Psychological Inquiry
2004, Vol. 15, No. 4, 257–278

TARGET ARTICLE

Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?"

Hal R. Arkes
Department of Psychology
Center for Health Outcomes, Policy, and Evaluation Studies
Ohio State University

Philip E. Tetlock
Haas School of Business
University of California, Berkeley

Measures of implicit prejudice are based on associations between race-related stimuli and valenced words. Reaction time (RT) data have been characterized as showing implicit prejudice when White names or faces are associated with positive concepts and African-American names or faces with negative concepts, compared to the reverse pairings. We offer three objections to the inferential leap from the comparative RT of different associations to the attribution of implicit prejudice: (a) The data may reflect shared cultural stereotypes rather than personal animus, (b) the affective negativity attributed to participants may be due to cognitions and emotions that are not necessarily prejudiced, and (c) the patterns of judgment deemed to be indicative of prejudice pass tests deemed to be diagnostic of rational behavior.

> "There is nothing more painful to me at this stage in my life than to walk down the street and hear footsteps and start thinking about robbery. Then look around and see somebody White and feel relieved."
>
> —Jesse Jackson

Survey research on racial attitudes in the general American population has shown a trend with potentially profound political implications: Overt White hostility toward African Americans began to decline markedly in the early 1960s and by the 1990s had reached historic lows (Schuman, Steeh, Bobo, & Kyrsan, 1997; Sniderman & Carmines, 1997). Whereas Americans were once deeply divided over whether African Americans and Whites should be allowed to drink from the same fountain, sleep in the same hotel room, attend the same schools, or intermarry, there is now close to consensus at the level of both mass and elite opinion that de jure segregation is unacceptable.

Survey results notwithstanding, the most influential observers of race relations in the United States—prominent academics, journalists, and political figures—are deeply divided over the prospects for overcoming traditional racial divisions (Black, 2002). On one side are those who see no relief in sight from continuing conflict between African Americans and Whites (Bell, 1992; Hacker, 1995) and an apocalyptic few who predict a coming race war (Rowan, 1996). These observers trace the problem to the pervasiveness and tenacity of White prejudice toward African Americans: "Racism lies at the center, not the periphery, in the permanent not in the fleeting, in the real lives of black and white people, not in the caverns of the mind" (Bell, 1992, p. 208). On the other side are those who paint a considerably more upbeat picture (Jacoby, 2000). Thernstrom and Thernstrom (1997) argued that African Americans have made substantial gains—economic, educational, and health—in overcoming the effects of past prejudice. They attribute pockets of persisting inequality not to White racism but rather to racial gaps in educational attainment, to the rise in African-American crime, and to the structure of the African-American family.

Most social psychologists who study racial attitudes seem to line up with the pessimists. They are skeptical of the depth and sincerity of the changes in racial attitudes shown in representative-sample surveys. Com-

menting on such surveys, they conjecture that it is only "plausible that prejudice was on the decline"; alternatively "it was also possible that prejudice was taking more subtle and insidious forms to which the available assessment methods were largely insensitive" (Brauer, Wasel, & Niedenthal, 2000, p. 79). Others echo this skepticism about whether the purported steep decline in prejudice is genuine (e.g., Dovidio & Gaertner, 2000; Greenwald & Banaji, 1995, p. 15; Rudman, Ashmore, & Gary, 2001.) Whites, they suggest, have learned to say the right thing, but they have not truly internalized the egalitarian ideals that would justify calling them nonracist (Jackman & Jackman, 1983).

To rise to the measurement and indeed moral challenge, social psychologists have developed new, more subtle techniques for tapping into the unexpressed and perhaps even unconscious racism that Whites may still harbor toward African Americans. These new techniques fall into several categories, including: (a) self-report scales designed to assess more indirect forms of hostility, such as resentment of welfare abuse and dislike of busing and affirmative action, and that can be justified by appeals to traditional American values such as self-reliance and individual responsibility. These controversial measures go by the names of modern or symbolic racism (Kinder, 1986; Sears, Sidanius, & Bobo, 2000; Sniderman & Tetlock, 1986); (b) unobtrusive indicators designed to pick up oblique manifestations of hostility that might manifest themselves when people think the sentiment cannot be traced to them personally (Kuklinski et al., 1997) or that might "leak out" in interpersonal encounters in the form of gaze aversion, physical distancing, facial expressions, and tone and content of speech (Crosby, Bromley, & Saxe, 1980); (c) most recently and arguably the most methodologically and theoretically sophisticated assault on the problem to date are measures of implicit prejudice that explore the associative linkages in memory between words with positive or negative valence and racial stimuli (Fazio & Olson, 2003).

The first two literatures have already been exhaustively reviewed and debated. But the third literature has not been subjected to the same critical scrutiny. This article redresses this omission. Our central points are these: (a) Measures of implicit prejudice that are based on associations between negative stimuli and minority group images or names may reflect shared cultural stereotypes rather than personal animus, (b) any *affective negativity* that can be personally attributed to participants may be due to cognitions and emotions that are not necessarily *prejudiced*, and (c) the patterns of judgment deemed to be indicative of prejudice pass tests deemed to be diagnostic of rational behavior.

The article is organized into three sections. First, we present a brief overview of the principal implicit associative measures, with special focus on the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and the affective priming paradigm (Fazio, Jackson, Dunton, & Williams, 1995). This section discusses the conceptual and empirical grounds on which researchers have justified their claims that particular assessment paradigms provide bona fide as opposed to bogus pipelines to attitudes that people are either unwilling to admit or unaware of possessing.

Second, we examine sources of construct-operation slippage in research that aims to measure implicit prejudice by adapting reaction time (RT) measures that memory researchers have used to explore properties of associative networks. Although researchers have made a strong case that automatically activated affective negativity plays a mediational role in these studies, researchers have been too quick to make the inferential leap from implicit associations to implicit attitudes, and then from implicit attitudes to value-laden characterization of those attitudes as prejudice. To appreciate why the case has yet to be made that implicit prejudice is prejudice, it is critical to be clear about definitions. One natural place to turn for guidance is to the classic writers on the topic. For example, Allport (1954) defined prejudice as "thinking ill of others without sufficient warrant" (p. 7). Kelman and Pettigrew (1959) defined group prejudice "as having two components: hostility and misinformation" (p. 436). Krech, Crutchfield, and Ballachey (1962) defined prejudice as "an unfavorable attitude toward an object which tends to be highly stereotyped, emotionally charged, and not easily changed by contrary information" (p. 214).

To satisfy definitional burdens of proof, one must show that the affective negativity tapped by implicit associative measures does not merely reflect culturally shared associations that might arise in any society with widespread inequality. First, one must show that the affective negativity is functionally intertwined with beliefs that indiscriminately attribute negative qualities to group members. Second, the negative affect must be grounded in hostility rather than in other aversive arousal states such as guilt, shame, embarrassment, or social anxiety that might plausibly accompany interracial relationships in a society trying to overcome a long, grim history of interracial tension. Third, the negative affect must be unwarranted as well as resistant to change in the face of new evidence. Fourth, the affect must be truly negative as opposed to merely less positive than the affect one has toward other groups. Insofar as implicit prejudice effects stem from widely shared cultural stereotypes rather than genuinely endorsed beliefs, guilt or shame rather than racial animus, accurate statistical associations rather than unwarranted conclusions, or relatively less positive affect rather than negative affect, then it is appropriate to question whether those effects bear on the concept of prejudice traditionally defined.

This article identifies three reasons why the gap between associationist and attitudinal interpretations proves exceptionally difficult to bridge: (a) the problem of distinguishing between RT facilitation and inhibition effects grounded in personal attitudes versus shared cultural stereotypes (that the respondent supposes that others believe but does not endorse and may even deem repugnant); (b) the problem that we call the "conceptual ambiguity of affective negativity"—a wide range of ideological-emotional perspectives on racial inequality (from the far left to the far right) are compatible with affective negativity; and (c) the justifiability of calling people prejudiced who pass classic correspondence and coherence tests of rationality (Hammond, 1996). If, for example, the base rate of criminality is higher for African Americans than for Whites, can Rev. Jackson be deemed prejudiced if he is relieved when he realizes that the footsteps behind him are those of a White?

Finally, the article closes by putting research programs on implicit prejudice in broader psychological and historical perspective. Examining evolving conceptions of prejudice in American history, we argue that work on implicit prejudice sets the threshold for making attributions of prejudice at an unprecedented low level. This methodological move raises largely ignored questions about how psychologists should manage the murky trade-offs between the risks of mistakenly accusing the unprejudiced and those of mistakenly exonerating the prejudiced. We propose instead that psychologists should approach racial cognition like any other aspect of judgment and choice. Insofar as researchers insist on making normative assessments, they should apply their benchmarks of rationality consistently. And if researchers insist on injecting special standards into particular content domains, then they should be explicit about the exact roles that their value judgments play in attributing prejudice to research participants.

Overview of Methods

The two primary methods used to detect negative associations to racial groups are the affective priming technique (Fazio et al., 1995) and the IAT (Greenwald et al., 1998). We acknowledge that other methods have been used (e.g., Devine, 1989; Kawakami, Dion, & Dovidio, 1998; Locke, MacLeod, & Walker, 1994; Nosek & Banaji, 2001; von Hippel, Sekaquaptewa, & Vargas, 1997), but because affective priming and the IAT are by far the most prominent methods, our discussion concentrates largely on them.

Affective Priming

Priming has been a common tool used by cognitive psychologists for decades to investigate the semantic properties of verbal material (e.g., Meyer & Schvaneveldt, 1971). In a typical priming experiment, a prime word is presented before a target word. The participant in the study is supposed to react to the target word quickly, usually by pressing a key or by pronouncing the target. The semantic relation between the prime and the target is the topic of interest to the investigator. To the extent the prime influences the response to the target the two stimuli are deemed to be associated. Some primes facilitate the response to the target word, as when the prime is *bread* and the target is *butter*. This facilitation is thought to be due to the automatic activation of the target by the earlier presentation of the highly related prime.

Fazio, Sanbonmatsu, Powell, and Kardes (1986) reasoned that this priming technique could be extended to attitudes. For example, a prime might be a negatively valenced word such as *murderer*. If the subsequently presented target is evaluatively congruent, such as the word *evil* would be, the evaluation of the target would be accomplished more quickly than if the target were *happy*. In other words, RT to hit either the *positive* or *negative* key would be quicker with congruent than incongruent prime-target pairs.

Fazio et al. (1995) extended this procedure to include pictures of White and African-American faces as the primes. The targets were adjectives with either a positive (e.g., *wonderful*) or negative (e.g., *annoying*) connotation. Participants had to push either the key labeled *good* or the one labeled *bad* as quickly as possible. The principal finding was that among the White respondents, RT to the good words was quicker following presentation of the White faces, and RT to the bad words was quicker following presentation of the African-American faces. Just as *bread* facilitated RTs to the target *butter* due to their close semantic relationship, Fazio et al. (1995) concluded that the White participants in this study associated positivity more closely with Whites and negativity more closely with African Americans. These results led the authors to suggest that this technique represented an unobtrusive measure of racial attitudes.

An important feature of the affective priming technique is the very short interval between the onset of the prime and the target. Because the interval is usually quite short—often 300 millisecs—there is a strong implication that the cognitive processes that cause the affective priming effect must be due to automatic semantic activation that is beyond the participant's conscious control. Further evidence for the automaticity of the processing is provided by studies such as that by Wittenbrink, Judd, and Park (1997), in which the primes were presented subliminally. Obviously no conscious control can occur in such instances.

