Contextual Variations in Implicit Evaluation

Jason P. Mitchell
Harvard University

Brian A. Nosek
University of Virginia

Mahzarin R. Banaji
Harvard University

Running Head: IMPLICIT EVALUATIONS

Abstract

The present research examined contextual variations in automatic attitudes that are assumed to be relatively stable. In four experiments, using two variations of measures of automatic attitudes, evaluative responses were shown to differ qualitatively by shifting focus on one or another aspect of social group membership (e.g., race or gender). For example, evaluations of liked Black athletes were positive when their occupation was the focus of judgment, but negative when their race was the focus of judgment (Expts 1 & 2). The opposite result was obtained for disliked White politicians (positive evaluations when race was the focus, negative evaluations when occupation was the focus). Variations in automatic attitudes as a function of context were obtained when the manipulation involved overt categorization (Expts 1 & 2) as well as more subtle contextual cues, such as category distinctiveness (Expts 3 & 4). Furthermore, participants were shown to be unable to predict such contextual influences on attitudes (Expt 2). Taken together, the results of these experiments support the idea of attitudes as continuous reconstructions rather than as the retrieval of invariant stored evaluations. Thus, multiple, independent evaluations of social objects can exist simultaneously. These experiments suggest that an ongoing process of attitude construction results in contextually-appropriate social responses that are inherently flexible, despite being outside of conscious control.

Contextual Variations in Implicit Evaluations

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 relied on 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 1980's mental representations were cast as things, capable of being stored or retrieved, like locating a can of beans in the pantry. More recently, in an effort to change that way of thinking, mental representations have been cast 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 the "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, 1988; 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 and 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). In viewing mental representations as dynamically reconstructed, not statically retrieved, distributed models have brought to the fore theoretical questions previously unexplored.

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 their very nature.

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 paper on conviction, Abelson (1988) includes 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. Moreover, Gross and Ellsworth (2001) report that a majority of Americans indicate that their attitudes toward the death penalty, an attitude of strong conviction, had not changed over time. They point 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. 7, manuscript version).

An alternative view of attitudes, as sensitive to context, has been pointed out as well. An early experiment on group perception and, in particular, 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 presented respondents with one of two rank-ordered lists of ten 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 the bottom of the peer-ranked list, and 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 them to be ranked last by their peers.

Importantly, Asch 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, as Lord and Lepper (1999) astutely note, the explanation was psychologically far more interesting. 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).

More recently, Schwarz and colleagues have conducted extensive research to show the sensitivity of explicitly stated attitudes to contextual variables. Self-reported attitudes, they show, 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, we know that contextual factors can systemtically 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 – of attitude change that occurs merely in laboratory studies because such studies tend to use attitude objects that have little bearing on strong attitudes that are held with conviction. 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, i.e., those that lie outside conscious awareness or control, are invariant (Bargh, 1994). 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. 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).

Quite recently, two experiments using the IAT have shown that features present in the evaluation context can shape even automatic attitudes. Contrary to expectation that automatic attitudes are unbending and invariant, exposure to positive African American exemplars produced less negative evaluations of that group than in a control condition (Dasgupta & Greenwald, 2001). Likewise, the very presence of an African-American experimenter produced more positive evaluations 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 the experiments reported here, we attempt to examine such contextual effects on automatic attitudes directly with an eye towards resolving an ambiguity regarding the processes that underlie such attitude fluctuations. 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 such automatic attitude "change." 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 following exposure to positive African American exemplars, they were characteristically negative when measured after a twenty-four hour 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 attenuation 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 thing as stable, precompiled attitudes exist in the first place, and that what appears as attitude "change" is in fact attitude "construction." Consistent with theories of mental representations as 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 following exposure to positive African American individuals, such attitudes may assimilate the positivity associated with those exemplars. 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 picked up from the environment. If one's attitude toward the same object is subsequently measured in a different context, it will also necessarily incorporate different informational material, potentially resulting in the construction of an evaluatively discrepant attitude. In this view, perceivers' negativity toward African Americans twenty-four hours 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, on-line constructions. Both 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 four 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.

In Experiment 1, we demonstrate that automatic attitude toward well-known Black athletes and White politicians can vary as a function of targets are categorized by race or occupation. Having provided initial evidence of such rapid automatic attitude change, we go on to examine whether perceivers are able to anticipate the effects of contexual changes in producing attitudinal shifts. Experiment 2 replicated the observation of contextual effects on automatic attitudes, and further demonstrates that rapid automatic attitude change occurs even under conditions where participants do not predict such change. Finally, Experiments 3 and 4 examined more dramatic changes in automatic attitude. Using a newly developed method for measuring automatic evaluations, these experiments observe even sharper dissociations in evaluation than previously demonstrated.

Multiply-categorizable social targets, i.e., 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 is an inevitable consequence of categorizing a person as a member of a social group (Brewer, 1988; Devine, 1989; Fiske and Neuberg, 1990). The question posed in these four experiments concerns the variation in automatic attitude 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: 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-construction views predict that automatic attitudes will likewise vary in step with manipulations that highlight one or another feature of a target.

