2009

Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT

Hart Blanton
University of Connecticut - Stamford

James Jaccard
New York University

Jonathan Klick
University of Pennsylvania Carey Law School

Barbara Mellers
University of Pennsylvania

Gregory Mitchell
University of Virginia - Main Campus

See next page for additional authors

Journal of Applied Psychology
2009, Vol. 94, No. 3, 567–582

0021-9010/09/$12.00 DOI: 10.1037/a0014665

Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT

Hart Blanton
Texas A&M University

James Jaccard
Florida International University

Jonathan Klick
University of Pennsylvania

Barbara Mellers
University of California, Berkeley

Gregory Mitchell
University of Virginia

Philip E. Tetlock
University of California, Berkeley

The authors reanalyzed data from 2 influential studies—A. R. McConnell and J. M. Leibold (2001) and J. C. Ziegert and P. J. Hanges (2005)—that explore links between implicit bias and discriminatory behavior and that have been invoked to support strong claims about the predictive validity of the Implicit Association Test. In both of these studies, the inclusion of race Implicit Association Test scores in regression models reduced prediction errors by only tiny amounts, and Implicit Association Test scores did not permit prediction of individual-level behaviors. Furthermore, the results were not robust when the impact of rater reliability, statistical specifications, and/or outliers were taken into account, and reanalysis of A. R. McConnell & J. M. Leibold (2001) revealed a pattern of behavior consistent with a pro-Black behavioral bias, rather than the anti-Black bias suggested in the original study.

Keywords: Implicit Association Test, predictive validity, discrimination, implicit bias

The Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) has become one of psychology's most popular exports to the wider social sciences and the law (e.g., Lane, Kang, & Banaji, 2007). The measure's popularity is easy to grasp: IAT researchers often inform test takers they harbor implicit or unconscious biases toward minorities that many test takers disavow at a conscious level (Dasgupta, Greenwald, & Banaji, 2003). According to the much-visited Project Implicit Web site, without constant vigilance these implicit biases may lead to unwanted behaviors:

> People who hold egalitarian conscious attitudes in the face of automatic White preferences may [be] able to function in nonprejudiced fashion partly by making active efforts to prevent their automatic White preference from producing discriminatory behavior. However, when they relax these active efforts, these nonprejudiced people may be likely to show discrimination in thought or behavior. (IAT Corporation, n.d., Question 16)

It is this claimed connection between implicit attitudes and discrimination that can make IAT feedback particularly disturbing to test takers. It also is the feature that makes research on the IAT of broad interdisciplinary interest. If the race IAT reliably predicts discriminatory behavior that cannot be consciously controlled, then society should take note. As but one example, the great majority of White Americans who have taken the IAT have been classified as anti-Black. This then points to an epidemic, either of unconscious racism (Greenwald & Krieger, 2006) or of false-positive accusations of unconscious racism (Mitchell & Tetlock, 2006).

Given the importance of the link between IAT scores and behavior, one might expect to find a large body of data establishing this connection—indeed, Greenwald and Krieger (2006, p. 961) described the evidence that implicit bias leads to discriminatory behavior as "already substantial." In fact, researchers in only a few studies have examined the ability of the IAT to predict behavior of any type, and researchers in just over a dozen published studies have examined the ability of the race IAT to predict outcomes that might be indicative of inappropriate or discriminatory behaviors (see Greenwald, Poehlman, Uhlmann, & Banaji, in press; Ziegert & Hanges, 2005). Yet the perception exists that the relationship between IAT scores and behavior has been much studied and well established. For instance, one prominent legal scholar, writing in the legal academy's preeminent law review, described the relationship as clear and strong:

> There is now persuasive evidence that implicit bias against a social category, as measured by instruments such as the IAT, predicts disparate behavior toward individuals mapped to that category. This occurs notwithstanding contrary explicit commitments in favor of racial equality. In other words, even if our sincere self-reports of bias

Hart Blanton, Department of Psychology, Texas A&M University; James Jaccard, Department of Psychology, Florida International University; Jonathan Klick, School of Law, University of Pennsylvania; Barbara Mellers and Philip E. Tetlock, Haas School of Business, University of California, Berkeley; Gregory Mitchell, School of Law, University of Virginia.

Authorship order was determined alphabetically.

Correspondence concerning this article should be addressed to Gregory Mitchell, University of Virginia School of Law, 580 Massie Road, Charlottesville, VA 22903-1738. E-mail: greg_mitchell@virginia.edu

> score zero, we would still engage in disparate treatment of individuals on the basis of race, consistent with our racial schemas. Controlled, deliberative, rational processes are not the only forces guiding our behavior. That we are not even aware of, much less intending, such race-contingent behavior does not magically erase the harm. (Kang, 2005, p. 1514)

This perception is reinforced in the popular-science bestseller, *Blink*, which describes the IAT as "more than just an abstract measure of attitudes. It's also a powerful predictor of how we act in certain kinds of spontaneous situations" (Gladwell, 2005, p. 85). In this article, we closely scrutinize claims that the race IAT predicts discriminatory behavior—and discover that the evidence is surprisingly weak.

Background

High Stakes and Strong Claims

Almost a half century ago, Congress passed the landmark Civil Rights Act of 1964. Title VII of the act forbade consideration of race, color, religion, sex, or national origin in employment decisions, with the goal of promoting equal employment opportunity and the expectation that socioeconomic gains for the newly protected groups would follow, albeit with delays due to historical disadvantages of these groups (Fiss, 1974). Forty-five years later, controversy persists over how much progress America has made. Optimists point to dramatic positive changes in public opinion and to market gains by women and minorities within this time period: The percentages of Americans who endorse overt prejudice sentiments has plummeted (Quillian, 2006), and wages for women and African Americans have grown substantially (Bodvarsson, 2001; Juhn & Potter, 2006). Pessimists point out that wages for Blacks in particular stagnated in the 1980s (Bodvarsson, 2001; O'Neill, 1990) and that large differences in socioeconomic outcomes persist, such as with respect to incarceration rates, educational achievement, and health outcomes (Franks, Muennig, Lubetkin, & Jia, 2006; Lucas & Paret, 2005; Western & Pettit, 2005). These stubborn disparities challenge the depth and sincerity of the newfound public support of racial equality in principle—and have led some social scientists to argue that prejudice has merely assumed more subtle but equally toxic forms (see Dovidio, 2001; Kang & Banaji, 2006; Quillian, 2006; Rudman, 2004).