IAT

Greenwald et al. (1998) introduced the IAT, which represents another attempt to exploit semantic activation to assess attitudes that one is not aware of or willing to admit that one possesses. Participants are asked to respond as quickly as possible to a series of tasks. In the first task, participants might be asked to respond to the left key when a first name associated with African Americans is presented (e.g., Latonya) and to respond to the right key when a first name associated with Whites is presented (e.g., Betsy). In the second task participants are asked to respond to two different keys when positive versus negative words are presented. These tasks are combined in two subsequent sessions. For example, in one session, the participants should respond to the left key when a White name or positive word is presented and to respond to the right key when an African-American name or negative word is presented. This represents the "compatible condition." In a final task, participants must respond to the left key when an African-American name or positive word is presented and to respond to the right key when a White name or negative word is presented. This represents the "incompatible condition." For the "vast majority" of White participants (Banaji, 2001, p. 137), RTs are faster in the compatible than in the incompatible condition. This led Greenwald et al. (1998) to conclude that among the White college students they tested "there was a considerably stronger association of White (than of African American) with positive evaluation" (p. 1474). However more explicit measures of prejudice suggested that these same participants did not harbor negative attitudes toward African Americans. This result fostered the conclusion that there was a dissociation between the benign attitudes openly expressed by the respondents and the more sinister implicit attitudes detected by the IAT.

Many researchers who use these two methodologies do not mince words in discussing the implications of their findings. For example, Greenwald and Nosek (2001) stated that the IAT can detect "unrecognized mental residues of a racist culture" (p. 86). What cognitive mechanisms might underlie these troubling findings?

Suggested Theoretical Mechanisms

Three mechanisms have been suggested as the cognitive bases for the affective priming and IAT results: association, response competition, and cultural stereotypes.

The association mechanism is predicated on the assumption that related items are located closer together in semantic memory than are unrelated items. Thus, if a person has a close association between negative words and words linked to a minority group, then words drawn from these two categories will be jointly accessed more quickly than would words drawn from a positive list and the same minority-group category. This differential reaction time is the index of implicit prejudice.

The response competition explanation is predicated on the fact that as soon as one is exposed to the first of two stimuli, one's reaction to that stimulus is initiated. If the second stimulus is congruent with the first, then the already-begun reaction to the first stimulus can proceed without interruption. However if the second stimulus is incompatible with the first, then the reaction to the first stimulus must be truncated, and a new response must be initiated. This is why a positive word following a White face is responded to more quickly by Whites than is a positive word following an African-American face. Presumably among White participants in such research, the responses to a positive word and a White face both require the same categorization response, whereas the responses to a positive word and an African-American face do not.

The associative and response competition mechanisms are both predicated on the core assumption that the people exhibiting prejudiced data maintain at some level the negative attitudes uncovered by either the affective priming or IAT procedures. Due to the fact that the prejudice uncovered by these techniques is implicit, the negative attitudes might not be consciously avowed or openly endorsed. In everyday language, to say that someone endorses a point of view is to say that person is aware of the opinion and is prepared (to some degree) to rise to its defense if challenged. People consider it reasonable to hold each other morally accountable for the views they endorse and to make unflattering character attributions—dumb, insensitive, selfish—for views they deem indefensible. In the implicit prejudice literature, the prejudicial attitude needs to be understood as occurring at some deeper level—perhaps not easily accessible to consciousness—of psychological functioning (Wilson, Lindsey, & Schooler, 2000).

At what level do implicit prejudice researchers think the attitude might lie? At a Harvard University Web site pertaining to the IAT (https://implicit.harvard.edu/implicit/demo/racefaqs.html), we read that

> Social psychologists use the word 'prejudiced' to describe people who endorse or approve of negative attitudes and discriminatory behavior toward various out-groups. Many people who show automatic white preference on the Black-White IAT are not prejudiced by this definition. These people are apparently able to function in non-prejudiced fashion partly by making active efforts to prevent their automatic White preference from producing discriminatory behavior.

We conclude that a person can refrain from explicit prejudice despite having implicit prejudice, but this

might require a vigilant effort to prevent the implicit prejudice from manifesting itself in overt behavior. Nosek, Banaji, and Greenwald (2002b) seemed to agree: "A stereotype may be maintained outside conscious awareness although it is neither wanted nor endorsed consciously, yet still influence both consciously and unconsciously held attitudes" (p. 55). Based on this quote, we suggest that these authors agree that implicit prejudice either qualifies as a genuine attitude or influences other attitudes.

Many contributors to the research literature take it almost for granted that their colleagues believe that that the IAT and affective priming methodologies reveal genuine prejudicial attitudes. Devine, Plant, Amodio, Harmon-Jones, and Vance (2002) stated: "Throughout this article, we deliberately avoid referring to the race bias indicated by implicit measures as a *prejudiced* response or an indicator of racial attitude, though many take these indicators to reflect racial attitudes (e.g., Fazio et al., 1995; Greenwald et al., 1998)" (p. 837). Devine et al. (2002) asserted that many researchers believe that the IAT and affective priming methods assess genuine racial attitudes. If we accept the customary definition of attitude as "a positive or negative evaluation of an object" (Franzoi, 2000, p. 148), then many researchers believe that the IAT and affective priming techniques do reveal a person's negative evaluations of a minority group. Similarly Payne, Lambert, and Jacoby (2002) wrote that "research showing the ubiquity of unconscious prejudice is both intriguing and disturbing. The fact that people who explicitly espouse egalitarian values may, nonetheless, be prejudiced carries the specter than anyone might be an implicit bigot without the power to know or control his or her own biases" (p. 384). Note that the authors said that the research on unconscious prejudice detects people who are "prejudiced" and "implicit bigots." Again, we suggest that these authors believe that people in whom unconscious prejudice exists are the owners of negative attitudes; they possess bigoted views of minority groups.

These examples could be multiplied. McConnell and Leibold (2001) stated: "The IAT has become a widely used instrument to measure attitudes in general, and prejudices toward groups in particular" (p. 435). Rudman et al. (2001) in their assessment of the purported benefits of diversity education reported that people who took such classes "showed a significant reduction in their implicit prejudice" which had been detected by the IAT (p. 865). Banaji (2001) stated that those who take the IAT note "a lack of synchrony between our view of ourselves as unbiased ('I am a morally good person') and evidence of ourselves as biased ('I am not a morally good person')" (p. 137). Fazio et al. (1995) suggested that the affective priming methodology "form[s] a valid, unobtrusive measure of attitudes toward African Americans" (p. 1019). Greenwald et al. (1998) interpreted IAT results as indicating "implicit racism" (p. 1476). These quotes lead us to surmise that there is a widespread assumption that people who manifest certain results on the affective priming or IAT methodologies are indeed guilty of harboring anti-African-American prejudice or racist attitudes—views that, if they were to take conscious form, would be grounds for censure in a society that (at least) formally endorses racially egalitarian norms.

The cultural stereotype mechanism differs sharply in that it is not necessary for the person exhibiting the prejudiced data to endorse at any level the reprehensible views uncovered by the affective priming or IAT methodologies. Our own view is that there is insufficient justification for labeling people as prejudiced if they exhibit certain patterns of response-time facilitation. Hence the core assumption of both the association and response competition explanations renders the two mechanisms equivalently opposed to our position. We therefore do not present any of the research bearing on the relative merits of the association and the response competition mechanisms in explaining the affective priming and IAT data (e.g., De Houwer, Hermans, Rothermund, & Wentura, 2002). Whether one mechanism is more facile in explaining pronunciation tasks versus key-press tasks is irrelevant for our purposes.

Obstacles to Bridging the Conceptual Gap Between Implicit Associations and Implicit Prejudice

Shared Stereotypic Associations: Are We Measuring Associations That Respondents Believe or Associations That Respondents May Reject but Are Aware That Others Hold?

To make the conceptual connection between data yielded by the affective priming and IAT paradigms on the one hand and implicit prejudice on the other, it is necessary to assume that the negative affectivity uncovered by the affective priming and IAT methods taps into prejudicial attitudes that the research participants hold at some conscious or unconscious level. These methodologies purportedly measure attitudes, and attitudes are generally believed to connote endorsement of a particular evaluative stance toward the world (Ajzen, 1987). However several authors have suggested that the participants tested using these methodologies are not providing responses indicative of their attitudes but instead are responding to cultural stereotypes to which they have been exposed but with which they may or may not agree (e.g., Karpinski & Hilton, 2001). For example, Greenwald et al. (1998) found that Korean and Japanese participants yielded IAT results that suggested that each held a negative view of the other.

However these results were stronger for those persons who were more steeped in the Asian culture. Greater awareness of one's Asian heritage would also be related to greater awareness of the group's mutual stereotypes of each other. Knowledge that my group thinks that the other group has more of trait X is all that is needed to generate an affective priming result in which the conjunction of the other group and X yields faster reaction times than the conjunction of my group and X. Whether I actually endorse this stereotype may be irrelevant. If I am aware of the cultural stereotype, I have all the cognitive software that I need to manifest prejudice on the IAT. In their review of the effects of stereotype activation on behavior, Wheeler and Petty (2001) concurred that awareness of a stereotype can influence one's behavior even if one disagrees with it. For example, African Americans who vigorously reject their own cultural stereotype may nevertheless find that it is detrimental to their performance on an academic test (Steele, 1997). In fact, Nosek, Banaji, and Greenwald (2002a) found that African-American respondents revealed preference for White over African American in their IAT data, albeit a less pronounced preference than was manifested by Whites. If we assume that the thousands of African Americans who took the Web-based IAT are not prejudiced against their own race, then these data strongly suggest that culturally stereotypic associations, which they do not endorse, are responsible for this result. We acknowledge the fact that several studies have shown that minority group members manifest some ambivalence concerning members of their own group (e.g., Jost, Pelham, & Carvallo, 2002). However a civil rights leader (Rev. Jackson) and Afro-American cab drivers who manifest reluctance only to pick up Afro-American male customers (Koren & Williams, 1999) are most unlikely to be prejudiced against their own race. Studies using other methodologies are congruent with the cultural stereotype explanation. Correll, Park, Judd, and Wittenbrink (2002) performed four experiments in which participants viewed a videogame in which objects were held in the hand of African-American or White target persons. Some of the objects were weapons, and some were innocuous items. In all studies persons were asked to "shoot" armed targets and not to "shoot" unarmed ones. In all studies the participants were more likely to shoot (or more quickly shoot) an armed target if the target was African American, but were less likely to shoot (or less quickly shoot) an unarmed target if the target was White. This "shooter bias" was equivalent in samples of White and African-American participants (Study 4). Furthermore, the magnitude of the bias did not vary with personal racial prejudice. It did vary with perceptions of the cultural stereotype, as measured by the participants' estimate of the prevalence of dangerousness, violence, and aggressiveness most White Americans would perceive among African Americans (Study 3). Again, being aware of the existence of the stereotype but not endorsing it is sufficient to engender biased responding, in the case of the Correll et al. (2002) studies, using explicit rather than implicit prejudice.

The cultural stereotype viewpoint also helps explain the relatively modest correlations between measures of implicit and explicit prejudice. For example, Brauer et al. (2000) reviewed studies in which a relation between implicit and explicit prejudice had been tested. The median correlation among the 21 tests was .24, with a range of –.07 to .60. Studies published since that review have also found low correlations between implicit and explicit prejudice (e.g., Karpinski & Hilton, 2001; Lowery, Hardin, & Sinclair, 2001; Wittenbrink, Judd, & Park, 2001). The most conspicuous exception is Cunningham, Preacher, and Banaji (2001) who showed that a latent variable approach that takes measurement error into account yields a correlation of .45 between an implicit attitude latent construct and the explicit attitude construct, which puts this result toward the upper end of the correlations reviewed by Brauer et al. (2000). However, it is worth stressing that the explicit measure, McConahay's (1986) Modern Racism Scale, hardly qualifies as a gold standard for prejudice. Critics have argued that, judging the scale on manifest item content and correlations with other measures, the scale can be more plausibly viewed as a measure of traditional values and conservative policy preferences (Sniderman & Tetlock, 1986).

