinate group blocks. Four pairs of GNAT blocks measured automatic attitudes toward superordinate race (Black, White) and gender (female, male) groups. Superordinate group blocks consisted of 30 inclusion trials and 30 distractor trials in random order.

Subgroup blocks. Automatic attitudes toward subgroup (Black female, White male) targets were measured in three different pairs of GNAT blocks. For each of the two subgroup targets, one pair of blocks used distractor items that differed from the subgroup only along the dimension of race (race construal), whereas another pair of blocks used distractors that differed only along the dimension of gender (gender construal). As an example, for the subgroup Black female, the race-construal condition used White female and White male names as distractors, whereas the gender-construal condition used Black male and White male distractors. A third pair of blocks used items that differed from subgroup targets along two dimensions (neutral construal). For both Black female and White male targets, the neutral-construal condition used Black male and White female distractors.

Each subgroup block consisted of 65 trials, the first 5 of which were pseudopractice trials excluded from analyses. Because category distinctiveness was established by the distractor names, these pseudopractice trials consisted of four distractor names, included to ensure that category distinctiveness was manipulated from the very beginning of the block. The remaining trials comprised 30 inclusion and 30 distractor trials in random order. Although described here as sequential pairs, participants actually completed all 20 GNAT blocks in random order.

Design

We measured automatic attitudes elicited by multiply categorizable social targets using a 2 (subgroup: Black female, White male) × 2 (word valence: bad, good) × 3 (category distinctiveness: gender, race, neutral) factorial. In addition, the experiment measured automatic attitudes toward superordinate group targets using a 4 (superordinate group: Black, White, female, male) × 2 (word valence: bad, good) design. All factors were manipulated within-subject.

Results and Discussion

Before testing the central idea of whether multiply categorizable targets elicited different automatic attitudes when gender, race, or neither feature was made salient through distractors, we first examined automatic attitudes toward superordinate gender (female, male) and race (Black, White) targets. Subsequently, we examined automatic attitudes toward multiply categorizable targets as a function of evaluative context.

Sensitivity, indexed by d', was calculated for each critical block and measured a participant's ability to discriminate targets from distractors within a block. The automatic attitude toward a group

was represented by the sensitivity difference (in d' scores) between (a) the block in which the group was paired with positive words (e.g., male names + positive words) and (b) the block in which the group was paired with negative words (e.g., male names + negative words). Negative word blocks were subtracted from positive word blocks; using this convention, positive automatic attitudes toward the social group were designated by d' difference scores greater than zero.

Automatic Attitudes Toward Superordinate Groups

The top panel of Figure 2 presents sensitivity scores observed in superordinate gender (female, male) and race (Black, White) GNAT blocks. Both female targets and White targets elicited positive automatic attitudes, whereas both male targets and Black targets elicited negative automatic attitudes. A Superordinate Group × Word Valence analysis of variance (ANOVA) confirmed that superordinate group targets elicited different automatic attitudes, $F(3, 27) = 31.11$, $p = 10^{-7}$, $f = 1.86$. A planned contrast analysis in which we applied a lambda weight of 1 to superordinate groups for which we expected to observe a positive automatic attitude (i.e., female and White) and a lambda weight of –1 to superordinate groups for which we expected to observe a negative automatic attitude (i.e., male and Black) confirmed the expected pattern of results in a more focused test, $t(9) = 10.03$, $p = 10^{-6}$, $d = 3.34$. As in earlier research using a variety of automatic attitude measures, the GNAT revealed positive automatic gender attitudes toward female targets relative to male targets and negative automatic race attitudes toward Black targets relative to White targets.

Automatic Attitudes Toward Multiply Categorizable Targets

In light of these superordinate group results, we expected to observe different automatic attitudes toward Black female and White male targets as category distinctiveness cues altered the evaluative context in which these targets were evaluated. The top panel of Figure 3 presents sensitivity scores observed in subgroup blocks under gender, race, and neutral distinctiveness conditions.