Placed in this historical context, it becomes clear why few psychological debates have higher policy stakes than those over the pervasiveness and potency of implicit prejudice: There is a direct logical—and legal—link between assessments of the lingering power of prejudice and assessments of how much society in general, and organizations in particular, should do to ensure equality of opportunity in workplaces. If, as Kang and Banaji (2006) argued, implicit prejudice biases the judgments of 75% or more of Americans and leads to discriminatory treatment at either the level of nonverbal behavior or overt personnel actions, then it is important to reexamine the types of policies and procedures organizations need to block the influence of this newly discovered form of prejudice. Such reexamination could lead judges to look at evidence in litigation differently, lead regulators to reform their best-practice recommendations, and lead legislators to reconsider the role of intentionality in antidiscrimination law. In short, if the claims being made for the pervasiveness and behavioral impact of implicit prejudice are correct, then wholesale changes to American law and American ways of doing business may be in order (see Blasi & Jost, 2006; Jolls & Sunstein, 2006).

Strong claims, however, require strong evidence—and this is all the more so when there are political and legal temptations to exaggerate or trivialize what researchers uncover combined with public commitments to particular viewpoints (e.g., Lodge & Taber, 2005; Munro, Leary, & Lasane, 2004; Tetlock, 2005). Not surprisingly, disputes have erupted in both psychological and legal journals over both the validity of unconscious-prejudice measures and the applicability of unconscious-prejudice research to real-world controversies (e.g., Arkes & Tetlock, 2004; Bagenstos, 2007; Banaji, Nosek, & Greenwald, 2004; Blanton & Jaccard, 2006; Greenwald, Rudman, Nosek, & Zayas, 2006; Mitchell & Tetlock, 2006). The debates involve a wide range of issues, including (a) construct-validity questions about what exactly implicit measures measure; (b) criterion-variable questions about the degree to which correlations between implicit measures of prejudice and outcome variables should count as support for a pervasiveness-of-prejudice interpretation or alternative more benign explanations; and (c) external-validity questions about the degree to which lab demonstrations of the alleged predictive power of measures of implicit bias will hold up in work settings in which decision makers work under accountability constraints, have incentives to assemble productive teams, and have access to large amounts of individuating information.

In this article, we shed light on two key debates within the implicit bias literature: (a) Does the race IAT reliably predict discriminatory behavior and (b) do published reports support the claim that the majority of people in the general public possess a level of bias likely to produce discriminatory behavior? Defenders of the IAT often point to studies of correlations between IAT scores and various behavioral criterion variables. They suggest these correlations validate the use of the IAT as a measure of implicit attitudes and stereotypes (e.g., Greenwald, Nosek, & Sriram, 2006) and prove its utility in predicting behavior (e.g., Greenwald & Krieger, 2006). We have argued elsewhere, however, that any purported linkage between implicit prejudice and discriminatory behavior found in these studies is weak and unstable, because it depends on measures of dubious reliability and validity and perhaps a small number of outlier respondents (Mitchell & Tetlock, 2006). Further, we have argued that researchers have not pursued the types of analyses that permit them to draw inferences about the prevalence of implicit biases (Blanton & Jaccard, 2006, 2008). The current project speaks to both of these issues.

We do not focus on the basic construct-validity question of what the IAT measures because we think scores on current forms of the IAT are confounded by too many influences to support a unified theoretical account. For instance, Uhlmann, Brescoll, and Paluck (2006) found that the IAT response patterns typically interpreted as reflective of antipathy toward Blacks may reflect sympathy for them. Rothermund and Wentura (2004; Rothermund, Wentura, & De Houwer, 2005) linked this same response pattern to differential familiarity with test stimuli. Frantz, Cuddy, Burnett, Ray, and Hart (2004) linked this pattern to respondents' fears of appearing racist (see also Vanman, Saltz, Nathan, & Warren, 2004, who found greater bias on the IAT among those with a higher motivation to control prejudice). Various nonattitudinal cognitive skills also influence IAT scores, with less cognitively proficient test takers typically appearing more implicitly racist (Blanton, Jaccard, Gonzales, & Christie, 2006; Hummert, Garstka, O'Brien, Greenwald, & Mellott, 2002; McFarland & Crouch, 2001; Mierke & Klauer, 2003). Accordingly, we adopt an applied perspective and focus on what we take to be both a more pressing and tractable question about IAT research: To what degree does alleged anti-Black bias on the IAT translate into a propensity to discriminate?

Data Analytic Considerations

Our initial goal was to obtain data from all published studies examining the IAT's power to predict workplace discrimination in actual or simulated environments. Surprisingly, however, we found only two studies reasonably precisely fitting this description—Rudman and Glick (2001), who reported their data as lost, and Ziegert and Hanges (2005), who shared their data but could not provide the raw IAT data necessary for applying standard scoring algorithms and testing underlying psychometric assumptions (Blanton & Jaccard, 2006). Accordingly, we expanded our project to encompass data from published studies that figure prominently in applications of IAT research to the law. In particular, we requested data from the specific studies cited by Kang and Banaji (2006)—in a special issue of the *California Law Review* devoted to implicit bias—to argue that the IAT predicts discriminatory behavior. We thus added Green et al. (2007; a study that focused on medical decision making and that was based on a data set that arrived too late for inclusion) and McConnell and Leibold (2001; a study that focused on the power of the IAT to predict interpersonal behavior and that was based on a data set that did arrive in time for inclusion). Because we again found the scope of relevant studies surprisingly small, we expanded our data requests to include studies methodologically similar to McConnell and Liebold (2001) and thus of potential relevance to racial bias in interviews, including Heider and Skowronski (2007), Richeson and Shelton (2005), and Shelton, Richeson, Salvatore, and Trawalter (2005), as well as to include Rudman and Ashmore (2007), which examines the relation of race IAT scores to overt acts of discrimination. We obtained only the Heider and Skowronski (2007) data from this final search (but again only after the editorial process was well underway and too late for inclusion in this study). Table 1 summarizes the data requests we made for this project and the results of these requests.[1]