To increase the generalizability of the current research, we employ two indirect measures of evaluation, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and a newly-developed derivative, the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001). 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

Experiment 1 examined how a particular exemplar can elicit qualitatively different automatic evaluations as a function of contexts that highlight different exemplar features. 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/politician. In the race categorization task, the same targets were categorized on the basis of race, using the category labels Black/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 different automatic evaluations of consciously-liked Black athletes and consciously-disliked White politicians as a function of whether targets were categorized according to race or occupation.

Method

Participants

A total of 91 volunteers at Yale University were either paid $10 or received partial credit in an introductory psychology course in exchange for participation. Results from 2 participants were excluded from analysis due to a computer malfunction that erased some of the critical data, and 7 participants were excluded because an excess of 10% fast responses (< 300ms) on the implicit measures (Greenwald, Nosek, & Banaji, 2002), leaving a total of 82 participants for the analysis.

Stimuli

For the race and politician categories, 3 liked Black athletes and 3 disliked White politicians were selected for each participant. 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 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 words and names for use in the original demonstration of the Implicit Association Test.

Implicit Association Test

The Implicit Association Test (IAT; Greenwald, et al., 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 (Fazio, et al., Bargh et al.), 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+evaluation pairs, and (d) that the strength of evaluative association as measured under conditions of speeded responding is a measure of automatic attitude.

More specifically, the IAT relies on a response latency indicator obtained in the process of pairing an attitude concept (say a social group such as *old/young*) with an evaluative attribute (*good/bad*), or specific attributes that may not be purely evaluative (such as *self-other, home-career, science-arts*). In the computerized version, the pairing is achieved by using a keyboard key (say a left key) to be pressed in response to items from the two paired categories, say *old+bad* while another key (the right key) is used for the other pair, *young+good*. The differential speed to complete these two types of pairings, i.e., 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 (attitude). The IAT effect is a difference score reflecting a relative attitude that shows both the direction (positive versus negative) of implicit attitude as well as the magnitude of the attitude (larger numbers reflecting larger differences between pairings in milliseconds). Besides traditional tests of significance, this measure has typically been reported with an additional test of effect size and the data show the many instances in which the IAT effect is a large one (see Nosek, et al., 2001).

In the experiments reported in this paper, the strength of evaluation of gender and race 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 a classification of these items more rapidly when they are artificially paired (by being assigned to the same computer key for responses) compared with the opposite pairing of White with bad and Black with good. In the half-million instances of the race IATs completed via the Internet (www.yale.edu/implicit) we have observed a strong overall association of Black+bad and White+good in addition to variations as a function of group membership and other demographic properties (Nosek, Banaji, & Greenwald, 2001).

Apparatus and program

Presentation of experimental stimuli was controlled by IBM (80486 processor) desktop computers running Inquisit software (Inquisit 1.00). Participants were instructed to give left responses with their left forefinger (using the "A" key) and right responses 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 millisecond inter-trial 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. Following 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/politician. In the race task, those same targets were categorized as Black/White. The task followed the basic procedure outlined by Greenwald et al. (1998; and available for demonstration at http://buster.cs.yale.edu/implicit/) with 40 response trials in each of the critical conditions. To facilitate the shift in salience of race and occupation, practice blocks for each IAT mixed the race of the targets in the occupation task (athlete/politician), and mixed the occupation of the targets in the race task (Black/White). That is, when discriminating athletes from politicians, the practice blocks included

Black and White athletes and Black and White politicians. In the critical block of the IAT, all athlete exemplars were Black and all politician exemplars were White with the exception of two response trials in which a White athlete or a Black politician were presented to discourage recoding of the occupation task as a race task (these two trials were not included in subsequent analysis).

Design

Experiment 1 employed a 2 (Block: Black[or athlete]+good, Black[or athlete]+bad) x 2 (Categorization Task: race, occupation) within-participant 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-participant counterbalancing factors.

Results and discussion

Data preparation. Data were prepared following the IAT scoring algorithm recommended by Greenwald, Nosek, and Banaji (2002). In brief, (i) trial response latencies greater than 10,000 ms were eliminated, (ii) participants who responded faster than 300 ms on more than 10% of the trials were excluded, (iii) all response latencies were included in the analysis, (iv) response latencies faster than 400 ms were excluded from the analysis and, (v) a difference score was calculated between the two critical blocks of trials (e.g., [athlete+bad/politician+good] – [athlete+good/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, Nosek, & Banaji, 2002 for more detail about the scoring algorithm). For ease of interpretation, 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 they were categorized. When the categorization task emphasized occupation (athletes and politicians) Black athletes were preferred to White politicians (IAT D = -.29, SD = .39, $t(81) = -6.7$, $p < .0001$, $d = -.74$). This is not surprising since the exemplars consisted of 3 Black athletes individually rated as well liked by each participant, and 3 White politicians rated as disliked by each participant. When the categorization task emphasized race (Black and White), White politicians were preferred to Black athletes (IAT D = .13, SD = .43, $t(81) = 2.8$, $p = .006$, $d = .31$). This finding is substantially more surprising since negativity toward the category Black relative to White was observed despite the relative 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 suggests that manipulating the categorization frame can indeed influence the elicited automatic attitude ($t(81) = 7.18$, $p < .0001$, $d = .80$).