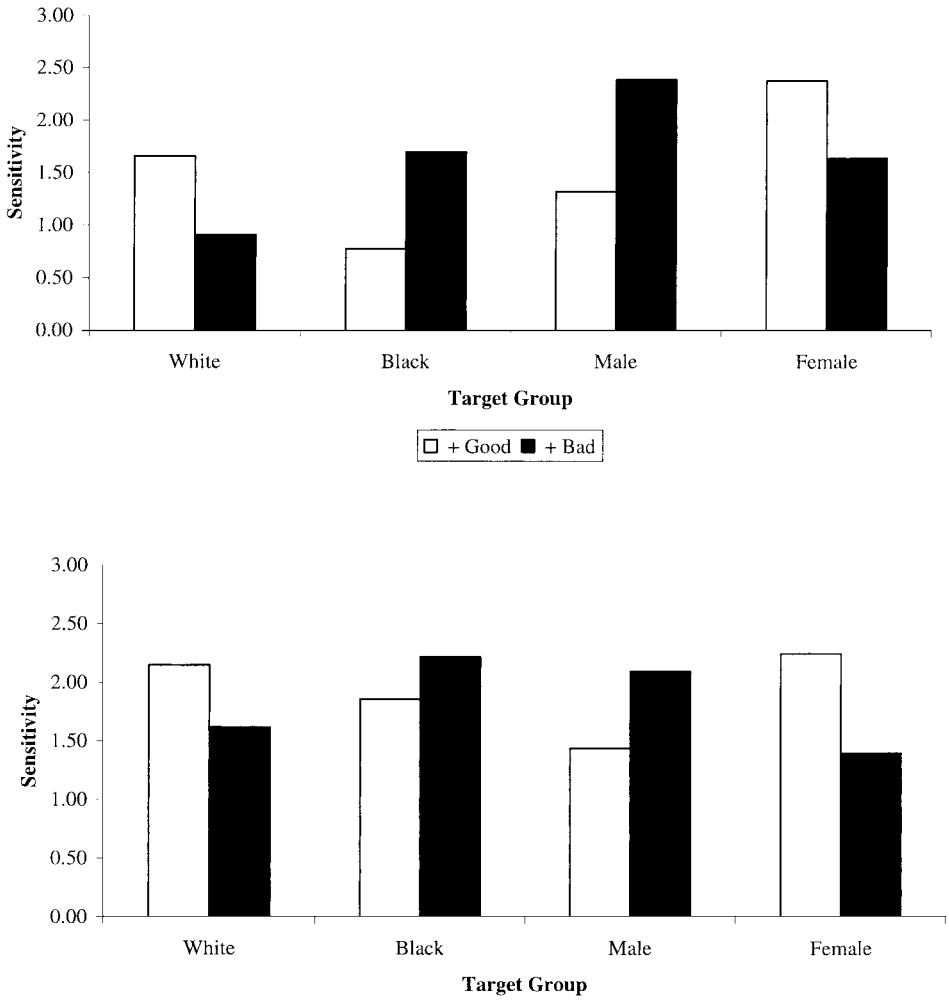

Figure 2. Sensitivity (d') by evaluative pairings for superordinate groups obtained in Experiment 4 (top panel) and Experiment 5 (bottom panel). Higher sensitivity scores indicate a stronger association between target group and evaluation.