The reliance on data from a small set of studies should give one pause about embracing strong claims about the predictive validity of the race IAT. Ideally, we would have performed identical statistical tests on all of the published reports and presented our summary results. It would have been valuable, for instance, to apply modern robust statistical procedures to all of the published data sets (Wilcox, 2005). Since researchers commonly claim that the published record indicates that the majority of people have implicit biases that can influence their behavior in negative ways (e.g., Greenwald & Krieger, 2006; Kang & Banaji, 2006), a summary statement about the role that statistical outliers plays in this research area would have told us whether effects being attributed to most of us are being driven by a few extreme individuals. Although we did explore robust statistical analyses in two separate studies, we cannot know whether the results from these studies represent the influence of outliers across all IAT–behavior studies because of our difficulty in obtaining other data sets for replication.

Nevertheless, there is substantial value in critically examining the data sets we did obtain. Data from McConnell and Leibold (2001) and Ziegert and Hanges (2005) play important roles in ongoing debates about the behavioral consequences of implicit biases, with the McConnell and Leibold study being particularly prominent. As noted above, Kang and Banaji (2006) cited these studies to lay a foundation for their claim that antidiscrimination law must be remade to address implicit biases, and they noted that McConnell and Leibold was "the first study to demonstrate relations among the IAT, intergroup discrimination, and explicit measures of prejudice" (p. 440). Furthermore, McConnell and Leibold is cited more often than any other study for which we requested data and appears to be the most-cited study on the IAT–behavior linkage (McConnell and Leibold has been cited over 150 times in the PsycINFO database). Ziegert and Hanges is an important study as well because, also as noted above, it is one of only two published studies examining the power of implicit measures to predict judgments in an applied context that is potentially related to discriminatory treatment, albeit in a simulated employment setting. Moreover, Ziegert and Hanges has already made its way into the legal and organizational behavior literatures as evidence that the implicit biases influence employment outcomes (e.g., Duehr & Bono, 2006; Haines & Sumner, 2006; Katz, 2006).

It is reasonable to ask what can be gained from data reanalysis that cannot be obtained from more traditional approaches to studying the robustness of effects, namely, meta-analysis or a qualitative review emphasizing methodological strengths and weaknesses of different studies. We conducted three key inquiries that required the underlying data and thus could not be done through qualitative or meta-analytic reviews. First, we formally explored the role of outliers on conclusions using modern-day robust methods of analysis (Wilcox, 2005). Second, we examined the effect of aggregation and data transformations on the reported trends in the data. Third, we examined the functional form of the relationship between individual IAT scores and individual behaviors and the incremental predictive validity provided by the IAT.

[1] Greenwald et al. (in press) identified 12 published articles involving some version of the race IAT and discriminatory behavior broadly defined, including indicators of brain activity on observing racial stimuli. Our requests included a number of the studies included in the Greenwald et al. (in press) meta-analysis as well as studies that would seem to fall within the scope of the meta-analysis but were omitted (Heider & Skowronski, 2007; Richeson & Shelton, 2005; Shelton et al., 2005; Ziegert & Hanges, 2005). Had we been able to obtain all of the data initially sought, then we would have attempted to include in this project data from all of the published studies involving behavioral and judgment criterion variables identified by Greenwald et al. (in press), although not the studies using neurological indicators as criterion variables because such studies do not directly address the linkage between IAT scores and behavior or judgments. Nonetheless, our findings still qualify the conclusions reached in Greenwald et al. (in press), which omitted seemingly relevant studies and took the reported results of McConnell and Leibold (2001) at face value. We continue to seek data for use in what we hope will be a more comprehensive future analysis, and, as noted in the text, we recently obtained two additional data sets for inclusion in a follow-up project.

Table 1
Requests for Data on IAT–Behavior Correlations

Study	Topic	Status
Green et al. (2007)	Prediction of biased medical decisions by race IAT	Provided data during editorial process
Heider & Skowronski (2007)	Prediction of subtle discrimination in interpersonal interactions by race IAT; prediction of biased behavior in Prisoner's Dilemma game by race IAT	Provided data during editorial process
McConnell & Leibold (2001)	Prediction of subtle discrimination in interpersonal interactions by race IAT	Provided data
Richeson & Shelton (2005)	Prediction of subtle discrimination in interpersonal interactions by race IAT	In continuing discussions to try to obtain data
Rudman & Glick (2001)	Prediction of biased ratings of hypothetical job candidates by gender IAT	Data unavailable
Rudman & Ashmore (2007)	Prediction of self-reported acts of discrimination and support for cuts to different student groups by IATs	Data unavailable
Shelton et al. (2005)	Prediction of subtle discrimination in interpersonal interactions by race IAT	Data unavailable
Ziegert & Hanges (2005)	Prediction of biased ratings of hypothetical job candidates by race IAT	Provided transformed IAT data and all other data; original raw IAT data unavailable

Note. IAT = Implicit Association Test.

Reanalysis of the McConnell and Leibold (2001) Study

Description of the Study

Overview. McConnell and Leibold (2001) examined whether explicit and implicit measures of racial bias predicted racially discriminatory behavior, with discrimination defined as less friendly interactions by White participants to Black versus White experimenters. They reported that scores on the race IAT "related to biases in intergroup social interactions" (p. 440). This relationship was evident based on ratings of the interactions by "objective" judges who watched videotapes of the interactions and by the experimenters who participated in the participant–experimenter interactions.[2] We focus here on the results found in McConnell and Leibold using the judges' ratings of interactions, because these are not filtered by the experience of actually participating in the interactions, are richer in the number and types of behaviors rated, and presumably provide a direct comparison of the relative treatment by participants of persons of different races.