The preferred explanation for these weak or inconsistent correlations among advocates of implicit measures of prejudice invokes the superiority of the implicit measures: Such measures are free of the impression management and other obfuscating strategies that contaminate the measures of explicit prejudice (Fazio et al., 1995). An alternative explanation for why the correlations are not stronger is, however, that different psychological constructs are being assessed: A person can be aware of cultural stereotypes, as indicated by the measure of implicit prejudice, but reject those same stereotypes, thereby manifesting low explicit prejudice.

The research program of Devine (1989) fosters a similar conclusion. In her first experiment Devine (1989) ascertained that both high and low prejudiced persons were equally knowledgeable of the cultural stereotype of African Americans. In her second study she primed participants with visually presented words, which were presented so that they were not recognizable on a subsequent memory test. Nevertheless, these words, which cued an African-American stereotype, influenced participants' judgments of the hostility present in a subsequent scenario. Those who had been primed with 80% stereotype-related words perceived more hostility than did those who were primed with 20% stereotype-related words. Most important, this re-

sult was equally strong among both high and low scorers on the Modern Racism Scale (McConahay, 1986). Devine (1989) concluded: "Study 2 suggested that automatic stereotype activation is equally strong and equally inescapable for high- and low-prejudice subjects" (p. 15).

We suggest that the same conclusion applies to affective priming and the IAT. Both high- and low-prejudice participants are aware of the stereotype, as Devine (1989) showed in her first study. When the IAT or the affective priming methods force both types of people to respond quickly without conscious monitoring of their responses, the "automatic stereotype activation," which is "equally strong and inescapable" (Devine, 1989, p. 15) in the two groups, makes nearly everyone appear to be "implicitly prejudiced." Devine was able to detect differences between low- and high-prejudiced persons in her third study in which quick reactions were not required. In Study 3 persons were asked to "list all of their thoughts" about African Americans. When low-prejudiced persons thus had the time needed to behave consciously in a nonprejudicial manner, their responses differed from those of the high-prejudiced persons who behaved in the same general manner regardless of time constraints. Note again that the presence of prejudice on explicit tasks does not necessarily bear any relation to the level of prejudice detected by implicit tasks.

Fazio et al. (1995) offered a possible objection to the cultural stereotype explanation of the affective priming and IAT results: "If … the shared cultural stereotype is activated in the presence of a minority group, one would expect little meaningful variation in the pattern of facilitation across participants. On the other hand, if it is one's personal evaluation that is activated in the presence of a minority group member, the variation across participants would be more substantial and predictive of race-relevant behaviors" (p. 1095).

We respectfully disagree with this analysis. There is no reason why people could not have varying levels of awareness of various facets of cultural stereotypes that are unevenly distributed throughout the population at large. For example, although the aforementioned study by Greenwald et al. (1998) found that Korean and Japanese participants yielded IAT results that suggested each group responded more quickly when negative words were associated with the other group than when positive words were associated with the other group, these results were stronger for those more steeped in Asian culture. Greater awareness of one's Asian heritage would also be related to greater awareness of the stereotype with which each of the two groups characterized the other. Therefore we suggest that interperson variability in the IAT or affective priming results is not necessarily contrary to the cultural stereotype explanation.

Note that the cultural stereotype explanation of the affective priming and the IAT results does not deny the role of associations or response competition in generating the results found within those paradigms. The cultural stereotype explanation only denies that those methods necessarily tap endorsement of prejudice.

We should also point out that of the two main implicit methodologies—the IAT and affective priming—the former may be more susceptible to the cultural stereotype explanation. De Houwer (2001) has shown that within an IAT paradigm, a participant's responses to members of a category are driven largely by the participant's evaluation of the category. Thus one's evaluation of *Princess Diana* is heavily influenced by one's evaluation of *British*. This feature of responding to the category when one is presented with a specific stimulus might make the IAT somewhat more vulnerable to the influence of category labels compared to the affective priming procedure, according to Fazio and Olson (2003). Thus, the culture's categorization of particular stimuli might seem to play a somewhat larger role in IAT experiments compared to affective priming ones.

Results consistent with this notion were presented by Olson and Fazio (2004). These authors contend that the results of a traditional IAT are contaminated by extrapersonal associations. These are associations that may or may not be personally accepted or endorsed by an individual taking the IAT but that instead exist due to culturally shared *environmental associations*, to use the phraseology of Karpinski and Hilton (2001). In an effort to reduce the influence of such extrapersonal or environmental associations, Olson and Fazio made a few modifications to the traditional IAT procedure to create a more personalized IAT. First, whereas the traditional IAT procedure uses the category labels *pleasant* and *unpleasant*, Olson and Fazio used the labels *I like* and *I don't like*. Categorization of words or images using the former pair of labels carries a normative implication, in that the participant in the experiment may think that there is a correct way to classify a stimulus; any entity must be either *pleasant* or *unpleasant*. On the other hand, *liking* is entirely subjective and carries no normative implication. Thus participants classifying stimuli using the latter pair of labels are less likely to think there is an environmentally driven way to categorize any stimulus. Second, rather than using universally pleasant (e.g., *love*) or unpleasant (e.g., *bombs*) stimuli, Olson and Fazio used more ambiguous stimuli (e.g., *coffee*) to reduce the participants' belief that there is any normatively correct classification scheme based on environmental contingencies. Finally, whereas the traditional IAT procedure provides feedback when a participant makes a classification error, the modified personalized procedure used by Olson and Fazio did not. Again, this should reduce

the participants' belief that the environment has unambiguously divided the presented stimuli into separable categories.

The personalized IAT procedure resulted in significantly lower prejudice against Blacks than did the traditional IAT procedure. Olson and Fazio (2004) argued that the personalized IAT prompts participants to concentrate more on their own attitudes and less on external considerations. We suggest that the personalized IAT represents a significant step in the right direction in that it helps to distill out the cultural stereotypes that we contend may contaminate the traditional IAT results.

Yet another effort to try to remove cultural stereotypic knowledge from measures of implicit prejudice was made by Lepore and Brown (1997). They showed in their Experiment 3 that when stereotype-related words were used as subliminally presented visual primes, differences between high- and low-prejudice people were minimal when both groups were asked to evaluate a stimulus person following the primes. This suggests that both high- and low-prejudiced people are equally aware of the stereotype that links stereotype-related words to the relevant minority group. However in their Experiment 2 Lepore and Brown showed that when category labels or neutral associates of those labels were used as the primes, then high- and low-prejudiced people did differ in their subsequent evaluation of a stimulus person. Lepore and Brown conjectured that this result may be diagnostic of stereotype endorsement, which does differ between high- and low-prejudice groups, rather than mere stereotype knowledge, which is common knowledge. If they are correct, then the IAT might not be well suited to assess pure stereotype endorsement, because the words to be paired with African-American or White names include such stereotype-related items as *poverty* and *prison* (Greenwald et al., 1998, p. 1479). Such primes tap mere stereotype knowledge, not endorsement, according to Lepore and Brown.

The first tier of our argument is thus that the IAT and affective priming methods measure associations, but we question whether these associations tap prejudicial attitudes. We grant that true bigots may show concordance between implicit and explicit prejudice, but our central point is that the overall modest correlations between the two should not be used to indict nonbigots—including members of minority groups—who are merely aware of historically rooted cultural stereotypes and, as a result, manifest incriminating RT results. We now turn to the two other tiers: (a) The affective negativity posited to underlie the RT data is open to alternative explanations other than racial animus, and (b) affective negativity may be grounded in inferences that pass standard benchmarks of rationality.

The Conceptual Multidimensionality of Affective Negativity: The Parable of the Two Jesses

Imagine two respondents who have markedly different associative-network structures for encoding information about African Americans. One respondent is politically sympathetic to a left-liberal policy agenda, believes that racial discrimination is an ongoing, not just a past, problem, supports aggressive affirmative action and even racial reparations, and believes that the major reason why African Americans make less money and have lower levels of educational achievement in America today derive directly from the historical legacy of slavery, continuing exploitation, and segregation. This respondent thinks that progress during the past half century has been too little and too slow. He thinks that Whites are intransigently hostile to African Americans. The other respondent is sympathetic to the right-conservative agenda and believes that just as other minority groups have had to work their way up the American success ladder, so it should be for African Americans. This respondent disapproves of affirmative action and rejects the idea of racial reparations. This respondent believes that the primary cause of African-American economic and educational inequality in America today is internal to the African-American community: the widespread abdication of personal responsibility within inner-city communities and the surge in the late twentieth century of out-of-wedlock births.

We might call our two respondents the two Jesses to represent those eponymous figures of late twentieth century American politics: Jesse Jackson and Jesse Helms. Although the two figures disagree profoundly on certain political issues, they do agree about certain basic facts. They agree that the African-American family is in trouble, that African-American crime rates are far too high, and that African-American educational test scores are too low. They experience varying mixtures of sorrow and anger about these facts. Is there any compelling theoretical reason for expecting these two individuals to exhibit differential responses to the types of affective priming manipulations reviewed in Fazio (2001)? Is it possible to translate this thought experiment into a computer simulation in which associative networks, with initial parameters set to "African Americans take advantage" versus "African Americans are taken advantage of," respond to the IAT? Would these two simulations produce identical RT results? In short, should we theoretically expect indexes of negative affectivity to differentiate people who share a considerable knowledge base but who differ only in their causal attributions for between-group inequality? We question whether the RT measures now in use can reliably make these more refined cognitive distinctions.

Reconsider the experiment by Fazio et al. (1995, Study 1). After completing the affective priming portion of the study, participants were debriefed by an African-American female experimenter who rated each person on their friendliness and interest in psychology and who was "especially attentive to such factors as smiling, eye contact, spatial distance, and body language" (p. 1016). The implicit prejudice of each participant as measured by the affective priming procedure correlated .31 with a composite measure of the African-American experimenter's ratings of the participant's interest and friendliness. This would seem to be an empirical basis for hoping that social psychology's old bogus pipeline is about to be replaced by a bona fide pipeline.

Which emotional-cognitive associations underlie negativity? An enormous range of cognitions can be subsumed under the rubric "automatically activated negativity," any of which might account for the Fazio et al. (1995) results. Negativity could mean anything running from unfortunate, tragic, and victimized to lazy, selfish, and violent. It could cover emotions ranging from sorrow to frustration to anger to despair to shame. Which configurations of associations justify the label *prejudice* or *implicit racism*? We know from the quote that began this article that Jesse Jackson would experience relief if he were walking down the street, heard footsteps, and saw a White person, whereas he would experience a less positive emotion if he were to see an African-American person. Presumably Jesse Jackson would be rated much less friendly by the African-American pedestrian who passed him uneventfully than by a White pedestrian who did so. Rev. Jackson would therefore fail the "friendly/interested" test used in Fazio et al. (1995). Apparently Jackson harbors implicit prejudice, according to the analysis offered by IAT researchers. (Researchers using the affective priming technique generally do not use the term *implicit prejudice*.)