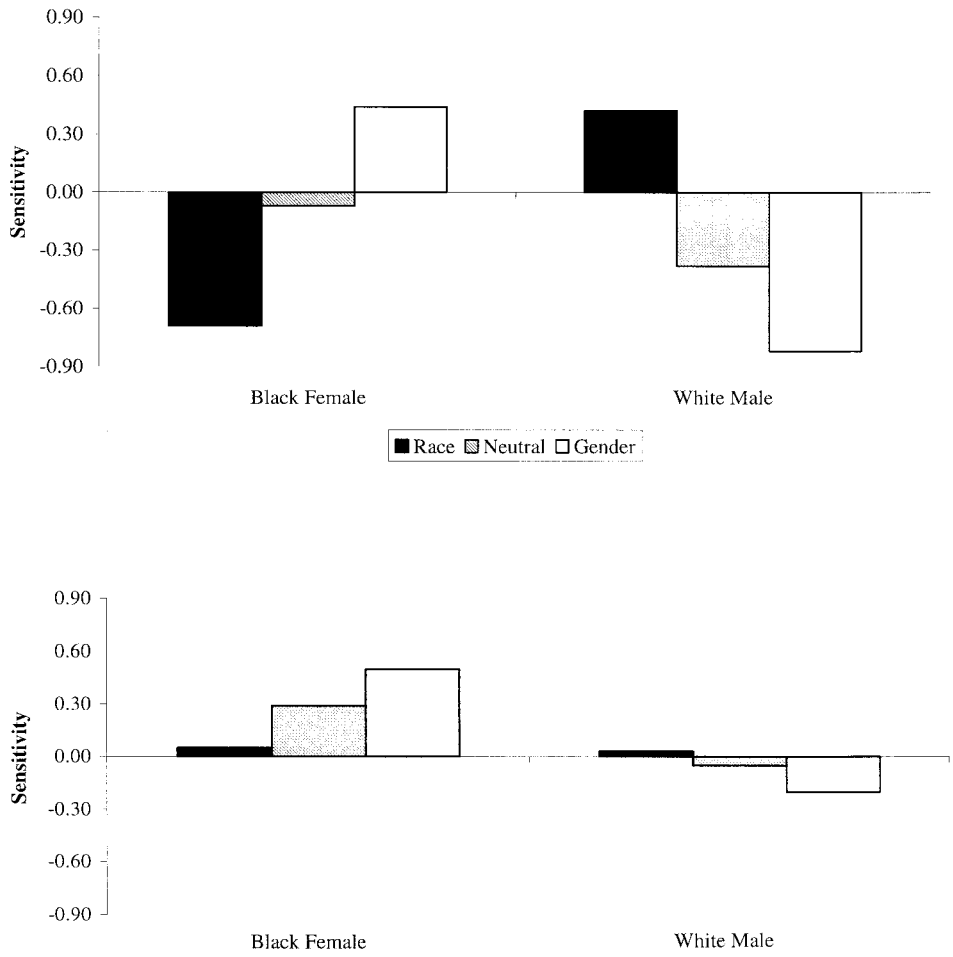

Figure 3. Sensitivity (d') difference scores for Black female and White male targets in Experiment 4 (top panel) and Experiment 5 (bottom panel). Values represent the difference between a subgroup paired with positive words and one paired with negative words. Positive values indicate a positive automatic attitude toward a subgroup.

Black female and White male targets produced a contrasting pattern of automatic attitudes as a function of whether gender, race, or neither feature was made distinctive. A three-way Subgroup × Word Valence × Category Distinctiveness ANOVA confirmed that Black female and White male targets elicited contrasting automatic attitudes as a function of category distinctiveness, $F(2, 18) = 12.45$, $p = 10^{-4}$, $f = 1.18$. An analysis on only the gender and race conditions confirmed the predicted pattern of results in a more focused comparison, $F(1, 9) = 22.85$, $p = .001$, $d = 1.59$. Furthermore, this pattern of different automatic attitudes as a function of category distinctiveness was obtained separately for both Black females, $F(2, 18) = 6.13$, $p < .01$, $f = 0.83$, and White males, $F(2, 18) = 7.93$, $p < .004$, $f = 0.94$.

The results of Experiment 4 demonstrated that manipulating the salience of one feature of a target by altering the gender and race characteristics of a contrasting group was sufficient to change the evaluation of a social target. When encountering a target against the backdrop of others who differed along the dimension of gender (i.e., Black female targets against male distractors or White male targets against female distractors), perceivers construed targets according to gender, and their automatic attitudinal responses were consistent with their positive evaluations of the superordinate gender group. For example, Black females were evaluated as positively as females in general. In contrast, when encountering a target against the backdrop of others who differed along the dimension of race (i.e., Black female targets among White distractors or White male targets among Black distractors), perceivers construed targets according to race, and their automatic attitudinal responses were consistent with their evaluations of the superordinate race group. For example, Black females were evaluated as negatively as African Americans in general (correspondingly, both effects were reversed for White male targets).

Using a novel measure of automatic attitudes, we extended previous findings in Experiment 4 by demonstrating that automatic attitudes toward the same targets can vary dramatically as a function of the context in which such attitudes are elicited. Unlike earlier research in which the targets making up a category were manipulated (Dasgupta & Greenwald, 2001) or a Black confederate was present or absent (Lowery et al., 2001), Experiment 4 was not designed to manipulate the composition of the evaluated group.