Procedure. Forty-one White college undergraduates participated in McConnell and Leibold's (2001) study, which was described as an experiment on word perception involving four unrelated tasks. All interactions between participants and experimenters were secretly recorded for later coding by observers.[3]

In the first task, a White female experimenter asked participants four scripted questions and told a scripted joke. This interaction lasted about 3 min. Next, participants completed a booklet of questionnaires, which included the explicit measures of prejudice. The experimenter stressed to participants to answer honestly and that their responses were private.

When participants completed the booklet, the White experimenter took the participant to a word experiment, where the participant completed the race IAT. Just before the participant started the IAT task, the experimenter looked at the clock and said her shift was over and that a new experimenter would take over after the word experiment. While participants completed the IAT, a Black female experimenter replaced the White experimenter and greeted participants at the completion of the IAT. The Black experimenter asked each participant seven scripted questions about the experiment, pausing for the participant's answers between each question and recording the responses on an interview form and told a scripted joke after the fourth question. McConnell and Leibold (2001) did not report the elapsed time for this final interaction.[4] They scored the IAT using the original algorithm (Greenwald, McGhee, & Schwartz, 1998), with larger positive scores indicating greater negativity toward Blacks than Whites.

[2] The finding with respect to experimenter ratings of interracial interactions, at least as reported by McConnell and Leibold (2001), is contrary to the finding of Shelton et al. (2005), where Black experimenters reported more favorable interactions with participants who showed greater anti-Black bias on the IAT.

[3] Forty-two persons participated initially, but 1 participant refused to allow the videotape of her interactions with experimenters to be used in the study. The text of the article suggests that both men and women participated in the study, but there was no breakdown of participants by sex nor was a test for sex differences reported.

[4] As Chugh (2004) noted, the second interaction differed significantly from the first interaction: Whereas the first interaction raised no racial issues, the second interaction involved a Black experimenter explicitly asking White participants race-related questions after the participants had completed the race IAT. Chugh incorrectly stated, however, that judges in McConnell and Leibold (2001) were blind to the race of the experimenter; in fact, McConnell and Leibold's procedure section makes clear that judges viewed both the participant and experimenter. McConnell and Leibold acknowledged that the order of tasks may have sensitized participants to racial attitudes and led to discomfort on the part of some participants. Furthermore, in the second interaction the experimenter asked three additional scripted questions, which may have impacted the quality of the interaction and certainly altered the opportunity for the expression of a number of the behaviors that the judges coded from the interactions (e.g., number of smiles and speech errors).

Two male judges viewed the videotapes of participant–experimenter interactions and rated (on 9-point scales) each participant's interaction with the White and Black experimenters with respect to the participant's friendliness during the interaction, the abruptness or curtness of the participant's responses to questions, the participant's general comfort level, how much the participant laughed at the experimenter's joke, and the amount of participant's eye contact with the experimenter. On 5-point scales, judges rated the participant's forward body lean toward the experimenter (vs. leaning away), the extent to which the participant's body faced the experimenter (vs. facing away), the openness of the participant's arms (vs. crossed arms), and the expressiveness of the participant's arms (vs. not moving at all). Judges also estimated the distance between the experimenter and the participant's chair at the end of the interaction to gauge social distance and recorded the participant's speaking time and the participant's number of smiles, speech errors, speech hesitations, fidgeting body movements, and extemporaneous social comments.

McConnell and Leibold (2001) used individual-level difference scores to define prejudice and discrimination. For example, they defined discriminatory behavior as the difference between a score representing some dimension of the interaction with the White experimenter minus a score characterizing the same dimension for the interaction with the Black experimenter. A score of zero reflected identical ratings of the participant's interactions with the White and Black experimenters. Higher scores ostensibly indicated more favorable behavior toward Whites than Blacks. This difference-score approach was adopted for almost all analyses. We show later how this approach masked important trends in the data and how transformations applied to these difference scores created additional confusion.

Results. McConnell and Leibold (2001) treated the judges' ratings of each participant's interactions with each experimenter on the dimensions of friendliness, abruptness, and comfort level as a molar judgment of interaction quality because ratings on these dimensions were supposed to capture overall interaction quality. These three ratings were averaged (once for the interaction with the White experimenter and once for the interaction with the Black experimenter), creating two molar scores, and these two scores were then standardized. Difference scores were calculated for each of these dimensions, such that more positive values reflected more positive behavior toward the White experimenter. These three mean standardized difference scores were then summed to create a single molar rating of relative interaction quality for each participant. Although the sum of the three behaviors (when scored in the same direction) yielded a statistically significant correlation with the IAT, none of the individual items forming the molar rating index did so. Overall, of the 16 behaviors rated by the two judges, only 5 correlated significantly with IAT scores (see Table 2). McConnell and Leibold treated the three target molar judgments of behavior as sufficiently correlated to justify aggregating them and focused their conclusions on the sum of the three molar behaviors. Accordingly, it is these molar judgments about interactions that we focus on in our reanalysis.

Table 2
Correlations Between Implicit Association Test Scores and Judges' Ratings of Interaction Behaviors in McConnell and Leibold (2001)

Judges' rating	r
Molar behaviors	
Judge rating of friendliness	.27
Judge rating of curtness or abruptness	−.27
Judge rating of general comfort level	.26
Other behaviors	
Judge rating of forward leaning	−.26
Judge rating of facing experimenter	−.03
Judge rating of body openness	.17
Judge rating of expressiveness	.09
Judge rating of eye contact	.25
Judge rating of seating distance	.26
Judge rating of speaking time	.51*
Judge rating of smiling	.39*
Judge rating of speech errors	.42*
Judge rating of speech hesitation	.35*
Judge rating of fidgeting	−.06
Judge rating of laughter at joke	.19
Judge rating of social comments	.32*

* $p < .05$.