Many other researchers have attempted to validate the measures of implicit prejudice by using nonattitudinal variables. Dovidio, Kawakami, and Gaertner (2002) reported that results from an affective priming procedure correlated .41 with nonverbal measures. The nonverbal assessment of White students in this study was made during a 3-min interaction with either a White or African-American confederate. Prior research by members of this research team (Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997, Experiment 3) found that the number of eyeblinks and amount of eye contact toward an African-American versus a White person was correlated with levels of implicit prejudice. McConnell and Leibold (2001) found that one's pro-White bias on the IAT was related to differential nonverbal behaviors toward an African-American versus a White experimenter. Significant differences were found for such nonverbal behaviors as speech errors, smiling, and speaking time, but not for others, such as seating distance, fidgeting, or expressiveness. Wilson, Damiani, and Shelton (1998) found that implicit prejudice was related to how often a person "handed a pen to an African American confederate, as opposed to placing it on a table" (cited in Wilson et al., 2000, p. 111).

Other validity studies have used behaviors other than nonverbal interaction. Fazio et al. (1995) reported a correlation of .32 between one's results on an affective priming task and one's assignment of responsibility for the 1992 Los Angeles riots primarily to African Americans. Using the affective priming procedure Fazio and Hilden (2001) were able to predict emotional reactions to a public service ad that led viewers to draw an unwarranted and prejudiced conclusion. Dunton and Fazio (1997) used the affective priming procedure and the Motivation to Control Prejudiced Reactions Scale to predict one's evaluation of an African-American male undergraduate. Fazio and Dunton (1997) reported a relation between racial attitudes detected by the affective priming procedure and the extent to which racial characteristics—as opposed to occupational or gender-related ones—were used to assess the similarity of photographs.

It is disconcertingly easy to construct alternative explanations of these IAT and affective priming results within the response-competition explanatory framework. What are the competing responses? One might be a tendency to be hostile and rejecting (as a prejudice or racism interpretation might have it). Or it might be shame or embarrassment linked to the many reasons that White respondents realize that certain ethnic-racial groups have for being angry at them. The former reprehensible motivation—bigotry—might cause a White person to sit further from an African-American or to avert one's gaze. However the latter motivations—shame or embarrassment—might have precisely the same effect. For example, Keltner and Buswell (1996) and Keltner and Harker (1998, Table 4.1) reported that a downcast gaze, halting speech, verbal silence, and slumped posture are characteristics of shame. A White person who is genuinely ashamed of society's treatment of African-Americans by Whites might well be scored as prejudiced by raters in many validation studies that probe links between implicit prejudice with nonverbal behavior such as gaze aversion and body language. Yet a person who is ashamed of Whites' treatments of African-Americans is not likely to be a bigot; the opposite is more likely.

Might another competing response be due to the social awkwardness that stems from the simple fact that some Whites have just had far less experience interacting with members of other ethnic-racial groups? Keltner and Buswell (1997) reported that a downcast

gaze is also characteristic of embarrassment, and Asendorpf (1990, p. 97) summarized evidence that speech disturbances also characterize this emotion. Given the highly segregated nature of many American high schools, a White undergraduate student being interviewed by an African-American might find that situation to be an unfamiliar one that fosters anxiety and embarrassment. A person experiencing such emotions and displaying the accompanying nonverbal behaviors (e.g., sitting further away) is not necessarily prejudiced.

A reviewer suggested that although the nonverbal behaviors in some studies might reasonably be attributed to shame, guilt, or some other emotion, prejudice is the most parsimonious explanation in that it is a possible cause of the targeted nonverbal behavior in every one of the studies. However, we suggest that given the fact that the overwhelming majority of White undergraduates score quite low on explicit measures of prejudice (e.g., Monteith, Voils, & Ashburn-Nardo, 2001), it seems equally as likely that guilt and shame concerning their race's past treatment of African Americans would be aroused in these individuals with high frequency.

Neither the IAT nor the affective priming methodologies can provide answers to these questions concerning which motivation is responsible for a particular nonverbal behavior. Yet some researchers maintain that these test results specifically index implicit prejudice, rather than guilt, nervousness, or any of a large number of other automatically activated negative reactions (Dovidio et al., 1997). It not clear to us how data such as differential eye gaze duration or eyeblink frequency (Dovidio et al., 1997) can be attributed confidently and specifically to implicit prejudice, given that these nonverbal behaviors are also characteristic of a host of other emotional and motivational states. Our position is that the criterion variables in such construct-validational studies are so open to alternative interpretations that even strong correlations are not necessarily diagnostic (in the Bayesian sense) of implicit prejudice. Racial animus is far from the only, or even the most plausible, explanation for findings of this sort.

Recently even more sophisticated techniques for detecting implicit prejudice have been employed (e.g., Chee, Sriram, Soon, & Lee, 2000). Phelps et al. (2000, Experiment 1) reported that differences in strength of amygdala activation to African-American versus White faces was correlated with bias detected on the IAT. The authors announced that "we have for the first time related indirect behavioral measures of social evaluation to neuronal activity" (p. 734). The authors noted that the amygdala is involved in signaling the presence of stimuli with emotional significance. However, it is not clear what emotions are implicated. We see no reason why the results should be attributed to bigotry as opposed to guilt, shame, or numerous other emotions.

The neuropsychological research is in sharp tension with the strong moralistic tone expressed by some researchers in the implicit prejudice literature.[1] If we assume that spreading semantic activation and amygdala activity are beyond one's conscious control, can we hold others blameworthy for such factors as "bad" amygdala behavior? Note that on explicit measures of prejudice such as the Modern Racism Scale (McConahay, 1986), those with "bad" amygdala behavior were exonerated (Phelps et al., 2000). If persons exhibit no explicit prejudice, if their behavior is above psychometrically detectable reproach, but their amygdala fires in a suspicious way, what moral stance should be taken toward such individuals? We question whether they should be censured for manifesting the residues of a racist culture.

Despite our profound skepticism about the ability of eyeblink frequency, amygdala activity, and slumped posture to provide adequate construct validity criteria for measures of implicit prejudice, we recognize room for reasonable disagreement. In this vein, it is useful to consider validational efforts from a Bayesian perspective. For those whose prior odds (e.g., political preconceptions) strongly support the hypothesis that we still live amidst a racist culture, differential eyeblink rates, although not very diagnostic, may incrementally raise their confidence in the implicit measure's construct validity. For those whose current hypothesis is that college students strive to be reasonably fair-minded on racial issues, even strong correlations with eyeblink rates may not be sufficient to produce much, even any, adjustments in their prior belief concerning the validity of the implicit measure. These scholarly observers see alternative explanations for blinking that possess as strong ex ante odds of being correct as the racism hypothesis and that are approximately equally consistent with the experimental data. For them, the construct validity of the implicit measure remains in doubt. In other words, skeptics, among whom we include ourselves, do not doubt that implicit primes can influence behavior (e.g., Dijksterhuis, Aarts, Bargh, & van Knippenberg, 2000). The issue is whether this influence is caused by racism.

Is the negativity really negative? There is another reason why many measures used in prior research need to be interpreted with caution as valid indicators

[1] Skeptics who doubt that implicit prejudice carries a stigma need only refer to the home page of tolerance.org, an organization whose aim is to "fight hate and promote tolerance." We learn here that "even if we believe in our hearts that we see and treat people as equals, hidden biases may nevertheless influence our actions. A new suite of psychological tests measures unconscious bias. … More than 1 million tests have been taken and in each category a large majority of respondents reveal unconscious bias. You may be disturbed by your own tests results." (Tolerance.org, n.d.). This organization's Web site invites visitors to take the IAT.

of prejudice. As pointed out by Brendl, Markman, and Messner (2001), the IAT, for example, provides "at best a relative measure of one target set against another. However, in contradiction to this constraint of relativity, the results of the IAT are often interpreted as reflecting an implicit prejudice for one group over another. The problem with this interpretation is that … prejudice connotes a negative attitude toward a group" (p. 771). It may be the case that Whites respond more slowly when *Latonya* requires responding to the *good* key than when *Betsy* requires responding to the *good* key. However Whites may nevertheless have a positive attitude toward African Americans, albeit not as positive as toward members of their own race. A relative difference in RT between two target sets does not necessarily imply hostility or prejudice toward either group. This same criticism applies to the affective priming paradigm and the neuropsychological measures that demonstrate relative differences in activation as a function of the race of the stimulus.[2]

Consider Study 1 reported by Fazio et al. (1995). Using the affective priming paradigm, the authors reported RT facilitation among White persons when photographs of Whites were used as primes for positive words and photographs of African Americans were used for negative words. The authors also administered the previously noted Modern Racism Scale (McConahay, 1986) to all participants and reported "the distribution of scores on the Modern Racism Scale was heavily skewed in our mass survey of nearly 500 students. Relatively few scores fell at the prejudiced end of the scale" (p. 1020). We assume that the Indiana University students tested by Fazio et al. (1995) are similar to students tested at Yale, University of Washington, and other universities with regard to their scores on the Modern Racism Scale: The large majority of the students do not manifest modern racism. Our suggestion therefore is that it is misleading to bemoan the very high proportion of participants who exhibit prejudice, unless the speaker clarifies that prejudice merely means a relative difference in RT and not any necessarily racial animus.

Monteith et al. (2001) illustrated the readiness to characterize people as prejudiced based exclusively on their IAT results. The University of Kentucky students tested by the authors were decisively nonprejudiced according to their scores on the Modern Racism Scale (McConahay, 1986), with 65% of them falling into the lowest third of the possible distribution and only 7% in the top third. The authors also administered the Should-Would Discrepancy Questionnaire (Monteith & Voils, 1998), which compares how respondents believe they should respond to African Americans in various situations compared to how the respondents think they would respond to African Americans in those situations. A difference between these two ratings is termed *discrepancy proneness* and denotes a failure to behave according to one's standards of appropriate racial behavior.

There were several results of interest. First, nearly all participants manifested implicit prejudice on the IAT. Approximately two-thirds of the participants noted that they responded more slowly when *Black* was paired with positive words than when *White* was. Such detection was associated with feelings of guilt among participants. However the detection of implicit prejudice in one's IAT responses was not related to the actual presence of implicit prejudice in one's IAT responses. In other words, taking the IAT made a number of people feel more guilty than was warranted granting that the test measures what it purports to measure. A second result was that those with low levels of discrepancy proneness did not feel guilty about their prejudiced IAT results, whereas the high discrepancy prone people did.

Monteith et al. (2001) emphasized that "participants generated truly biased responses in the experimental session" (p. 412). This refers to the IAT responses, of course. The authors are disturbed that these recalcitrant low discrepancy proneness people did not feel guilty about their IAT results. Nevertheless the authors suggested that the IAT can be used to provide many people with self-insight into their prejudice; after all, the majority of the respondents when asked did note that they manifested prejudiced response patterns, even though most of them were exonerated on the Modern Racism Scale (a scale that itself sets a low threshold for calling people racists by measuring conservative policy preferences such as opposition to affirmative action; see Sniderman & Tetlock, 1986). However the authors are still wary: "We believe the important take-home message is that even if people report being discrepancy-not-prone, and they typically are highly effective at responding in non-prejudicial ways … , nonconscious biases still may reside in the mind" (pp. 414–415). The detection of nonconscious biases in the mind rests solely on the IAT results. A nonprejudiced score on the Modern Racism Scale plus a low discrepancy- proneness score plus a high level of effectiveness in responding in unprejudiced ways does not absolve anyone if the IAT yields suspicious results.