Rather, automatic attitudes were elicited in response to the same Black female and White male targets across all experimental conditions. Moreover, in Experiment 4, the way that exemplars were categorized was not manipulated. Even across blocks in which participants consistently categorized targets as Black females, qualitatively different automatic attitudes were observed toward this group. Taken together, these results demonstrate that the evaluative context in which an automatic attitude is elicited can be altered substantially by incidental environmental information that renders a target feature more or less distinctive than other individuals.

Although the GNAT introduces several methodological advantages over the IAT, the two tasks nevertheless share a number of important limitations. In particular, although the GNAT allows increased flexibility in choosing distractor stimuli, participants may nevertheless spontaneously impose a second, contrasting category on the distractors if the items happen to form a coherent set. In such cases, the GNAT may operate in a manner very similar to that of the IAT. However, although participants in the present study could conceivably have adopted such a strategy in blocks that measured automatic attitudes toward superordinate group targets (e.g., Black or male), the experimental design diminishes the likelihood that they used this strategy for subordinate blocks (i.e., Black female and White male). Automatic attitudes toward subgroups were measured three different times in blocks that were not temporally adjacent and that used different distractor items. As such, to maintain a spontaneous categorization strategy, participants would be required to keep track of all preceding stimuli in order to discern the single dimension along which the distractors could be categorized. Given the online demands of the task (e.g., responding within a 600-ms window), it seems highly unlikely that participants consciously adopted such a strategy during task performance.

Experiment 5: Testing Category Distinctiveness With Pictures

Others have argued and we agree that social judgments are often triggered by visual encounters with social group members rather than lexical representations of such targets (Gilbert & Hixon, 1991). Notwithstanding, a majority of experiments on person perception have exclusively used verbal information such as names to activate group membership judgments. This overreliance on verbal stimuli may be a shortcoming of contemporary social cognition research: Abundant evidence in other areas certainly shows that mental representations of pictures and words differ (e.g., Farah, 1992; Glaser & Glaser, 1989; Israel & Schacter, 1997).

Indeed, as suggested by recent theorists (Macrae & Bodenhausen, 2000; Zárate & Smith, 1990), a dependence on verbal stimuli may partially account for the stunted development of social cognition research on multiple categorizability. These theorists point out that by their very nature, verbal stimuli often provide perceivers with a built-in solution to the construal problem by presenting only one salient dimension along which a target can be construed. In light of these criticisms and to establish the replicability of Experiment 4, we examined in Experiment 5 whether the effect of category distinctiveness on automatic attitudes extends to pictorial representations of social groups.

Method

Participants

A total of 22 White female undergraduates at Harvard University participated in exchange for $5 each.

Stimuli

In Experiment 5, we used color images from the Corel Mega Gallery (1997, disc 3) clip art CD-ROM. They consisted of 15 positive objects (e.g., a trophy, a balloon, flowers), 15 negative objects (e.g., a gun, poison, a spider), 15 Black female faces, 15 Black male faces, 15 White female faces, and 15 White male faces.

Procedure

The procedure was identical to that of Experiment 4, except for two changes. First, clip art images were used in place of the words and names used earlier. Second, the response window was shortened to 500 ms. Because images could generally be classified more quickly than verbal stimuli, the shorter response window was used to keep overall accuracy comparable to that found in Experiment 4 (approximately 60%).

Results and Discussion

Automatic Attitudes Toward Superordinate Groups

The bottom panel of Figure 2 presents sensitivity scores observed in superordinate gender (female, male) and race (Black, White) GNAT blocks. As in Experiment 4, both female targets and White targets elicited positive automatic attitudes, whereas male targets and Black targets elicited negative automatic attitudes. A Superordinate Group $\times$ Word Valence ANOVA confirmed that superordinate group targets elicited different automatic attitudes, $F(3, 63) = 26.68$, $p = 10^{-11}$, $f = 1.13$. A planned contrast analysis (see Experiment 4) confirmed the expected pattern of results in a more focused test, $t(21) = 8.49$, $p = 10^{-8}$, $d = 1.85$. These results mirror those of Experiment 4 while using pictorial representations of social group members.