Methodological Comments

Before turning to the reanalysis, we comment on two methodological limitations of the study, one pertaining to the timing of assessments and another pertaining to the training of the judges.

Timing of measures. As noted, participants completed explicit racism measures and the race IAT after interacting with the White experimenter but before interacting with the Black experimenter. Completion of these measures likely caused some participants to suspect they were participating in a study on discrimination. Consider that the explicit measures asked individuals to report how "pleasant" and "ugly" they find Black people, among other things. Participants were then presented with a 10-min task in the form of the IAT that linked stereotypical Black names (e.g., Jamal and Yolanda) and White names (e.g., Fred and Mary Ann) to positive and negative words. These experiences may have made participants feel nervous about appearing racist, particularly if they thought they had scored badly on the IAT (see Frantz et al., 2004). Because the experience of failing the IAT may cause people to act nervously around Black experimenters (as reflected by speech hesitations, speech errors, etc.), the IAT score–behavior correlation may reflect a method artifact rather than the influence of preexisting implicit attitudes on behavior.[5]

Low interjudge reliability. Charter (2003) reported a median interjudge reliability coefficient of .90 across a wide range of studies and recommended .70 as a minimal acceptable correlation between judges' ratings. Low interjudge reliability makes it diffi-

[5] This same artifact was present in Richeson and Shelton (2005) and Shelton et al. (2005), both of which administered the race IAT before participants had key interactions with Black people. Ideally, order of presentation of the attitude measures and order of interaction with the White versus Black experimenter would be counterbalanced and any order effects examined.

cult to determine whether observed differences reflect true differences among the participants or simply differences among the judges (see, e.g., Lunz, Stahl, & Wright, 1994).

McConnell and Leibold (2001) reported interjudge reliability coefficients of .48, .43, and .53 for the three items in the molar rating (they characterized these levels of agreement as good; p. 439), and only four of the reliability coefficients for the other 13 behaviors rated by the judges exceeded .65 (see Table 3 in their study). Such low interjudge reliability suggests that the judges were not given adequate training or guidance on how to code target behaviors or that the judges were subject to differential expectancy effects or motivational influences. In McConnell and Leibold, the judges knew the race of both the participant and experimenter when making ratings (contrast this procedure with that in Richeson & Shelton, 2005, where raters viewed only the target and were blind to the race of the other interacting party), and many of the ratings that the judges were asked to make explicitly drew attention to both the participant and the experimenter. Thus, it is possible that a judge's idiosyncratic expectations about White–Black interactions or a judge's own prejudices introduced systematic bias into the ratings, biases that were not sufficiently eliminated when McConnell and Leibold aggregated across judges. In any event, the low interjudge reliabilities in McConnell and Leibold suggest substantial differences in perceptions of the participant–experimenter interactions and bring into question the propriety of aggregating the two judges' ratings.

Reanalysis 1: The IAT–Behavior Correlation as a Function of a Single Judge's Ratings on a Subset of the Behaviors

We noted above that one of McConnell and Leibold's (2001) main findings was a statistically significant correlation between participants' IAT scores and judges' aggregated molar ratings of the overall interaction quality between White participants and White versus Black experimenters. Because the interaction quality variable was composed of two distinct ratings (molar ratings of the quality of the interaction with the White experimenter and molar ratings of the quality of the interaction with the Black experimenter), the significant IAT–molar correlation may have been driven by the treatment of the White experimenter, the treatment of the Black experimenter, both, or neither. Also, because molar ratings were made with two distinct judges who disagreed often in their ratings, the correlation between IAT scores and molar ratings may reflect different judgment contexts rather than differences driven by implicit attitudes. At a minimum, substantively different patterns across the two judges would indicate that the results are more fragile or conditional than McConnell and Leibold's aggregated results suggest.

Table 3 presents the IAT–behavior correlation as a function of judge and race of the experimenter. Whereas the molar ratings by Judge 2 correlated statistically significantly with IAT scores (r = .38, $p < .05$), those of Judge 1 did not (r = .19). Further, the molar ratings for Judge 2 correlated with IAT scores primarily due to how this judge assessed participants' interactions with the White experimenter compared to the Black experimenter, but neither correlation was statistically significant by traditional standards (r = –.27 vs. r = .10).[6] It is noteworthy, also, that the IAT–behavior correlation depended more on behavior toward the White experimenter than the Black experimenter. These results suggest that the significant correlation of IAT scores and molar ratings is attributable to the idiosyncratic reactions of one judge.[7]

Table 3
Correlation of Molar Score With Implicit Association Test Score From McConnell and Leibold (2001)

Judges' score	r	p
Judge 1		
White minus Black	.19	.24
White experimenter	–.06	.72
Black experimenter	.14	.38
Judge 2		
White minus Black	.38*	.02
White experimenter	–.27	.09
Black experimenter	.10	.54

* $p < .05$.

Reanalysis 2: The IAT–Molar Rating Correlation Is Sensitive to Outlier Effects

We noted earlier that past studies have found that individuals with slower processing speed (e.g., older individuals, persons with low IQ) tend to be diagnosed on the IAT as more implicitly biased than those who respond more quickly, even though they are not necessarily more biased (see Hummert et al., 2002; McFarland & Crouch, 2001). In recognition of this problem, in 2003 the creators of the IAT introduced a new system for scoring the IAT that allegedly corrects for the processing confound (but see Blanton & Jaccard, 2006, 2008). Because McConnell and Leibold published in 2001, they used the original method for scoring the IAT and thus did not remove this potential source of method variance from IAT scores before conducting statistical tests.