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

One possible interpretation of these results would suggest that the IAT merely measure attitudes toward the category labels, and that the exemplars comprising the categories do not contribute to the

elicited automatic attitude. That is, by manipulating the labels used to categorize targets on the IAT, Experiment 1 might not have measured automatic attitudes toward multiply-categorizable targets, but rather, attitudes toward two different attitude objects, i.e., 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 Experiments 3 and 4), we conducted a follow-up data collection that demonstrated the sensitivity of the IAT to the identity of the exemplars used in measuring automatic attitudes.

A total of 58 participants (4 were removed for excessive fast responses [more than 10% of trials faster than 300ms; Greenwald et al., 2002]) first indicated 3 liked and 3 disliked Black names and 3 liked and 3 disliked White names from a list of 45 Black and 57 White entertainers, athletes, and politicians/leaders. Participants then completed two race categorization IAT measures that were identical with the exception of the exemplars representing the categories Black and White: one IAT included the three liked Black and three disliked White names (the liked-Black task), while the other IAT included the three disliked Black and three liked White names (the disliked-Black task). In both tasks, participants categorized names using the same category labels, Black/White.

Participants showed a strong preference for White compared to Black when those categories were represented by disliked Black and liked White targets (mean IAT D = .44, SD = .27, $t(53) = 12.22$, $p < .0001$). When the categories were represented by liked Black and disliked White targets, participants showed a non-significant preference for the White targets (mean IAT D = .08, SD = .37, $t(53) = 1.7$, $p = .10$). Showing that the IAT is sensitive to the specific exemplars used to represent social groups, automatic race evaluations were significantly stronger when the Black exemplars were disliked rather than liked (and the White exemplars were liked rather than disliked), $t(53) = 5.57$, $p < .0001$.

Together with the main results of Experiment 1, these data suggest that automatic evaluations indexed by the IAT can be influenced by the exemplars comprising a social group as well as the categories into which those exemplars are classified. However, we acknowledge 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, Experiments 3 and 4 revisit the issue of contextual effects on automatic attitudes with a newly-developed, more flexible measure of automatic associations.

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 a target (e.g., Michael Jordan) will elicit a highly positive automatic attitude in one context and a neutral/negative automatic attitude in another? 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 Experiment 1 are not entirely automatic, but rather reflect the operation of more controlled, explicit mechanisms. While the IAT has been shown to be resistant to self-presentational demand (e.g., Kim & Greenwald, 1998), previous research has yet to examine such rapidly-shifting attitudes. 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 do not typically bring to bear on more traditional IAT tasks.

Experiment 2: Explicit predictions of automatic attitudes

Overview

Experiment 2 examined whether perceivers have any conscious insight that changes in evaluative context can alter the automatic evaluations they will express toward a target. 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 a space bar at the same speed they anticipated being able to respond to items. In this way, Experiment 2 measured each participant's automatic attitude toward targets as a function of categorization task (replicating Experiment 1) as well as their predicted attitude

Method

Participants

Of 32 participants, six 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. Following the rating task, they engaged in three IAT categorization tasks. These IATs were identical to those described above except for the introduction of a "prediction" block. The first IAT measured automatic evaluations toward the Coke and Pepsi colas, and served as a practice phase to introduce the IAT and prediction tasks. In this task, participants first practiced classifying good from bad words and then practiced classifying Coke-related from Pepsi-related pictures. Following these two practice blocks, participants were shown one of two dual-classification configurations (e.g., Coke+good/Pepsi+bad) and were instructed that they would be pressing one key in response to Coke-related pictures and good words and another key in response to Pepsi-related pictures or bad words. Prior to performing the actual classification task, however, participants were asked to demonstrate the speed with 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 later be able to respond during the actual classification task. A counter on the screen began at 20 and decreased by 1 upon each key response, and the reaction time between each key press were recorded. Trials 3-15 were used to calculate predicted response latencies. Following the prediction block, participants performed the actual classification task for 20 trials with a pause and then again for 40 trials. Following Greenwald et al. (2002), all trials from these blocks were used 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 classifying Pepsi- from Coke-related pictures with the new key mappings, and were then shown the remaining dual-classification configuration. Participants once again predicted their reaction times within this new classification block, practiced the actual task, and then performed the critical block.

Following this practice IAT, participants completed race (Black/White) and occupation (athlete/politician) IATs, in counterbalanced order. The details of these IATs was exactly as described for the practice IAT, except that stimuli consisted of the individually-selected liked Black athlete and disliked White politician names.