Automatic Attitudes Toward Multiply Categorizable Targets

Consistent with these superordinate group results and as in Experiment 4, we expected to observe different automatic attitudes toward Black female and White male targets as category distinctiveness cues altered the evaluative context in which these targets were evaluated. The bottom panel of Figure 3 presents sensitivity scores observed in subgroup blocks under gender, race, or neutral construal conditions. A three-way Subgroup $\times$ Word Valence $\times$ Category Distinctiveness ANOVA confirmed that Black female and White male pictorial representations elicited contrasting automatic attitudes as a function of category distinctiveness, $F(2, 42) = 8.69$, $p = 10^{-4}$, $f = 0.64$. Further analysis on only the gender and race conditions confirmed the predicted pattern of results in a more focused comparison, $F(1, 21) = 7.35$, $p = .01$, $d = 0.59$. Finally, this pattern of different automatic attitudes was obtained marginally in Black female blocks, $F(2, 42) = 2.54$, $p = .09$, and significantly in White male blocks, $F(2, 42) = 9.51$, $p = 10^{-4}$, $f = 0.67$.

Using pictorial representations of social objects, we replicated in Experiment 5 the observation that manipulations in category distinctiveness can define the evaluative context of a social target and, subsequently, evoke different automatic attitudinal responses. We note that substantially smaller automatic attitude effects were obtained using pictorial stimuli than verbal stimuli; for example, the effect size associated with the focused comparison for multiply categorizable targets was substantially smaller in Experiment 5 ($d = 0.59$) than in Experiment 4 ($d = 1.59$). However, even this smaller effect is still of substantial magnitude, exceeding the cutoff for a medium-sized effect (Cohen, 1988). Furthermore, pictorial stimuli are generally associated with less extreme automatic attitude effects on the IAT (Nosek et al., 2002). Given the underlying conceptual similarity between the GNAT and the IAT, it is unsurprising but reassuring that we observed comparable differences between pictorial and verbal stimuli between Experiments 4 and 5.

General Discussion

Recent research has suggested that even automatic attitudes may be shaped by recent orienting experiences (Dasgupta & Greenwald, 2001; Lowery et al., 2001). The present research showed that fluctuations in such attitudes can be even more dramatic than these recent reports suggest. Across five experiments, rapid, qualitative shifts in the valence of automatic attitudes were elicited in response to the same attitude objects. To a greater or lesser degree, researchers in all previous studies on this topic (Dasgupta & Greenwald, 2001; Lowery et al., 2001) have manipulated the exemplars used by perceivers to represent a group. For example, by introducing an African American experimenter in a position of authority, Lowery et al. (2001) may have induced participants to include capable or positive individuals in their representation of the category Black. Likewise, Dasgupta and Greenwald (2001) explicitly manipulated the exemplars used to represent race categories. In contrast, in the current experiments, we circumvented this limitation by using identical exemplars to represent social categories (e.g., Black athletes or White males), demonstrating that the very same attitude objects could nonetheless elicit opposing automatic evaluations.

Furthermore, Experiment 1 demonstrated that automatic attitudes could vary in a small period of time within an individual perceiver, even when targets were well-known and strongly liked (or disliked). In addition, Experiment 2 demonstrated that evaluations of the categories Black and White shift dramatically when the targets are changed from disliked Black (and liked White) exemplars to liked Black (and disliked White) exemplars.

These data directly address a point of considerable interest and some confusion regarding the nature of effects measured by the IAT. For example, De Houwer (2001) and Fazio and Olson (2003) suggested that "the IAT seems to assess associations to the category labels, not automatically activated responses to the individual exemplars" (Fazio & Olson, 2003, p. 315). Taken together, Experiments 1 and 2 demonstrate that this conclusion was premature. Instead, the IAT clearly measures automatic attitudes that depend on both the contextual frame (provided by the categories) and the target exemplars (stimulus items). Specifically, we have identified conditions under which the typically observed automatic race attitudes (i.e., relative negativity toward Black compared with White targets) can shift to a neutral preference when these groups are represented by liked Black and disliked White exemplars (see Experiment 2 in this article). In much the same way, Nosek, Greenwald, and Banaji (2003) reported a similar shift in IAT effect on the basis of changes to individual exemplars. By changing just two of eight category exemplars representing the category gay (from ones showing male couples to ones showing female couples), these authors successfully reduced the magnitude of automatic gay bias by over 30%. In sum, it is now clear that successfully eliciting an IAT effect depends on both the category frame as well as the individual exemplars. The exact way in which these two dimensions interact to measure automatic associations using the IAT remains a task for future research.