However, participants in McConnell and Leibold (2001) did report their age, making it possible to test for the influence of at least this documented contaminant of IAT scores (Hummert et al., 2002). Analysis of the age distribution of participants revealed 1 participant whose age placed her more than five standard deviations above the mean age of study participants (participant age of 50 compared to modal and mean ages of approximately 19). Consistent with the finding that IAT scores increase with age, this older participant's IAT score was 2.5 standard deviations above the mean IAT score for the sample. When this participant was excluded from the data analysis, the correlation between IAT scores and judges' molar ratings of participants' interactions with experimenters fell from significance (r = .32, $p < .05$) to nonsig-

[6] When the IAT was regressed on the two judges' ratings in the two conditions (White and Black experimenter), the multiple R was significant (r = .41, $p < .03$). However, when the ratings of Judge 2 from the White experimenter condition were dropped, the multiple R became nonsignificant (r = .20, $p > .65$). This finding is consistent with the proposition that the race IAT–behavior correlation is primarily due to the behavioral ratings of Judge 2 for the White experimenter.

[7] If the sample size in the McConnell and Leibold (2001) study (N = 41) were larger, one could test whether an observed correlation for the one judge differs significantly from the corresponding observed correlation for the other judge.

nificance ($r = .24$, $p < .13$).[8] This drop in statistical significance is not attributable to a loss of statistical power (because only one degree of freedom was lost by excluding the individual). Figure 1 presents a scatter plot of IAT scores and judges' aggregate molar ratings. Note that 2 participants anchor the lower and upper quadrants of the scatter plot, with the bulk of the cases in between showing no discernable linear relation between the constructs in question. Examination of the correlation between IAT scores and molar ratings using the percentage bend correlation, an outlier resistant correlation index (Wilcox, 2005), revealed that the original correlation was dependent on outliers ($r_{pb} = .24$, $p < .13$).

Reanalysis 3: The IAT Data Suggest a Pro-White Bias But the Behavioral Data Suggest a Pro-Black Bias

Ninety percent (37 of 41) of McConnell and Leibold's (2001) participants had positive IAT scores (the IAT is traditionally coded such that positive scores reflect relatively slower reaction times when African American stimuli are paired with positive words and/or when European American stimuli are paired with negative words; Greenwald et al., 1998). If 90% of respondents are "biased" against Blacks and the IAT metric is not arbitrary (see Blanton & Jaccard, 2006), one would expect a large proportion of the sample to discriminate against the Black experimenter. But that was not the case. In fact, over 70% of the sample acted more positively toward the Black experimenter.

This trend in the data was not apparent in McConnell and Leibold's (2001) reporting of results due to a data transformation and the reporting of results using only this transformed data. As noted above, McConnell and Leibold defined discrimination as the difference in molar ratings of interactions with the White versus the Black experimenter. Thus, more positive scores would indicate more positive interactions with the White experimenter, and the sample mean difference score on the molar judgment variable should be positive. However, because McConnell and Leibold standardized the two separate molar ratings for each interaction, each separate molar rating had a mean of zero and the mean of the summed molar ratings approximated zero. Only by returning these two scores to their original metric and recomputing the molar difference score were we able to explore whether the White or Black experimenter was perceived as receiving better treatment from the participants. We found that, on average, the judges assigned more favorable molar judgment ratings to the interactions with the Black experimenter than to that with the White experimenter ($M = 3.80$ vs. $M = 2.94$, respectively). This difference was statistically significant, $t(40) = 4.28$, $p < .001$, $\eta_p^2 = .31$.

When one examines the untransformed data, it appears that those with higher IAT scores were the least behaviorally biased in the sample. As IAT scores increased, the tendency to behave in a biased way (against Whites relative to Blacks) decreased. This trend can be seen by regressing the unstandardized molar judgment difference scores onto the IAT score, as shown in Figure 2. This graph estimates differences in the molar treatment of the White versus the Black experimenter for individuals ranging from 2 standard deviations above and below the mean IAT score ($M = 0.18$, $SD = 0.13$). Negative molar difference scores indicate a tendency to act more positively toward the Black rather than White experimenter.

As Figure 2 shows, individuals tended to be most biased behaviorally against Whites across most of the IAT distribution. Participants at the high end of the IAT distribution showed little or no racial bias (i.e., difference scores near zero). Thus, it is not accurate to say that high IAT scores predicted discrimination against the Black experimenter. Instead, high IAT scores appear to have predicted more egalitarian behavior toward both the Black and White experimenters, and lower scores appear to have predicted more discrimination toward the White experimenter. There is thus a disconnect between the attitudinal and behavioral data, and the usual interpretation given to the McConnell and Leibold (2001) study as showing that the IAT predicts discrimination against Blacks is dubious.[9]

McConnell and Leibold (2001) provided no rationale for standardizing the molar ratings, which does not appear to be standard practice within similar interaction studies (see, e.g., Bargh, Chen, & Burrows, 1996; Heider & Skowronski, 2007; Richeson & Shelton, 2005; Shelton et al., 2005), and we see no rationale for doing so given that the interest was in the relationship between IAT scores and observed behavior of individuals across the spectrum of

[8] Table 3 in McConnell and Leibold (2001) reported a correlation of $r = .34$ rather than the $r = .32$ that our reanalysis yielded. Both correlations were significant at the .05 level before elimination of the outlier. We could replicate the correlation reported in Table 3 using the IATFULL variable from the McConnell and Leibold database, but we could not replicate the IAT effect correlations reported in Table 2 of McConnell and Leibold using the IATFULL variable. However, we were able to replicate the results reported in Table 2 using the IATTRIM variable from McConnell and Leibold's database. (The IATFULL variable included Blocks 3 and 4 vs. Blocks 6 and 7 from the IAT session, whereas the IATTRIM variable drops Blocks 3 and 6 but keeps Blocks 4 and 7. The differences in results produced by the variables were small.) We focused our reanalyses on the IATTRIM variable that was used to report the IAT effect.