Design

Experiment 2 employed a 2 (Block: Black[or athlete]+good, Black[or athlete]+bad) x 2 (Categorization Task: race, occupation) x 2 (Response Phase: predicted, measured). 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. Data were examined following the same procedures described in Experiment 1. Positive IAT D scores reflect positive evaluations of White politicians relative to Black athletes in both the race salient and occupation salient conditions.

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 scoring algorithm described by Greenwald, Nosek, and Banaji (2002). Replicating the previous experiment, actual automatic evaluations of the positive Black athletes and negative White politicians was dependent on the salience of ethnicity or occupation ($t(25) = 2.63$, $p = .01$). Participants preferred Black athletes to the White politicians when occupation was salient ($d = -.41$, $p < .05$), but showed a slight preference for White politicians when race was salient ($d = .10$, $p = .63$). However, participants did not predict this evaluative shift in the simulation exercises ($t(25) = .93$, $p = .36$). Participant simulations showed a directional preference for Black athletes over White politicians whether occupation ($d = -.41$; $p = .03$) or race ($d = -.20$; $p = .32$) was salient.

Even though mean level effects suggest little sensitivity to changing 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 first compared simulated and actual performance for each of the tasks. Zero-order correlations between simulated and real task performance for all three tasks indicate that participants were unable to anticipate their task performance (Coke/Pepsi $r = .01$, $p = .97$; Black/White $r = .08$, $p = .68$; Athlete/Politician $r = .19$, $p = .35$). Then to more directly examine participants' ability to anticipate shifts in evaluation between the race salient and occupation salient conditions, 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 non-significant 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, while evaluations shift as a function of changes to category salience, participants are unable to predict the nature of those evaluative shifts.

Experiment 3: Context as category distinctiveness

While Experiments 1 and 2 demonstrated 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 even more distinct attitudinal shifts should be possible. That is, while the first two experiments succeeded in shifting attitudes from negative (or positive) to neutral, a truly convincing demonstration of contextual influences on automatic attitudes requires a more complete qualitative shift in evaluations of an attitude object, i.e., from highly negative to highly positive (or vice versa). The observation of limited attitude shifts (typically, from negative to neutral race attitudes) includes previous studies that demonstrate the malleability of

automatic attitudes (Dasgupta & Greenwald, 2001; Lowery et al., 2001). By using unfamiliar exemplars to represent social groups, Experiments 3 and 4 attempt to provide initial evidence of the full reversibility (e.g., from negative to positive) of automatic attitudes.

Experiments 3 and 4 also extend earlier findings that the exemplars used to represent a group can change perceivers' automatic attitudes toward that group. 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 male along the dimension of gender but, alternately, use some other construal (e.g., race or age) to individuate the female targets. Experiments 3 and 4 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 "token" African-American or female enough to engender a different evaluative context and thus produce different automatic attitudes?

To this end, Experiments 3 and 4 measured automatic attitudes toward targets that were multiply-categorizable, 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, Jackson, Dunton, & Williams, 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) as well as 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), while White males would be evaluated positively. In contrast, however, when targets' gender was salient, Black females should be evaluated positively (consistent with positive automatic attitudes toward females generally), while White males should be evaluated negatively.

Experiments 3 and 4 tested these predictions under conditions in one feature of Black females and White males was made salient through a "category distinctiveness" manipulation in which target stimuli differed from distracters along either race or gender. These experiments made use of the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001) to measure automatic attitudes toward Black female and White male targets, as a function of varying gender and race composition of surrounding target individuals.

By adopting the GNAT method, Experiments 3 and 4 did not explicitly manipulate the way that exemplars were categorized. Experiments 1 and 2 were limited by the methodological requirements of the IAT to force participants to switch the way in which they categorized targets, i.e., 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." While the data collection reported in the discussion of Experiment 1 demonstrated that the exemplar stimuli comprising 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.

The Go/No-go Association Task

The Go/No-go Association Task (GNAT) derives 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 measuring evaluations toward a single category without necessitating an explicit, contrasting category. Participants are instructed to respond within a prescribed deadline to items that fall into either of two concept+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, performance is indexed by signal detection theory's estimate of sensitivity, d' (Green & Swets, 1966). Within a concept+evaluation pairing, d' indexes a participant's ability to discriminate targets (signal) from distracters (noise). For example, when a group is strongly associated with a valence pairing, participants should more easily discriminate targets from distracters, resulting in higher d' scores than when a group is dissociated or weakly associated with a valence pairing.

Unlike the IAT, the GNAT allows measurement of automatic attitudes toward a category of targets without the requirement 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, distracter items in the GNAT can be freely manipulated (i) without the need to form a unitary category and (ii) without drawing observers' attention to changes in the set of distracter items. Experiments 3 and 4 capitalized on these advantages of the GNAT procedure by manipulating distracter items to make distinct either the gender or race of Black female and White male targets. Automatic attitudes toward Black females and White males were measured three times in blocks that made distinctive either gender, race, or neither feature of these targets.