In Experiment 3, we replicated the results of the first experiment and further demonstrated that participants do not anticipate the effects of context on their evaluations of targets. Participants' failure to predict the effect of evaluative context demonstrates that such shifts in attitude can take place completely outside of explicit awareness or conscious control. Indeed, the inability to predict the malleability of attitudes may help explain the pervasiveness of the view that attitudes are inherently inflexible and stable.

Experiments 4 and 5 extended these findings in two ways. First, in these experiments (as well as in Experiment 1), we observed a more dramatic demonstration of automatic attitude change than previously reported, such that an attitude object that elicited strongly negative evaluations in one context (e.g., Black females when race was made salient) elicited strongly positive evaluations in another (e.g., Black females when gender was made salient). It is important to note that previous research (Dasgupta & Greenwald, 2001; Lowery et al., 2001) has demonstrated moderation of but not qualitative shifts in automatic attitudes.

Second, Experiments 4 and 5 highlighted the power of subtle changes in the evaluative context to produce substantial changes in automatic attitudes. A context in which Black female targets were the lone African Americans among a group of White distractors was sufficient to elicit automatic attitudes toward Black females that reflected superordinate race attitudes, even though perceivers did not attend explicitly to the composition of the distractor stimuli. A similarly subtle manipulation in which Black females were the lone women among a group of male distractors produced automatic attitudes reflecting superordinate gender attitudes. Both effects were reversed for White males. These experiments suggest that shifts in the race or gender composition of a roomful of people may be enough to elicit very different automatic attitudes toward an individual. For individuals belonging to two (or more) superordinate groups associated with opposing evaluations (e.g., Black females), these shifts may provoke qualitatively different attitudinal responses from perceivers.

The Notion of Attitude Change

Although it is tempting to think of such effects as representing attitude change, a more parsimonious explanation must be considered. A qualitative difference in the evaluation elicited by an attitude object does not represent a change in attitude, for that would require a real or stable attitude from which the new attitude may be said to represent a change. Rather, the present experiments suggest that automatic attitudes are defined within the context established by the situation. The appearance of stability or the existence of a single real attitude arises from the high consistency

in environments that masks the fact that evaluations are continuously and actively being constructed against the backdrop of the current situation.

This view of automatic attitudes as constructed rather than retrieved is a close homologue of current views regarding episodic memory. Although folk psychological and early theoretical stances approached memories as high-fidelity historical recordings that could be played back more or less verbatim (see Roediger, 1980, for a discussion), research has since demonstrated that the act of remembering is an intrinsically constructive process that causes memories to be highly susceptible to distortions introduced by context (Loftus, Miller, & Burns, 1978).

In much the same way that misleading information can distort memory, one's explicit attitudes have also been shown to introduce retrospective biases. For example, Ross, McFarland, and Fletcher (1981) persuaded participants that frequently brushing one's teeth was associated with either positive or negative health outcomes. When later asked to recall the number of times they had brushed their teeth in the preceding 2 weeks, those participants who had been persuaded of the benefits of the behavior reported more frequent brushing than did those who were led to believe that moderate amounts of brushing were optimal. Much like the misleading suggestion that an object was present in an earlier scene, a change to one's explicit attitude brought about through persuasive messages was subsequently incorporated into memory for past events.

The Notion of a True Attitude

A common thought experiment asks the question, "What color would a chameleon appear in a room of mirrors?" Of course, this brainteaser relies on the assumption that chameleons do, in fact, possess one true color, which, because the chameleon rapidly assimilates to its environment, is never directly observed. In much the same way, social psychologists have tacitly assumed that for any given attitude object, a perceiver must possess one true attitude, although expression of this authentic attitude is prevented by self-presentational biases or the impossibility of accurate introspection. Just as the chameleon may have one true but rarely observed color, so too have people been assumed to have one true, rarely observed attitude toward an attitude object.