[9] For example, Chugh (2004, p. 211) read McConnell and Leibold (2001) as showing an "implicit bias favoring Whites" in participants' nonverbal behaviors. Chugh's reading is perfectly reasonable given that McConnell and Leibold reported that "as participants' IAT scores reflected relatively more positive attitudes toward Whites than Blacks, social interactions were more positive toward the White experimenter than the Black experimenter as assessed by both trained judges and by the experimenters themselves (p. 439)." Although it would not salvage the result as reported by McConnell and Leibold, one could argue that the IAT is still predictive of discriminatory behavior against Blacks, because, even though the Black experimenter was treated more positively than the White experimenter for most individuals, those with a lower IAT score tended to behave even more positively toward the Black experimenter relative to the White experimenter. Of course, such a finding hardly indicates that the IAT predicts discriminatory behavior against minorities, which, at a minimum, would require that a White person be treated more favorably than a Black person. It is possible that the Black experimenter was more likable than the White experimenter and that those with low IAT scores were more open to liking the Black experimenter than those with higher IAT scores. But presumably McConnell and Leibold controlled for relative likeability of the experimenters. If not, then this uncontrolled variable confounds the interpretation of McConnell and Leibold's findings. It also is quite possible that the computed difference score for both the judges' ratings and the IAT difference score have arbitrary zero points, such that neither value should be used to estimate the degree of anti-Black bias in the sample. At a minimum, however, problems with the judges' ratings point to the need for caution in interpretation.

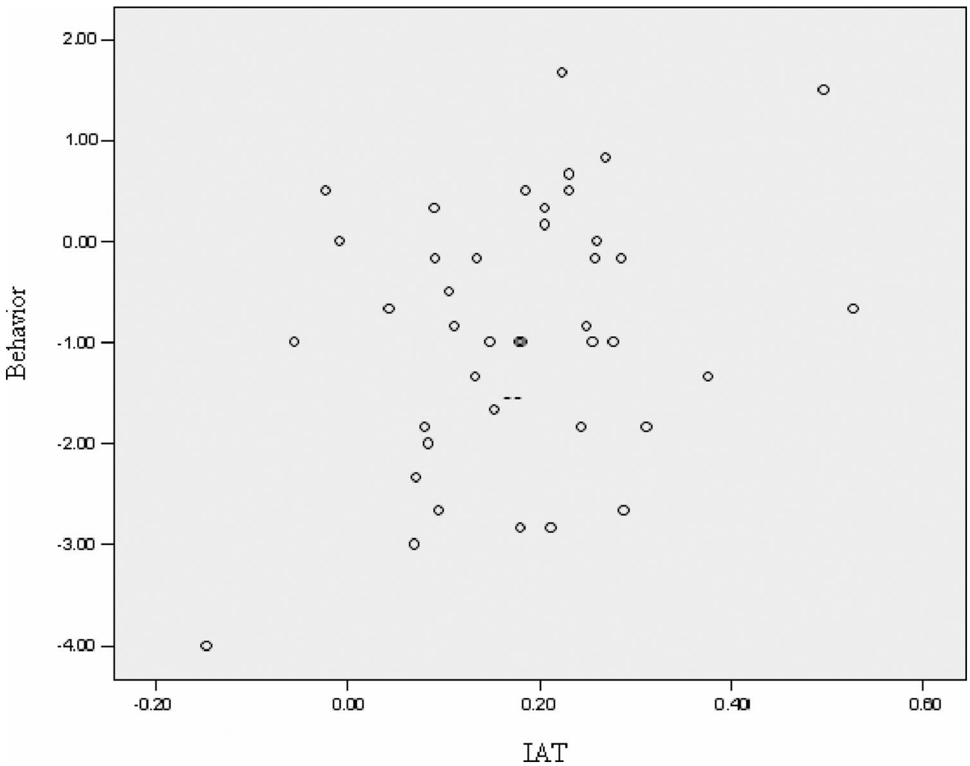

Figure 1. Scatter plot of unstandardardized molar judgment ratings and Implicit Association Test (IAT) scores from McConnell and Leibold (2001). The *y*-axis (behavior) refers to the difference in overall interaction quality for treatment of the White versus Black experimenter as measured by unstandardized molar judgments. A positive score represents more positive interaction with the White than Black experimenter (as rated by independent judges). Positive IAT scores reflect greater relative negativity toward Blacks.

behaviors as defined by the rating measures (which McConnell and Leibold treated as meaningful measures of interaction quality) rather than the behavior of individuals relative to the mean level of discrimination within a sample.[10] Indeed, using this approach effectively ensures relative differences in behavior that may appear equally discriminatory despite the actual nature of, and differences in, the interactions.

In any event, the key point is that the standardization of judges' ratings obscured the nature of participants' behavioral preferences as originally rated by the judges, which might cause readers to draw inaccurate conclusions: that the study documented disparate treatment of Blacks relative to Whites (e.g., Kang, 2005, p. 1514) and that the IAT predicted behavioral tendencies that will likely "disadvantage Black job applicants" (Greenwald & Krieger, 2006, p. 962). Such inferences do not appear warranted. Moreover, by focusing readers' attention only on the tendency for 90% of the sample to show IAT scores that IAT researchers traditionally interpret as indicative of an anti-Black implicit bias and by not at the same time reporting the corresponding tendency for 70% of the sample to act more positively towards the Black experimenter than the White experimenter, the published report could give readers the mistaken impression that the distribution of IAT scores in the study correctly characterized the behavioral tendencies of the study sample. Such was not the case.

Reanalysis 4: The Predictive Utility of an IAT Score

Following Blanton and Jaccard's (2006) call for indexing the IAT metric to observed behaviors to make the IAT metric less arbitrary, one may ask whether knowing a person's race IAT score allows one to forecast with any degree of precision how that

[10] A reviewer raised the possibility that McConnell and Leibold (2001) standardized the judges' ratings to control for a lack of equivalence in the interactions with experimenters (i.e., the White experimenter asked fewer scripted questions, and presumably interacted less with the participant, than did the Black experimenter). Although these differences may well have affected the dependent variables that involved behavioral counts (e.g., number of smiles and speech errors by participants during their interactions), these differences should not have affected the molar judgment ratings, which were supposed to be ratings of overall interaction quality. Even if we assume the molar judgments were biased by the differences in interaction content and duration, the standardization approach taken by McConnell and Leibold—standardizing molar judgments of interactions with the White experimenter, standardizing molar judgments of interactions with the Black experimenter, and then computing difference scores from these two standardized scores—could not correct any bias in the difference score. This is because McConnell and Leibold's approach merely put each separate interaction score onto its own standardized metric (each with a mean of zero). This method guarantees that the two interactions will be rated as equivalent by the sample of judges.