Method

Participants

Ten White female undergraduates at Harvard University participated in exchange for $5. Participation was restricted to White females because earlier research (e.g., Carpenter & Banaji, 2000) has demonstrated that this population holds strongly positive evaluations of both females 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 3 used 15 positive words, 15 negative words, 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., Peggy, Meredith), and 15 stereotypic White male names (e.g., Brandon, Todd).

Procedure

At the beginning of each GNAT block, two labels appeared on screen, e.g., Black Female and good. Participants were obliged 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 (distracter trials). The trial was scored as correct if the participant responded to an inclusion trial before the 600-millisecond deadline (hit) or if they avoided responding to a distracter trial (correct rejection). A buzzer sounded to indicate an error if the participant failed to respond to an inclusion trial before the 600 millisecond deadline (miss) or responded to a distracter trial (false alarm). Sensitivity in each block was indexed by using the number of hits and false alarms to calculate d' (Green & Swets, 1966).

Following a set of 6 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, e.g., male+good. Each GNAT block had a companion block in which the same social group was paired with words of the opposite valence, e.g., 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. More accurate responding in the male+bad block, for example, 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 distracter 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 employed distracter items that differed from the subgroup only along the dimension of race (race construal), while another pair of blocks employed distracters that differed only along the dimension of gender (gender construal). A third pair of blocks employed items that differed from subgroup targets along two dimensions (neutral construal). Each subgroup block consisted of 65 trials, the first 5 of which were pseudo-practice trials excluded from analyses. Because category distinctiveness was established by the distracter names, these pseudo-practice trials consisted for 4 distracter names, included to ensure that category distinctiveness was manipulated from the very beginning of the block. The remaining trials comprised a random order of 30 inclusion and 30 distracter trials. Although described here as sequential pairs, participants actually completed all twenty GNAT blocks in random order.

Design

We measured automatic attitudes elicited by multiply-categorizable social targets using a 2 (group: Black female, White male) x 2 (word valence: + bad, + good) x 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) x 2 (word valence: +bad, +good) design. All factors were manipulated within-subjects.

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 distracters, we first examine automatic attitudes toward superordinate gender (female, male) and race (Black, White) targets, and then, subsequently, 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 distracters 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, while both male targets and Black targets elicited negative automatic attitudes. A Superordinate Group x Word Valence ANOVA confirmed that superordinate group targets elicited different automatic attitudes, $F(3, 27) = 31.11$, $p = 10^{-7}$. A planned contrast analysis in which we applied the lambda weight +1 to superordinate groups for which we expected to observe a positive automatic attitude (i.e., female and White) and the lambda weight –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 observed 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. Consistent with 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 neither distinctiveness conditions. 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 3-way Subgroup x Word Valence x Contextual Cue ANOVA confirmed that Black female and White males elicited contrasting automatic attitudes as a function of category distinctiveness, $F(2, 18) = 12.45$, $p = 10^{-4}$. 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 obtained separately for both Black females, $F(2, 18) = 6.13$, $p < .01$, and White males, $F(2, 18) = 7.93$, $p < .004$.

The results of Experiment 3 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 evaluative context in which a social target was evaluated. When encountering a target against the backdrop of others who differed along the dimension of gender (i.e., Black female targets against male distracters or White male targets against female distracters), perceivers construed targets according to gender, and their automatic attitudinal responses were consistent with their positive evaluations of the superordinate gender group, e.g., Black females were evaluated as positively as females in general. In contrast, encountering a target against the backdrop of others who differed along the dimension of race (i.e., Black female targets among White distracters or White male targets among Black distracters), perceivers construed targets according to race, and their automatic attitudinal responses were consistent with their evaluations of the superordinate race group, e.g., 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, Experiment 3 extended previous findings 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 comprising a category were manipulated (Dasgupta & Greenwald, 2001) or a Black confederate was present or absent (Lowery, et al., 2001), Experiment 3 did not manipulate the composition of the evaluated group. Rather, automatic attitudes were elicted in response to the same Black females and White males across all experimental conditions. In addition, Experiment 3 did not manipulate the way that exemplars were categorized. 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 rendered a target feature more or less distinctive than other individuals.

Experiment 4: 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. This overreliance on verbal stimuli may be a shortcoming of contemporary social cognition research: certainly there is abundant evidence in other areas that mental representations of pictures and words differ (e.g., Israel & Schacter, 1997; Farah, 1992; Glaser & Glaser, 1989).

Indeed, as suggested by recent theorists (Zárate & Smith, 1990; Macrae & Bodenhausen, 2000), 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 3, Experiment 4 examined whether the effects 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.