Attitudinal versus associative interpretations of the results. Banaji (2001) asserted that it is reasonable to suppose that implicit attitudes are being assessed in paradigms such as those of Fazio et al. (1995) and Greenwald et al. (1998). She took the definition of attitude as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken, 1998, p. 269) and

[2] This criticism does not apply to the recently developed Go/No-go association task (Nosek & Banaji, 2001).

the definition of implicit memory as something that is "revealed when previous experiences facilitate performance on a task that does not require conscious or intentional recollection of these experiences" (Schacter, 1987, p. 501). She then combined these propositions to define implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald & Banaji, 1995, p. 8).

Banaji (2001) also argued that, regardless of whether evaluative priming and IAT measures are highly correlated, they do share four key assumptions: (a) It is possible to assess the strength of evaluative associations; (b) the extent to which concepts share connotative/evaluative meaning (independent of denotative/semantic meaning) is manifested in the ease with which concepts can be mentally paired; (c) a good indicator of the strength of an evaluative association is the speed of object plus evaluation pairs; and (d) the strength of evaluative association as assessed under conditions of speeded responding is a reasonable measure of automatic attitude. In Banaji's words: "Both tasks measure the strength of evaluative association in some way, and both take the strength of that evaluative association to reflect the strength of automatic attitude—that is their fundamental commonality" (p. 124).

It is worth stressing that virtually no one denies that implicit attitudes are plausible theoretical constructs that serve plausible psychological functions. Insofar there is controversy over exactly what various priming measures assess, it revolves around what types of inferences can be drawn from these measures: associative or attitudinal. Do these measures detect mere associations whose implications one may or may not endorse, or do they tap into one's attitudes toward various groups? What type of content would an implicit attitude have to possess—or for that matter an explicit attitude have to possess—to justify labeling it an instance of prejudice or racism? Banaji's fourth assumption is the politically and theoretically controversial one because it converts an association one has to an attitude one endorses at some level.

Is it Possible to Pass Classic Correspondence and Coherence Benchmarks of Rationality and Still Be Prejudiced?

To call someone prejudiced or racist in early twenty-first century America is to comment on both the cognitive competence and moral standards of that individual. The cognitive indictment runs through the professional literature: The prejudiced are too quick to jump to conclusions about target groups, too slow to acknowledge disconfirming evidence and to update beliefs in response to such information, and prone to see relationships between variables and group membership that are weakly or not at all connected (Allport & Postman, 1947; Henderson-King & Nisbett, 1996; Huici, Ros, Carmona, Cano, & Morales, 1996; Maass, Montalcini, & Biciotti, 1998; Rothbart, Evans, & Fulero, 1979; Rothbart & John, 1985; Ybarra, Schaberg, & Keiper, 1999; Ybarra, Stephan, & Schaberg, 2000). The moral indictment is no less explicit. There is something mean-spirited or selfish about those who harbor prejudiced attitudes (Sniderman & Tetlock, 1986).

This article does not dispute the obvious: Prejudice is often linked to rigid beliefs, unattractive motives, and hostile emotions that, in turn, are often linked to horrific social consequences. We do, however, raise questions—politically sensitive ones—about the criteria that implicit prejudice researchers implicitly use in labeling judgmental tendencies as evidence of prejudice. In all complex societies, there are identifiable subgroups that observers can readily classify along ethnic, linguistic, religious, and racial lines and that differ from each other in a host of ways that observers might deem evaluatively significant. Differential crime rates, differential sexual mores, differential educational test scores, and differential levels of socio-economic achievement are all plausible bases for differential reactions to groups (putting the "justifiability" of such reactions to the side). Researchers who treat implicit associative measures as presumptive indicators of prejudice argue, in effect, that people are prejudiced whenever (a) they live in societies in which inequalities across groups exist, (b) they correctly perceive those inequalities, and (c) the dimensions on which the inequalities exist have been vested with evaluative significance. Using current cognitive standards for identifying implicit prejudice in social psychology, we would be required to label realistic Bayesian information processors as prejudiced in all but the most homogeneously egalitarian societies that have succeeded in eradicating all differences of evaluative significance among identifiable groups.

In this section, we show that people who pass classic tests of rationality in the experimental literature—tests such as expected utility maximization, the attributional logic of discounting, Bayesian utilization of base rate information, the efficiency of cue utilization strategies in stochastic environments, and Bayesian belief updating—would almost certainly qualify as prejudiced on implicit associative measures if we assume the correctness of the theoretical mechanisms postulated by implicit prejudice researchers themselves.

Expected utility theory. Expected utility (EU) theory has long been accepted as a cornerstone of ratio-

nal decision making (von Neumann & Morgenstern, 1947). Persons who follow the consistency and combinatorial axioms of EU to maximize utility are, by definition, acting rationally. Of course, this assertion presumes the reasonableness of the inputs into the decisional calculus of EU theory. If a psychotic person assigned very peculiar probabilities or utilities, we might dispute the attribution of rationality. However if the assigned probabilities were based on accurate actuarial data and the utilities were widely endorsed, then a person who combined these factors according to the maxims of expected utility could not be deemed to be acting irrationally.

We begin by considering again the quote by Jesse Jackson, which suggests that he feels relieved when he realizes that the steps on the sidewalk behind him are those of a White person. Is he acting irrationally if he experiences more anxiety when the footsteps behind him are those of an African-American person? Based on Jackson's quote, we hypothesize that the affective priming or IAT methodologies would yield the typical White result in his data: He associates African American with some negative characteristics more than he associates White with those characteristics.

One way to measure the rationality of Jackson's thinking is to apply the standards of EU theory. We do not suggest that Jackson actually does use EU theory when he contemplates the footsteps behind him, nor do we suggest that anyone else consciously does either. However it is still possible to use the theory to decide if it is rational for Jackson to be more likely to take evasive action when the follower is African-American rather than White.

We begin with several assumptions. First, the following example draws on the 1999 Uniform Crime Statistics from the Department of Justice pertaining to robbery, on the assumption that this is the crime that Jackson fears. Second, we assume that the utility of the possible outcomes vary from the worst possible (–1,000) to uneventful (zero for being overtaken by a neutral pedestrian). Third, we assume that fleeing before the follower arrives will be successful in preventing a robbery and that this action has a small disutility (–2). The question is whether an entirely rational utility maximizer should flee when the approaching footsteps belong to an African American but not when they belong to a White. The 1999 data from the Department of Justice, when combined with the 2000 census data (U. S. Census Bureau, 2000), indicate that it is 7.49 times more likely for the decision maker to be in danger of being robbed if the follower is African American than if the follower is White. We assume that our decision maker is aware of these data in a general sense. Of course, our rational decision maker should flee if the EU of fleeing is greater than the EU of not fleeing. We begin by solving for the magnitude of disutility at which fleeing has precisely the same EU as not fleeing, given that the follower is African American.

The EU of any outcome is equal to its probability of occurrence multiplied by its utility or, in this case, its disutility. If one chooses not to flee, there are two possible outcomes. The first is that the follower is a neutral individual, who passes without incident. The probability of this outcome is .9936, and the disutility of this outcome is zero. The second is that the follower is a robber, which has a probability of .0064. The disutility of this outcome is designated U_r. The EU of these two possible outcomes of not fleeing must be added to obtain the aggregate disutility of not fleeing.

$$\text{Utility of not fleeing} = (.9936)(0) + (.0064)U_r$$

If one chooses to flee, there are also two possible considerations. The first is that the follower is a neutral individual, and fleeing was unnecessary. The probability of this outcome is .9936, and the disutility of this outcome is –2. The second possible outcome is that the follower would have been a robber, but successful evasive action prevented that large disutility. The probability of this outcome is .0064, and the disutility of this outcome is again –2, due to the success of the evasive action. The EU of these two possible outcomes of fleeing must be added to obtain the aggregate disutility of fleeing.

$$\text{Utility of fleeing} = (.9936)(–2) + (.0064)\,(–2)$$

If we set the EU of not fleeing equal to the EU of fleeing and solve for U_r, we find that the disutility of being robbed would have to be worse than –312.5 for the person to flee, assuming that this person is a utility maximizer. If a person's disutility of being robbed were exactly –312.5, then this individual would be undecided whether to flee when the steps were those of an African-American person.

What if the steps belonged to a White? Substituting the probabilities that the White follower was a robber in the previous equation, the disutility of being robbed would have to be worse than –2347.4 for fleeing to be a utility maximizing choice. Because the scale goes only as low as –1,000, it would never be rational to flee the White.[3] Thus Rev. Jackson (or anyone else) might be entirely rational to take evasive action only if the follower were African-American.

It is possible to modify this equation in unusual ways to help resolve this unfortunate state of affairs. For example, if the person being followed had extremely serious coronary disease such that fleeing would be very likely to cause death, then not fleeing

[3] Even if the worst possible outcome were more negative than –1,000, differential crime rates would still cause a rational decision maker to manifest different behavior toward members of different ethnic groups if the probability of criminal behavior of those two groups differed.

might be a rational choice no matter what the race of the follower might be. However given the differences in the probabilities, it is not easy to generate utilities which prevent a differential response depending on the race of the follower. The conclusions of this analysis would be the same if any of several crimes were substituted for robbery. The general principle is this: As long as there is a differential crime rate between racial groups, a perfectly rational decision maker may manifest different behaviors—explicit and implicit—toward members of different races.

Many social psychologists might concede that differential behavior toward pedestrians of different races is consistent with narrowly self-interest-based conceptions of utility theory but insist that a truly nonracist person would enter more public-spirited considerations into the utility calculus. For example, racial harmony would be furthered if pedestrians refrained from fleeing or even becoming nervous when they saw that the follower was African-American. Racial harmony and societal cohesion surely are outcomes with large positive utilities.

A related argument is that fleeing and excessive nervousness might create the very traits considered undesirable in the stereotype. Behavioral confirmation (Snyder, 1981) and stereotype threat (Steele, 1997) are two ways in which such unfortunate outcomes would occur. For example, if I believe that a person will act in a congenial way toward me, my actions toward that person may elicit the friendly behaviors that I predicted that person would exhibit (Snyder, 1981). Thus if I act as if the person following me was undesirable, that person would be more likely to act less positively toward me that if my actions did not communicate such a negative view. Is there not substantial disutility in causing the bad behavior I fear? If there is such disutility, this might tip the EU calculations toward the rationality of remaining calm, thus contradicting the rationality of the demeanor of the anxious Rev. Jackson.

We agree that additional factors could be inserted into the expected utility formula and that such factors could be assigned utilities such that fleeing would cease to be the optimal choice. However we question whether psychologists should expect others to modify their expected utility calculations in this manner. Rev. Jackson, who we assume is highly motivated to behave in a nonracist manner, stated that he was relieved to learn that a follower was White. Presumably his greater anxiety toward the African-American pedestrian would also motivate other incriminating indicators of implicit prejudice such as eyeblinking and amygdala firing. Should Jackson be blamed for not assigning enough utility to racial harmony or disutility to behavioral confirmation? Some psychologists might say so, but we are not tempted to impose our ivory tower judgments on everyday people coping with situations in their own lives.

Several helpful reviewers of this article have pointed out that although a person might be justified in feeling nervous when being followed by a person belonging to a subgroup with a higher probability of committing a crime, this does not mean that all prejudiced behaviors can be "explained away" by appeals to rationality. We wholeheartedly agree. A person who simultaneously manifests a negative attitude toward a large number of groups including Arabs, conservatives, Americans, Canadians, political moderates, White Anglo-Saxon Protestants, and even nonexistent people known as Meblus (Fink, 1971) is likely to be prejudiced, and nothing in our article should be construed as a denial of the existence of prejudice. Our point is that if a person exhibits quicker RT to *prison* and *jail* when the response to such words occurs on the same response key as African-American names as opposed to White names, the person doing so is not necessarily implicitly prejudiced. When this individual blinks more frequently when speaking with an African American than when speaking to a White, the person doing so is not necessarily implicitly prejudiced. And finally, as we argue in this section, when this person, as in the words of Rev. Jackson, "hear[s] footsteps and start[s] thinking about robbery [and t]hen look[s] around and see[s] somebody White and feel[s] relieved" this person is not necessarily implicitly prejudiced. In fact, he or she may be acting rationally.