To some extent, measures of automatic attitudes have been offered up as the chameleon's mirror for social cognition (Fazio et al., 1995). In an attempt to measure attitudes in isolation from obscuring influences, it has been assumed that stable, genuine attitudes exist and that implicit measures provide a lens through which authentic responses can be observed. However, coupled with earlier research (Blair, 2002; Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Lowery et al., 2001), the current findings cast doubt on the belief that there exist single, unitary attitudes awaiting authentic observation by implicit measures. With variation in context, multiple evaluations of an attitude object may be evoked, but none of those evaluations is more true than any other, even though some that are culturally privileged may be observed in the vacuum of the laboratory.

This view is reminiscent of the well-known baseball anecdote in which the plate umpire suffers a moment of hesitation before calling a crucial pitch. Anxiously, the batter whirls around and demands, "Well, was it a ball or a strike?" The umpire responds, "What do you mean, *was*? Son, it ain't nothing until I call it!" In the same vein, we suggest that, like for baseball pitches, no hidden, platonic form of automatic attitudes exists, waiting to be measured. Both positive and negative evaluations are possible, even probable, given a perpetually shifting context of evaluation. Abandoning a search for singular, true evaluations of social objects may be necessary to pursue an understanding of the true nature of evaluation.

References

Abelson, R. P. (1988). Conviction. *American Psychologist, 43,* 267–275.

Asch, S. E. (1940). Studies in the principles of judgments and attitudes: II. Determination of judgments by group and by ego standards. *Journal of Social Psychology, 12,* 433–465.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117–150). Washington, DC: American Psychological Association.

Banaji, M. R. (in press). The opposite of a great truth is also true. In J. Jost, D. Prentice, & M. R. Banaji (Eds.), *The yin and yang of progress in social psychology: Perspectivism at work.* Washington, DC: American Psychological Association.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes toward homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48,* 145–160.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology, 62,* 893–912.

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

Blair, I. V., Ma, J., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of automatic stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81,* 828–841.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. J. Wyer (Eds.), *Advances in social cognition* (pp. 1–36). Hillsdale, NJ: Erlbaum.

Carpenter, S., & Banaji, M. R. (2000). *Implicit gender attitudes: Group membership, cultural construal, and malleability*. Unpublished manuscript, Yale University, New Haven, CT.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, & Computers, 25,* 257–271.

Corel Mega Gallery. (1997). [Computer software]. Ottawa, Ontario, Canada: Corel Corporation.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800–814.

De Houwer, J. A. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37,* 443–451.

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

Draine, S. C. (1997). Inquisit (Version 1.0) [Computer software]. Seattle, WA: Millisecond Software.

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15,* 543–558.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit

association test for assessing anxiety. *Journal of Personality and Social Psychology, 83,* 1441–1455.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science, 1,* 164–169.

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

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54,* 297–327.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). San Diego, CA: Academic Press.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60,* 509–517.

Glaser, W., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118,* 13–42.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48,* 85–93.

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

Gross, S. R., & Ellsworth, P. C. (2003). Second thoughts: Americans' views on the death penalty at the turn of the century. In S. P. Garvey (Ed.), *Beyond repair? America's death penalty* (pp. 7–57). Durham, NC: Duke University Press.

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4,* 577–581.

Kim, D.-Y., & Greenwald, A. G. (1998, May). *Voluntary controllability of implicit cognition: Can implicit attitudes be faked?* Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Lemm, K. M., & Banaji, M. R. (1998, May). *Implicit and explicit gender identity and attitudes toward gender*. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31.

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 265–343). London: Academic Press.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81,* 842–855.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51,* 93–120.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McClelland, J. L., & Rumelhart, D. E. (1998). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1988). The appeal of parallel distributed processing. In A. M. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence* (pp. 52–72). San Mateo, CA: Kaufmann.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4, pp. 1–18). New York: Plenum.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19,* 625–666.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration Web site. *Group Dynamics, 6,* 101–115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2003). *Understanding and using the Implicit Association Test: 2*. Unpublished manuscript.

Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition, 8,* 231–246.

Ross, M., McFarland, C., & Fletcher, G. J. (1981). The effect of attitude on the recall of personal histories. *Journal of Personality and Social Psychology, 40,* 627–634.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 143–179). Boston: McGraw-Hill.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70,* 893–912.

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74,* 21–35.

Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 446–496). Boston: McGraw-Hill.

Zárate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition, 8,* 161–185.

Received October 11, 2001
Revision received February 5, 2003
Accepted March 24, 2003 ■