Stimuli

Experiment 4 used color images from Acme Clipart© CD and consisted of 15 positive objects (e.g., trophy, balloon, flowers), 15 negative objects (e.g., gun, poison, 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 3, except for two changes. First, clipart images were used in place of the words and names used earlier. Second, the response window was shortened to 500 milliseconds. Because images could generally be classified more quickly than verbal stimuli, the shorter response window was used to keep overall accuracy comparable to Experiment 3 (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 3, both female targets and White targets elicited positive automatic attitudes, while male targets and Black targets elicited negative automatic attitudes. A Superordinate Group x Word Valence ANOVA confirmed that superordinate group targets elicited differed automatic attitudes, $F(3, 63) = 26.68$, $p = 10^{-11}$. A planned contrast analysis (see Experiment 3) 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 3, using pictorial representations of social group members.

Automatic attitudes toward multiply-categorizable targets. Consistent with these superordinate group results, and as in Experiment 3, 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 neither construal conditions. A 3-way Subgroup x Word Valence x Contextual Cue ANOVA confirmed that Black female and White male pictorial representations elicited contrasting automatic attitudes as a function of category distinctiveness, $F(2, 42) = 8.69$, $p = 10^{-4}$. Further analysis on only the gender and race conditions confirmed the predicted pattern of results in a more focused comparison, $F(1, 21) = 7.35$, $p = .01$, $d = .59$. Finally, this pattern of different automatic attitudes obtained marginally in Black female blocks, $F(2, 42) = 2.54$, $p = .09$, and significantly in White male blocks, $F(2, 42) = 9.51$, $p = 10^{-4}$.

Using pictorial representations of social objects, Experiment 4 replicated the observation that manipulations in category distinctiveness can define the evaluative context in which a social target is evaluated and subsequently produce different automatic attitudinal responses. We note that substantially smaller automatic attitude effects were obtained using pictorial stimuli than verbal stimuli, e.g., the effect size associated with the focused comparison for multiply-categorizable targets was substantially smaller in Experiment 4 ($d = .59$) than in Experiment 3 ($d = 1.59$). Even this smaller effect, however, 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 (Nosek, et al., 2002). Given the underlying conceptual similarity between the GNAT and the IAT, it is unsurprising but reassuring that we

observed comparable difference between pictorial and verbal stimuli between Experiments 3 and 4 .

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 could be more dramatic than these recent reports suggest. Across four experiments, rapid changes in the valence of automatic attitudes were elicited in response to the same attitude objects. Experiment 1 demonstrated that automatic attitudes toward well-known Black athlete and White politicians could vary in a small period of time within an individual perceiver as a function of whether targets were categorized by race or occupation. Experiment 2 replicated these results, and further demonstrated that participants do not anticipate the effects of context on their evaluations of targets.

Experiments 3 and 4 extended these findings in three ways. First, these experiments 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) alternately elicited strongly positive evaluations in another (e.g., when gender was made salient). Previous research (Dasgupta & Greenwald, 2001; Lowery et al., 2001) had demonstrated moderation of, but not qualitative shifts in, automatic attitudes.

Secondly, Experiments 3 and 4 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 "token" African Americans among a group of White distracters was sufficient to elicit automatic attitudes toward Black females that were reflective of superordinate race attitudes, even though perceivers did not attend explicitly to the composition of the distracter stimuli. A similarly subtle manipulation in which Black females were the token women among a group of male distracters produced automatic attitudes reflective of 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.

Finally, Experiments 3 and 4 demonstrated that the identical attitude object could elicit such qualitative shifts in automatic attitudes. To greater or less degree, all previous studies on this topic (Experiment 1 above; 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) and Experiment 1 both explicitly manipulated the exemplars used to represent race categories. In contrast, our final two experiments circumvented this limitation by employing identical exemplars to represent the categories Black female and White male, demonstrating that the very same attitude objects could nonetheless elicit opposing automatic evaluations.

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 notion of a real or stable attitude from which the new attitude may be said to represent a change. Rather, the present experiments demonstrate 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 (Roediger, 1980), 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 the retrieval context. In one classic example, the mere suggestion that a yield sign had been present in an earlier scene was enough to induce participants to incorporate the sign in their memories (Loftus, Miller, & Burns, 1978).

In much the same way that such 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 two weeks, those participants who had been persuaded of the benefits of the behavior reported more frequent brushing than those who were led to believe that moderate amounts of brushing were optimal. Much like the misleading suggestion that a yield sign had been present at the scene of an accident, a change to one's explicit attitude brought about through persuasive messages provided information that 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 (Fazio, et al., 1995) have been offered up as the chameleon's mirror for social cognition. 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, 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 no one of those evaluations is more 'true' than any other.

This view is reminiscent of the well-known baseball joke in which the plate umpire suffers a moment of hesitation before calling a crucial pitch. Angrily, the batter whirls around and barks, 'Well, was it a ball or a strike?,' to which the umpire responds, 'What do you mean,

was it a ball or a strike? Hell, it ain't nothing until I call it!' In the same vein, we suggest that no hidden, platonic form of automatic attitudes (or baseball pitches) exists waiting to be measured (or called). 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.