To be clear, we reemphasize that we do not justify the irrational aspects of racism; they cannot be justified. The question is whether any datum that might show differential IAT results, eyeblink frequency, or pedestrian anxiety as a function of race constitutes proof of prejudice. Our position is that it does not.

The discounting principle. In one of the original statements of attribution theory, Kelley (1971) posited people to be intuitive scientists who use logically defensible rules, such as the discounting principle, in their efforts to master the causal structure of the interpersonal world. According to the discounting principle, attribution of an outcome to dispositional or internal causes should be tempered to the degree that observers believe plausible external or situational causes are also present. Ross (1977) stated that "the Discounting Principle requires a 'psychologist' able to assess the role of various social pressures and situation forces" (p. 181).The person who successfully applies this principle thereby demonstrates sound social perception.

The fundamental attribution error (Ross, 1977) is rightly called an *error*, because it represents a failure to use the discounting principle as much as one should. Therefore it is incongruous—from a traditional intuitive-scientist perspective—to criticize observers who reduce the role of dispositional causes in explaining

why target actors get benefits whenever affirmative action arises as an alternative situational explanation. Our confidence that we have learned anything diagnostic about a focal actor's abilities and character should fall to the degree that the behavior or outcomes to be explained can be plausibly attributed to situational or external forces. However some writers take it to be a tell-tale sign of modern racism when observers express greater doubts about the competence of affirmative action beneficiaries than of nonbeneficiaries (e.g., McConahay, 1986). Is the use of the discounting principle in this instance a manifestation of racism, whereas its use in other contexts is an index of sound judgment?

Numerous studies have shown that observers disparage the abilities of persons who are perceived as having benefited from affirmative action (e.g., Garcia, Erskine, Hawn, & Casmay, 1981; Heilman, Block, & Lucas, 1992; Jacobson & Koch, 1977; Northcraft & Martin, 1982). Other studies have shown that the beneficiaries also may denigrate their own abilities (e.g., Heilman, Battle, Keller, & Lee, 1998; Heilman, Simon, & Repper, 1987). These results are all consistent with the discounting principle. To the extent plausible external causes are present, (i.e., the affirmative action policy), any hiring, promotion, or other benefit will be less likely to be attributed to an internal factor, such as ability or merit.

Consider Heilman et al's (1998) series of studies. Participants learned about selection policies in which merit was the only consideration, given equal weight to demographic characteristics of the candidate, assigned less importance than demographic characteristics, or assigned no importance. Key dependent variables assessed perceptions of those who were or were not selected. The results were straightforward. When merit was central to the selection process, beneficiaries and nonbeneficiaries alike gave higher ratings to the competence of the person selected. When merit was deemphasized, participants thought less highly of the selected individual. Heilman, Block, and Stathatos (1997) found that denigration of the competence of affirmative action beneficiaries only occurred when their performance level was ambiguous. When their performance level was unambiguous, the affirmative action status of the ratee did not influence competence. All of these results are precisely what the discounting principle would predict. However when observers denigrate the abilities of the beneficiaries of affirmative action in accord with the discounting principle, many psychologists no longer think that use of the principle represents sound thinking. In fact, some psychologists suspect that modern racism is responsible for doubts about the abilities of the beneficiaries of affirmative action or about affirmative action itself (Dovidio, Mann, & Gaertner, 1989; Jacobson, 1985).

What happens when people fail to use the discounting principle? If this lapse occurs in a nonracial context, it is deemed diagnostic of poor thinking. Ross, Amabile, and Steinmetz (1977) assigned persons to either a questioner or contestant role in a quiz game. The questioners were asked to make up difficult general knowledge questions to pose to the other person. Needless to say, most contestants did not fare well in answering these esoteric questions. Nevertheless people in both roles rated the questioners to be more knowledgeable than the contestants, apparently not taking into account the terribly unfair roles the two persons were obliged to assume. The two participants should not have attributed the poor performance of the contestant to an internal cause, such as lack of knowledge, but should have instead discounted that cause in light of the powerful situational reason for the poor performance, namely the disadvantageous role in which the contestant was placed. Participants who failed to heed the discounting principle in this experiment were deemed "consistently biased and erroneous" (Ross, 1977, p. 194). Note that this description is akin to those used in racial contexts when people do use the discounting principle and question the competence of persons who benefit from affirmative action. Is use of the discounting principle rational or is it not? Can it be rational in one setting but irrational in another?

One could argue that using the discounting principle is inappropriate in racial contexts because fair-mindedness requires offsetting applications of the discounting principle with equally forceful or even stronger applications of the mirror-image augmentation principle (also proposed by Kelley, 1971). According to the augmentation principle, those who succeed in the face of inhibitory causes such as poverty or discrimination should enjoy an augmentation of the internal attribution for their accomplishments. After all, to succeed despite these factors would require extraordinary ability and motivation.

Female managers seem to think that the discounting principle is more applicable than augmentation. Heilman et al. (1997) asked both male and female managers to evaluate the job performance of men and women, some of the women being identified as an "affirmative action hire." Male and female managers rated the job performance of the female affirmative action hires as less competent than the performance of the men and the nonaffirmative action women and also recommended lower salary increases for the affirmative action women. Presumably the female managers were well-positioned to judge whether the performance of female affirmative action hires should be augmented or discounted, and they opted to discount.

Bayesian inference and base rates. Decision theorists routinely invoke Bayes' theorem as the appro-

priate principle for aggregating base-rate and case-specific information (cf. Fischhoff & Beyth-Marom, 1983). The underlying idea is simple: After a datum has been observed, the prior odds of an event are multiplied by the diagnosticity of the datum to generate the posterior odds of the event. In the famous lawyer–engineer problem, Kahneman and Tversky (1973) asked some participants to consider a set of 100 thumbnail descriptions of men, 70 of whom were engineers and 30 of whom were lawyers. The prior odds that a randomly selected description would be that of an engineer are thus 70/30. If the specific description seemed as though it was twice as likely to be that of an engineer (his hobbies include home carpentry), then the likelihood ratio is 2. Multiplying the prior odds by the likelihood ratio gives posterior odds of 140/30. The probability that this person is actually an engineer is thus 140/(140+30), which is .82. The description was diagnostic in that it moved the posterior odds away from the prior odds, thus providing an updated estimate, which would be more informative than the original estimate of .70 that the person was an engineer.

The surprising result from the Kahneman and Tversky (1973) demonstration is that when the 100 descriptions were said to be those of 70 lawyers and 30 engineers, this substantial change in the base rate of lawyers and engineers had relatively small impact on the estimate that a given description was that of an engineer. However Bayes' Theorem requires that this change should have had a substantial effect, dropping the posterior probability that the description was that of an engineer to .46. Kahneman and Tversky (1973) stated: "The failure to appreciate the relevance of prior probability in the presence of specific evidence is perhaps one of the most significant departures of intuition from the normative theory of prediction" (p. 243).

This Bayesian combinatorial principle is so widely accepted within the research community that psychologists have for decades felt justified in positing that people who fail to factor base rates into their predictions have committed an error. Base-rate utilization became a formal benchmark of rationality—a benchmark that empirical work repeatedly revealed that people often failed. For many years, the base-rate fallacy has been regularly trotted out in influential textbooks as a lead exhibit in the case for human irrationality (e.g., Nisbett & Ross, 1980). The standard explanation has been that people make subjective-likelihood judgments by relying on simple error-prone heuristics such as representativeness, in which judgments about the probability of category membership hinge entirely on the perceived similarities of the target to the defining features of the category (Kahneman & Tversky, 1972).

It is ironic therefore when people are censured not for failing to use base rates but rather for using them. We saw one example in the earlier discussion of EU maximization. The source of the expectation that one was at greater risk if followed by a young African-American man was population base-rate statistics compiled by the federal government of the United States. Change the identity of the Rev. Jackson to a law enforcement officer deciding whether to detain a suspect, we have a potentially legally actionable case of racial profiling.

The decision to use or not to use base rates in a legal context (i.e., racial profiling) is a decision based partly on highly tangible trade-offs. In a respected econometric study, Farmer and Terrell (2001) showed that, as long as differential crime rates exist across groups in society, minimizing the overall crime rate will result in far more convictions of innocent members of the minority group even if racism is not at work. It follows mathematically (within a wide range of plausible assumptions) that by requiring less evidence to convict members of a smaller but higher crime group, one will simultaneously lower the overall crime rate and increase the overall probability of convicting an innocent person. This troubling state of affairs must be considered in light of the opposite option: to rectify racial inequality in the probability of erroneous convictions, society must tolerate a higher crime rate, whose victims will predominantly come from the minority group. Indeed, Farmer & Terrell (2001) have estimated that approximately 1,900 more murders per year will occur if racial inequality is removed from the erroneous conviction rates. These are exceptionally serious considerations. We do not dispute the prerogative of people who believe that using the base rate is a poor policy. Such people may have considered the trade-offs and have decided that the preservation of important principles is worth the cost of 1,900 lives. We respect their decision. What we do dispute is whether those who do use racial base rates, such as the nervous pedestrian Rev. Jackson, automatically deserve our censure for exhibiting prejudiced judgment.

It is not difficult to construct experimental scenarios in which observers denounce decision makers who factor race-charged base rates into their deliberations but render no objections when decision makers use exactly the same statistical information but now with the racial charge removed. Tetlock, Kristel, Elson, Green, and Lerner (2000) obtained exactly this pattern of results among observers who had been asked to judge the propriety of decisions by insurance executives to issue home insurance policies in different neighborhoods, or to charge different rates as a function of base rate statistics on risk of property damage. Observers, especially the most liberal among them, directed moral outrage at the executive who used base-rate information that was revealed to be closely correlated with the percentage of African Americans in the neighborhoods but no outrage at the executive who used the same base-rate information minus the racial correlates. Tetlock et al. (2000) used the term *forbidden base*

rates to characterize this phenomenon and went on to show that people who are lured into using a base rate that was subsequently revealed to have racial implications felt an increased need to engage in symbolic acts of atonement, such as volunteering for good causes. These acts of "moral cleansing" appear designed to distance self from potential attributions of racism.

Tetlock et al. (2000) defined forbidden base rates as any statistical generalization that devoted Bayesians would not hesitate to enter into their probability calculations but that deeply offends a religious or political community. The obstacle to using such base rates is not cognitive, but moral. Putting the accuracy and interpretation of such generalizations to the side, people who use these base rates in judging individuals are less likely to be applauded as savvy intuitive statisticians than they are to be condemned for their moral insensitivity.

The rationality of fast and frugal heuristics. In his classic statement of probabilistic functionalism, Egon Brunswik (1949) sought to model how people cope with the massive cognitive challenge of making sense of an environment in which the states of the world that people can observe (proximal cues) are related in a probabilistic rather than deterministic manner to states of the world that have the strongest implications for their future well-being (e.g., distal cues such as whether the food they are about to eat is poisoned or whether a prospective mate will be faithful). The likelihood of a rewarding or punishing outcome can only be estimated from past correlations between proximal and distal cues and the current levels of proximal cues in the environment. (Hammond & Stewart, 2001, p. 3). To survive, organisms must become adept at learning cue–outcome relations.

Recently Gigerenzer, Todd, and the ABC Research Group (1999) have furthered the Brunswikian tradition by enumerating a number of "fast and frugal" cognitive tools that humans use to solve the cue–outcome relations with which they are confronted. Some of these tools are shockingly simple and equally shockingly effective. For example, Goldstein and Gigerenzer (2002) discussed the recognition heuristic. When deciding in which company to invest, one possible rule would be to choose companies one recognizes and avoid those one does not recognize. When deciding which of two foreign cities is larger, choose the one whose name one recognizes. Using an ecological cue no more sophisticated than one's recognition of the stimulus, people can achieve very high accuracy in categorization, choice, and other cognitive tasks. Note that the proximal cue may be imperfectly related to the criterion, because not all recognized companies enjoy financial success nor all recognized cities are large. Mistakes will be made. Due to the probabilistic nature of the task, mistakes are inevitable. The question is whether people have the cognitive tools necessary to do a good job in this difficult, probabilistic environment. The conclusion from the many experiments discussed in Goldstein and Gigerenzer is that people are quite good, indeed. In some cases, using very simple cognitive tools in a probabilistic choice task enables people to achieve a rate of success equal to that of multiple regression (Gigerenzer & Goldstein, 1999)!

Interestingly, about 70 pages before the Goldstein and Gigerenzer (2002) article in *Psychological Review* is an article by Greenwald et al. (2002). That article contained a theoretical discussion that drew on empirical data from the IAT. One assumption of the IAT articles during the last several years is that people who note probabilistic relations between groups and various attributes are exhibiting implicit prejudice, that is, deficient judgment. For example, *prison* and *jail* are two of the negative words used by Greenwald et al. (1998). Whites exhibit quicker RT when the response to such words occurs on the same response key as African-American names as opposed to White names. The Federal Bureau of Investigation's (1997) arrest data for 1997, which is the year prior to the publication of the Greenwald et al. study, shows that violent crime per 100,000 persons was 3.59 times higher for African Americans than for Whites (Farmer & Terrell, 2001, p. 356). The probabilistic relation between race and crime is a well-known national problem. How should we characterize the cognitive processes of persons who are aware of that relation? Are such people bigots? The analyses presented by Goldstein and Gigerenzer (2002) plus the expanded discussion in Gigerenzer et al. (1999) suggest that people should have little trouble in noticing such relations, and they will do so using very simple cognitive tools. The juxtaposition of the two research programs is noteworthy: One portrays the vast majority of people as implicitly prejudiced, which we take to mean having a negative attitude with an insufficient evidentiary basis; the other portrays them as cognitively adroit, which we take to mean having the acumen to notice cue–outcome relations. Both research programs suggest that people are sensitive to the ecological validity of cues in their environment, but the implicit prejudice research program places an unmistakably uncomplimentary value judgment on this behavior.

The tension between the two viewpoints is closely related to the Bayesian inference problem mentioned earlier. In that discussion we point out that the first component of Bayes' Theorem, the base rate, must be used to arrive at sound posterior probability estimates. Rev. Jackson is apparently aware of racial differences in base rates of robbery, and Bayesians would trace his differential level of anxiety to the race of the follower. The research by Gigerenzer and his colleagues (1999) is also pertinent to another component of Bayes' Theo-

rem, the likelihood ratio, by which people consider new data to update their probability estimates. For example, if a person knows that women comprise about 50% of the American population, a person might begin by assuming this base rate reflects the proportion of female criminals in the population. However as a person reads the newspapers or has personal experience with crime, this subjective base rate is updated, so that the person should eventually come to realize that women represent far less than 50% of criminals. People who fail to adjust to the base rate would be making a clear-cut error of probabilistic reasoning in the Bayesian framework. However, people who adjust to the base rate would be making an error in the Greenwaldian framework: They should find it increasingly easy to associate *male* with bad things (such as *murder*, *prison*, etc.) and increasingly difficult to do so for *female*. Whether we view good belief updaters as Bayesians or bigots hinges on whether we deem it appropriate to apply an epistemic norm, rooted in probability theory, or a moral norm, rooted in egalitarian political sentiment. In fact, Banaji (2003) used the term "Bayesian bigots" to describe persons who use base rates in providing probabilistically defensible but morally insensitive dependent measures. This term completely captures the tension between the two norms.

Note that the etymological origin of *prejudice* is to *prejudge*: to render a judgment before data are collected. Prejudice presupposes an unwillingness to change one's opinion in response to new veridical information. A person who has accurate statistical knowledge of demographic variation in crime would, by definition, knows about racial differences in crime rates. This veridical information should cause one to update one's prior probability—the base rate—if one believes that members of all races are not equally likely to commit crimes. If a person does the appropriate probability updating and then takes the IAT, he or she should be more likely to associate crime-related words with African-American names than with White names. Now the person is deemed to have exhibited not good judgment by appropriately updating probability estimates in light of veridical information but prejudiced judgment by exhibiting differential RT. Again, the standards of rationality are not just distorted; they are inverted.

The Political Psychological Context of the Debate

Banaji (2001) overlooked the most obvious and politically unsettling alternative explanation of the results of the affective priming and IAT research: actual differences in social reality. There are many groups in American society—identifiable by ethnicity, language, religion, and race (among other things)—that vary on dimensions that various subsets of observers might judge to carry evaluative import. Differentials in family breakdown, educational test scores, crime rates, socioeconomic achievement, and mortality statistics are all plausible bases for differential reactions to groups. Equally relevant may be differential histories of how groups have been treated. A history of cruel exploitation—widely publicized images of shackled prisoners on slave ships, despairing families split at auctions, and public floggings and lynchings—could be linked easily to affective negativity. Assuming that those who participate in the affective priming and IAT methodologies all live in societies in which inequalities exist and are perceived, then as long as the participants are sensitive enough to imbue those inequalities with evaluative significance, nearly everyone will exhibit implicit prejudice and the residues of a racist culture, precisely the results that have been reported. Sentient organisms aware of their environment will be accused of harboring hidden biases.

The central argument here is part psychological, part philosophical, and certainly part political: The need to work through the implications of acknowledging that prejudice is a value-laden socio-political construct and that its identification is not on an epistemological par with mapping patterns of neuronal activation in various regions of the brain. It is possible to show determinants of all the variables mentioned in this article, and the patterns of interrelations among them, without recourse to tendentious political labels of uncertain and shifting meaning. For our part, we recommend one of two paths: either abandoning vague, value-laden labels or, if investigators insist on using such terms, specifying exactly which of the foregoing thresholds of proof they are employing in labeling a person or point of view as prejudiced.

Stable Disposition or Person-by-Situation Interaction?

Arguments over what should count as prejudiced often revolve around how crude, undifferentiated, and sweeping the disparaging generalizations must be (Sniderman & Tetlock, 1986). The assessment target of most implicit prejudice research has been attitudes toward Whites and African Americans as undifferentiated wholes. However, there are good reasons for supposing that attitudes toward racial groups in America have under the sheer press of evidence become more complex and qualified. In segregated society, interactions between Whites and African Americans were briefer and took a narrower range of social forms than today (Sniderman, 2002).

The research of Mischel, Shoda, and Mendoza-Denton (2002) suggests that rather than treating implicit prejudice as a stable dispositional

property, researchers might consider the contexts in which actors and perceivers of different races interact. For example, Wittenbrink et al. (2001) and Barden, Maddux, Petty, and Brewer (2004) have separately shown the impact on IAT or affective priming results of the context in which the African-American individual appeared. In the first experiment of Wittenbrink et al. (2001), all participants were first given the IAT. Half were then shown a 2-min segment of a movie depicting African-American gang members. The other half of the participants saw a 2-min segment of a movie of an African-American family at a barbecue. Subsequently all participants were given a second IAT, during which brief clips of the movie each participant had seen were presented. Finally all participants filled out explicit measures of prejudice. The principal result was that participants who saw the film of a harmonious African-American family manifested a significant reduction in the typical prejudiced result as assessed by the second IAT. The second experiment of Wittenbrink et al. (2001) yielded the analogous result, but this time using the affective priming paradigm. In this study the positive context was a church, and the negative one was a dilapidated street corner.

The Wittenbrink et al. (2001) results are problematic for the position that the IAT or affective priming methodologies tap rigid, broadly undifferentiated group-related beliefs. African-American individuals were in all movies (Experiment 1) and in all photographs (Experiment 2). All respondents were Whites. Due to randomization of participants it is most unlikely that the less prejudiced individuals happened to be in the barbecue and church groups rather than the gang and street corner groups. It is more likely that the IAT and the affective priming methodologies assess the extent to which a stereotypic association is manifested. With the context held constant, which is the typical procedure, Whites manifest the usual IAT results, given the fact that virtually all White members of the society are aware of the stereotypic associations. When the contexts differentially cue those stereotypic associations, as in the Wittenbrink et al. (2001) and Barden et al. (2004) studies, the results do not suggest stable levels of bigotry within the attitudes of the participants. Instead, Wittenbrink et al. (2001) and Barden et al. (2004) demonstrated that context is a powerful moderator of the prejudice participants manifest. Note that the attribution of prejudice to persons who manifest incriminating IAT results is the most stigmatizing explanation of such data. Shared cultural knowledge of the relative dangerousness of gang members and barbecue attendees is the least.

de Tocqueville's Prediction

In his classic *Democracy in America* de Tocqueville (1835/2001) argued 160 years ago that the natural long-term trend within democracies is for citizens to become ever more alert to, and intolerant of, sources of social inequality. It is a sign of how disconcertingly accurate this prediction is in our era that cognitive research programs now attempt to gauge prejudice not by what people do, or by what people say, but rather by millisecs of response facilitation or inhibition in implicit association paradigms. Viewed historically, the progression toward an ever more comprehensive egalitarianism began in de Tocqueville's era with the abolitionist struggle to abolish slavery in the 1840s and 1850s; the struggle made landmark advances in the 1950s and 1960s with the successful civil-rights campaigns to eliminate de jure segregation and into the 1970s and 1980s with the post-civil-rights campaigns for affirmative action to eliminate de facto inequalities; and the movement now appears to have culminated in the early twenty-first century with the psychological quest to eliminate subjective sources of inequality. Some would levy such moral-political judgments (e.g., "Fight hate," "dig deeper, "test yourself for hidden bias") not for what we actually do or consciously think but rather for what cognitive psychologists infer we must be thinking from how readily affective negativity can be primed by various stimuli.

Even if we believed the cognitivist claims to have discovered a window into our souls, is this a policy path that we, as a society, should take? Imagine, by way of a thought experiment, that the IAT and affective priming methodologies had been available in the heyday of McCarthyism in the early 1950s. Imagine that researchers proposed to adapt these measures for the purpose of measuring implicit Marxist attitudes by using as primes photos of political figures—Lenin, Stalin, Truman, Churchill—and then observing patterns of RT facilitation or inhibition. Would this not have struck many in the academic community as Orwellian? Would not the indignant reaction have been multiplied if follow-up research focused on using behaviorist principles for extinguishing differential implicit evaluative responses?