---

[1] Because of a procedural variation in the practice blocks, the IAT D effect for Experiment 1 only the 40 trials in the two 'critical' blocks were used for analysis. In the two follow-up studies described in the discussion of Experiment 1, trials from both 'practice' and 'critical' blocks were used following that recommendation by Greenwald et al. (2002).

References

Abelson, R. P. (1988). Conviction. American Psychologist, 4, 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 D.C.: American Psychological Association.

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

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., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. Journal of Personality and Social Psychology, 70, 1142-1163.

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: Lawrence Erlbaum Associates, Inc.

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

Cohen, J. (1988). Statistical power analysis for the Behavioral Sciences. Lawrence Erlbaum Associates: Hillsdale, NJ.

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

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. Psychological Science, 12, 163-170.

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.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. Journal of Experimental Social Psychology, 36, 316-328.

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

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. Journal of Experimental Social Psychology, 22, 22-37.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. Journal of Experimental Social Psychology, 33, 510-540.

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

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., 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.

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.

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 (Eds.), Advances in Experimental Social Psychology (pp. 1-74). San Diego, CA: Academic Press, Inc.

Gilbert, D. T., & Hixon, G. J. (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. (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. (2002). Scoring procedures to improve implicit association test measures. Manuscript submitted for publication.

Gross, S. R. & Ellsworth, P. C. (2001). Second thoughts: Americans' views on the death penalty at the turn of the century. In S. P. Garvey (Ed.) The modern machinery of death: The future of capital punishment in America. 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.

Lemm, K. M., & Banaji, M. R. (1998). 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 & 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, England: 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: The 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: Morgan Kaufmann, Inc.

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 website. Group Dynamics, 6, 101-115.

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. Journal of Personality and Social Psychology, 59, 475-486.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114, 510-532.

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.

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: The hidden coasts to women of a kinder, gentler image of middle managers. Journal of Personality and Social Psychology, 77, 1004-1010.

Rudman, L. A., Greenwald, A. G., McGhee, D. E. (in press). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. Personality and Social Psychology Bulletin.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In Gilbert, D. T., Fiske, S. T., & Lindzey, G. The Handbook of Social Psychology, Vol. 1 (4th ed.). (pp. 143-179). Boston, MA, US: McGraw-Hill.

Shih, M., Pittinsky, T. L., Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. Psychological Science, 10, 80-83.

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. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), Handbook of social psychology (4th ed, Vol. 1, pp. 446-496). New York: McGraw-Hill.

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

Figure 1. Response latencies in occupation and race tasks by evaluative pairings. (Experiment 1). Error bars represent the standard error of the mean