We wish to be clear that we do not believe that scientific research on promising leads should halt because someone can concoct a horror story of how the resulting technology could be abused. We do believe, however, that the enthusiasm for cranking up the magnification in search of covert prejudice needs to be placed in historical perspective. If we think of racial prejudice as the primal blemish on America's collective reputation as a just society, and if we agree with de Tocqueville (1835/2001) that Americans have a uniquely "lively faith in the perfectability of man" (p. 359), the IAT can be viewed as a quintessentially grass-roots project to use the best available science to expunge racist sentiments not only from the consciousness of Americans but from their unconscious as well. (See de Tocqueville, 1835/2001, Part I, p. 18.) How-

ever, if the decades of representative-sample surveys cited in the first paragraph of this article are correct and racism is in steep decline, then hunting for its vestiges using the millisec precision of modern computers appears in a different light: a project that requires attaching increasingly tendentious interpretations to implicit associative measures that are well-suited for answering precisely formulated psychological questions about the working of human memory but that are less suited for tackling political questions about the tenacity of prejudicial behavior.

Notes

We thank Russ Fazio, Michael Sargent, Kris Preacher, and William von Hippel for their helpful comments on earlier drafts of this manuscript.

Hal R. Arkes, Department of Psychology, Ohio State University, 240N Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43201. E-mail: arkes.1@osu.edu. Philip E. Tetlock, Haas School of Business, University of California–Berkeley, Berkeley, CA 94720-1900. E-mail: tetlock@haas.berkeley.edu

References

Ajzen, I. (1987) Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 20* (pp. 1–63). San Diego: Academic.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Allport, G. W, & Postman, L. (1947). *The psychology of rumor*. Oxford, UK: Holt.

Asendorpf, J. (1990). The expression of shyness and embarrassment. In W. R. Crozier (Ed.), *Shyness and embarrassment: Perspectives from social psychology* (pp. 87–118). New York: Cambridge University Press.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, III, 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. (2003, June). *Mind bugs: The psychology of ordinary prejudice.* Colloquium presentation at the Ohio State University, Columbus.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of implicit racial attitudes. *Journal of Personality and Social Psychology, 87,* 5–22.

Bell, D. (1992). *Faces at the bottom of the well.* New York: Basic Books.

Black, A. (2002). African-American and White elites confront racial issues. *Society, 39,* 39–46.

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. *Review of General Psychology, 4,* 79–101.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 760–773.

Brunswik, E. (1949). Remarks on functionalism in perception. *Journal of Personality*, *18,* 56–65.

Chee, M. W. L., Sriram, N., Soon, C. S., & Lee, K. M. (2000). Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport, 11,* 135–140.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329.

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin, 87,* 546–563.

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

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

De Houwer, J., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorization responses. *Cognition and Emotion, 16,* 643–666.

de Tocqueville, A. (2001). *Democracy in America* (H. Mansfield, Ed., & D. Winthrop, Trans.). Chicago: University of Chicago Press. (Original work published 1835)

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

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82,* 835–848.

Dijksterhuis, A., Aarts, H., Bargh, J. A., & van Knippenberg, A. (2000). On the relation between associative strength and automatic behavior. *Journal of Experimental Social Psychology, 36,* 531–544.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11,* 315–319.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62–68.

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.

Dovidio, J. F., Mann, J. A., & Gaertner, S. I. (1989). Resistance to affirmative action: The implication of aversive racism. In F. A. Blanchard & F. J. Crosby (Eds.), *Affirmative action in perspective* (pp. 83–102). New York: Springer-Verlag.

Dunton, B. C., & Fazio. R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23,* 316–326.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske, & G. Lindsay (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 269–322). New York: McGraw-Hill.

Farmer, A., & Terrell, D. (2001). Crime versus justice: Is there a trade-off? *Journal of Law and Economics, 44,* 345–366.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15,* 115–141.

Fazio, R. H., & Dunton, B. C. (1997). Categorization by race: The impact of automatic and controlled components of racial prejudice. *Journal of Experimental Social Psychology, 33,* 451–470.

Fazio, R. H., & Hilden, L. E. (2001). Emotional reactions to a seemingly prejudiced response: The role of automatically activated racial attitudes and motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 27,* 538–549.

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.

Federal Bureau of Investigation. (1999). *Crime in the United States*. Washington, DC: Government Printing Office.

Fink, H. C. (1971). Fictitious groups and the generality of prejudice: An artifact of scales without neutral categories. *Psychological Reports, 29,* 359–365.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

Franzoi, S. L. (2000). *Social psychology* (2nd Ed.). New York: McGraw-Hill.

Garcia, L. T., Erskine, N., Hawn, K., & Casmay, S. R. (1981). The effect of affirmative-action on attributions about minority-group members. *Journal of Personality, 49,* 427–437.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: Take the best heuristic. In G. Gigerenzer, P. M. Todd, & the ABC Research Group. *Simple heuristics that make us smart* (pp. 75–95). New York: Oxford University Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Journal of Personality and Social Psychology, 109,* 3–25.

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.

Hacker, A. (1995). *Two nations: Black and White, separate, hostile, and unequal*. New York: Ballantine.

Hammond, K. R. (1996). *Human judgment and social policy*. New York: Oxford University Press.

Hammond, K., & Stewart, K. R. (2001). *The essential Brunswik*. New York: Oxford University Press.

Heilman, M. E., Battle, W. S., Keller, C. E., & Lee, R. A. (1998). Type of affirmative action policy: A determinant of reactions to sex-based preferential selection? *Journal of Applied Psychology, 83,* 190–205.

Heilman, M. E., Block, C. J., & Lucas, J. A. (1992). Presumed incompetent? Stigmatization and affirmative action efforts. *Journal of Applied Psychology, 77,* 536–544.

Heilman, M. E., Block, C. J., & Stathatos, P. (1997). The affirmative action stigma of incompetence: Effects of performance information. *Academy of Management Journal, 40,* 603–625.

Heilman, M. E., Simon, M. C., & Repper, D. P. (1987). Intentionally favored, unintentionally harmed? The impact of gender-based preferential selection on self-perceptions and self-evaluations. *Journal of Applied Psychology, 72,* 62–68.

Henderson-King, E. I., & Nisbett, R. E. (1996). Anti-Black prejudice as a function of exposure to the negative behavior of a single Black person. *Journal of Personality and Social Psychology, 71,* 654–664.

Huici, C., Ros, M., Carmona, M., Cano, J. I., & Morales, J. F. (1996). Stereotypic trait disconfirmation and positive-negative asymmetry. *Journal of Social Psychology, 136,* 277–289.

Implicit Association Test Background Information. (2004). Retrieved November 2, 2004, from https://impicit.harvard.edu/implicit/demo/racefaqs.html

Jackman, M. R., & Jackman, R. (1983). *Class awareness in the United States*. Berkeley: University of California Press.

Jacobson, C. K. (1985). Resistance to affirmative-action: Self-interest or racism? *Journal of Conflict Resolution, 29,* 306–329.

Jacobson, M. B., & Koch, W. (1977). Women as leaders: Performance evaluation as a function of method of leader selection. *Organizational Behavior and Human Performance, 20,* 149–157.

Jacoby, T. (2000). *Someone else's house: America's unfinished struggle for integration*. New York: Basic Books.

Johnson, D. (1999, June 19). Police racism charges defy a pattern. *New York Times*, p. A12.

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology, 38,* 586-602.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology, 81,* 774–788.

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin, 24,* 407–416.

Kelley, H. H. (1971). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Kelman, H. C., & Pettigrew, T. (1959). How to understand prejudice. *Commentary, 28,* 436–445.

Keltner, D., & Buswell, B. N. (1996). Evidence for the distinctiveness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expression of emotion. *Cognition & Emotion, 10,* 155–172.

Keltner, D., & Buswell, B. N. (1997). Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin, 122,* 250–270.

Keltner, D., & Harker, L. (1998). The forms and functions of the nonverbal signal of shame. In P. Gilbert & B. Andrews (Eds.), *Shame: Interpersonal behavior, psychopathology, and culture* (pp. 78–98). New York: Oxford University Press.

Kinder, D. R. (1986). The continuing American dilemma: White resistance to racial change 40 years after Myrdal. *Journal of Social Issues, 42,* 151–171.

Koren, L., & Williams, C. (1999, November 12). Recent slaying prompts cabbies to demand police protection, understanding from fares. *The Washington Times*, pp. C1–2.

Krech, D., Crutchfield, R. S., & Ballachey, E. L. (1962). *Individual in society: A textbook of social psychology*. New York: McGraw-Hill.

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., & Mellers, B. (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science, 41,* 402–419.

Locke, V., MacLeod, C., & Walker, I. (1994). Automatic and controlled activation of stereotypes: Individual differences associated with prejudice. *British Journal of Social Psychology, 33,* 29–46.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72,* 275–287.

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.

Maass, A., Montalcini, F., & Biciotti, E. (1998). On the (dis-)confirmability of stereotype attributes. *European Journal of Social Psychology, 28,* 383–402.

McConahay, J. P. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Orlando, FL: Academic.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science, 11,* 50–54.

Monteith, M. J., & Voils, C. I. (1998). Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of Personality and Social Psychology, 75,* 901–916.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19,* 395–417.

Nisbett, R. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Northcraft, G. B., & Martin, J. (1982). Double jeopardy: Resistance to affirmative action from potential beneficiaries. In B. Gutek (Ed.), *Sex role stereotyping and affirmative action policy* (pp. 81–130). Los Angeles: Institute of Industrial Relations, University of California.

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. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6,* 101–115.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math is not equal to me. *Journal of Personality and Social Psychology, 83,* 44–59.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653–667.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology, 38,* 384–396.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12,* 729–738.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz, (Ed.), *Advances in experimental social psychology, Vol. 10* (pp. 173–220). New York: Academic.

Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology, 35,* 485–494.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology, 15,* 343–355.

Rothbart, M., & John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of Social Issues, 41,* 81–104.

Rowan, C. (1996). *The coming race war in America: A wake-up call.* Boston: Little Brown.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81,* 856–868.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13,* 501–518.

Schuman, H., Steeh, C., Bobo, L., & Kyrsan, M. (1997). *Racial attitudes in America: Trends and interpretations*. Cambridge, MA: Harvard University Press.

Sears, D. O., Sidanius, J., & Bobo, L. (2000). *Racialized politics: The debate about racism in America*. Chicago: University of Chicago Press.

Sniderman, P. M., & Carmines, E. G. (1997). Reaching beyond race. *Political Science & Politics, 30,* 466–471.

Sniderman, P. M., & Tetlock, P. E. (1986). Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues, 42*, 129–150.

Snyder, M. (1981). Seek, and ye shall find: Testing hypotheses about other people. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium* (Vol. 1, pp. 277–303). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Tetlock, P. E., Kristel, O., Elson, B., Green, M., & Lerner, J. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78,* 853–870.

Thernstrom, S., & Thernstrom, A. (1997). *America in black and white*. New York: Simon & Schuster.

Tolerance.org. (n.d.) *Dig deeper: Test yourself for hidden bias.* Retrieved September 22, 2004, from http://www.tolerance.org/hidden_bias/index.html

U.S. Census Bureau. (2000). *Profiles of general demographic characteristics 2000*. Washington, DC: Government Printing Office.

von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1997). The linguistic intergroup bias as an implicit indicator or prejudice. *Journal of Experimental Social Psychology, 33,* 490–509.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797–826.

Wilson, T. D., Damiani, L. M., & Shelton, N. (1998). [Dual attitudes of Whites toward African Americans]. Unpublished raw data.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101–126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72,* 262–274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81,* 815–827.

Ybarra, O., Schaberg, L., & Keiper, S. (1999). Favorable and unfavorable target expectancies and social information processing. *Journal of Personality and Social Psychology, 77,* 698–709.

Ybarra, O., Stephan, W. G., & Schaberg, L. (2000). Misanthropic memory for the behavior of group members. *Personality and Social Psychology Bulletin, 26,* 1515–1525.