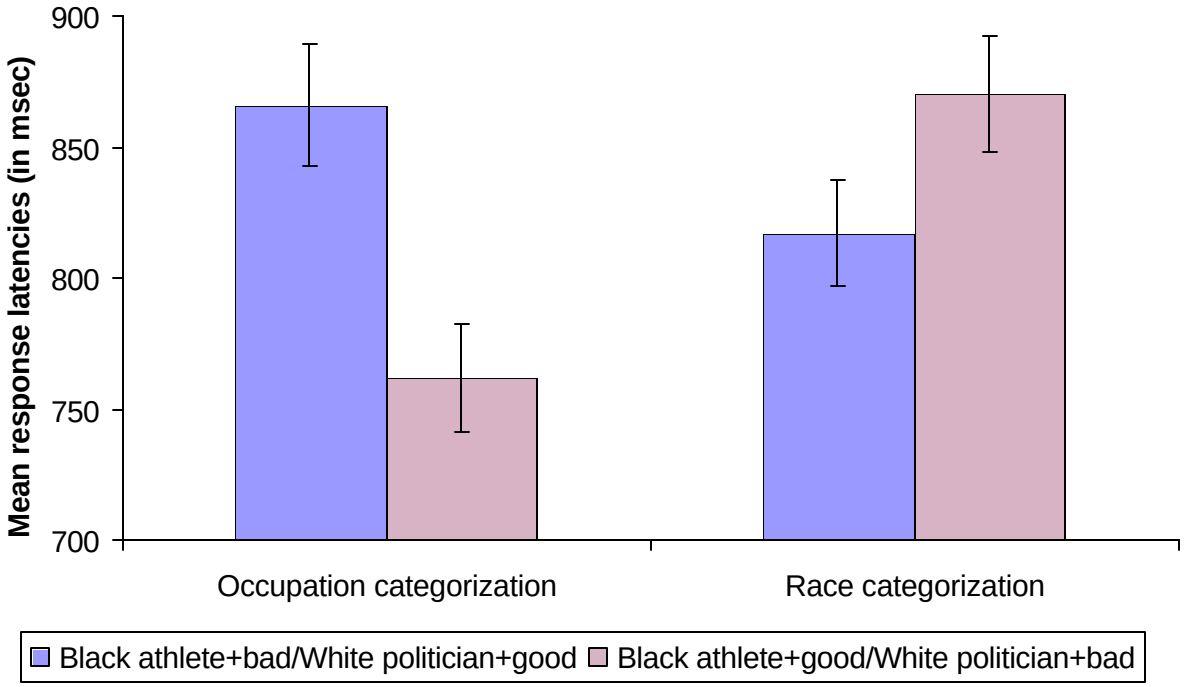


Figure 2. Sensitivity (d') by evaluative pairings for superordinate groups obtained in Experiment 3 (top panel) and Experiment 4 (bottom panel). Higher sensitivity scores indicate a stronger association between target group and evaluation. Error bars represent the standard error of the mean.

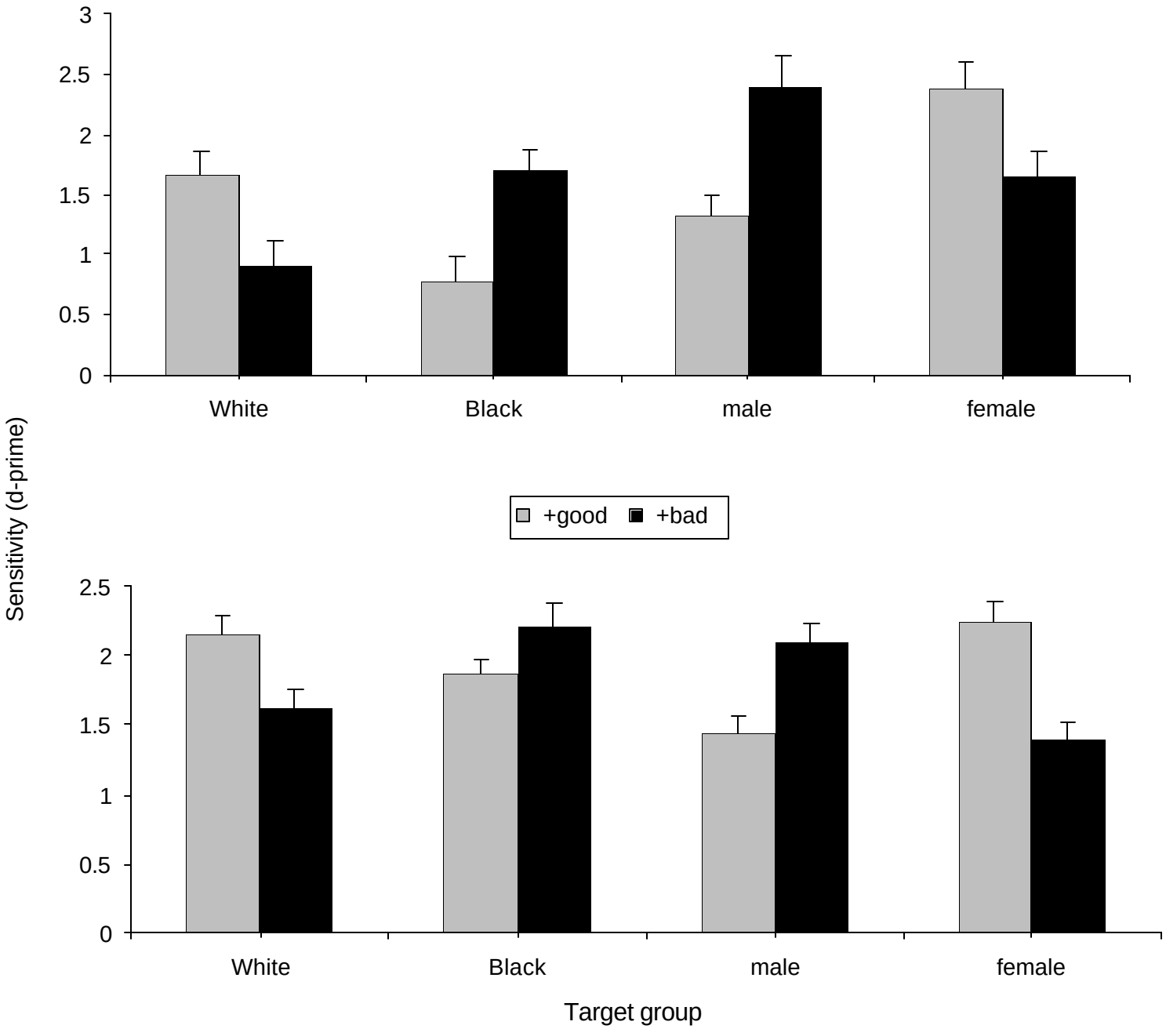


Figure 3. Sensitivity (d') difference scores for Black female and White male targets in Experiment 3 (top panel) and Experiment 4 (bottom panel). Values represent the difference between a subgroup paired with positive words and with negative words. Positive values indicate a positive automatic attitude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