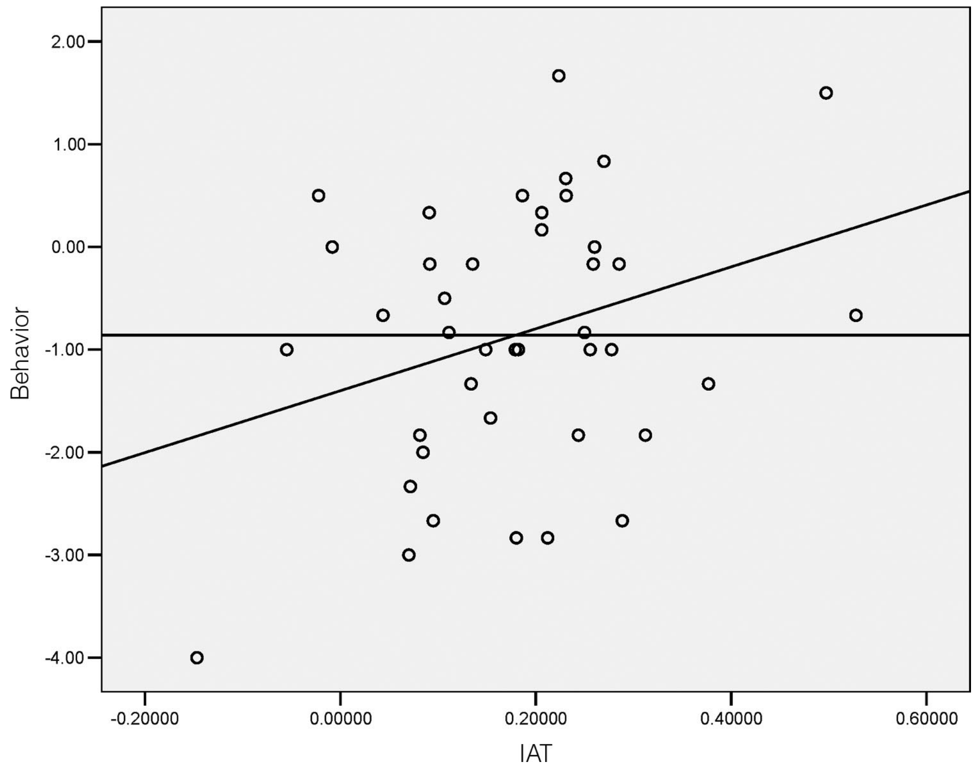

Figure 2. Scatter plot of unstandardized molar judgment ratings and Implicit Association Test (IAT) scores with predicted values (sloped line) and sample mean (flat line) from McConnell and Leibold (2001). The *y*-axis (behavior) refers to the difference in overall interaction quality for treatment of the White versus Black experimenter as measured by unstandardized molar judgments. A positive score represents more positive interaction with the White than Black experimenter (as rated by independent judges). Positive IAT scores reflect greater relative negativity toward Blacks.

person will act in the presence of African Americans. Analysis of McConnell and Leibold's (2001) individual-level data sheds light on this question, as well as on the question of whether IAT researchers should give individual test takers feedback that labels them racially biased. Consider that one of the participants in this study had an IAT score of 0.27 and a molar rating of 0.83. This means that this individuals' behavior tended to be 0.83 of a unit more positive toward the White experimenter than the Black experimenter. Another individual also had an IAT score of 0.27, and this individual's molar behavior score was –1.00. This means that the latter individual's behavior tended to be one unit more negative toward the White experimenter than the Black experimenter. Although these two individuals had identical IAT scores suggesting implicit bias against Blacks, their discriminatory behavior was in opposite directions.

We can more formally examine the predictive utility of the IAT at the individual level using prediction intervals derived from IAT scores (prediction intervals take into account, among other things, the amount of unexplained variance in the criterion, the number of predictors, and the overall sample size; for elaboration, see Neter, Kutner, Nachtsheim, & Wasserman, 1996). We calculated the 95% prediction interval for a participant with an IAT score of 0.21 and found an interval ranging from –3.3 to 1.8, which means that we can be 95% confident that an individual with an IAT score of 0.21 will have a molar rating score somewhere between –3.3 and 1.8.[11] This wide prediction interval is to be expected given that 90% of the variation in behavior has nothing to do with the IAT ($1 - r^2 = .90$). Note also that the prediction interval contains the value of zero (the value that supposedly indicates identical treatment of the Black and White experimenter). We thus cannot say with any confidence whether a person with an IAT score of 0.21 would discriminate against Blacks, Whites, or neither given this prediction interval spanning negative to positive numbers.

The same result holds for all IAT scores in McConnell and Leibold's (2001) study. We calculated 95% prediction intervals for the value of every individual within the McConnell and Leibold study, and in each case the interval spanned zero. For example, for the highest IAT score observed (0.53), the 95% prediction interval was –2.6 to 2.9 and, for the lowest observed IAT score (−0.15), the prediction interval was −4.6 to 0.9. If we lower our confidence interval to 80%, the prediction intervals still span zero for the IAT scores of 40 of the 41 individuals in McConnell and Leibold's study. The one exception was the case of an IAT score of −0.15, obtained by the participant anchoring the lower quadrant of Fig-

[11] This is an informal characterization of the prediction interval. As with all confidence intervals, the strict interpretation is in terms of the true population value being contained within the calculated interval across an infinite number of replications of the experiment.

ure 1, about whom we can say with 80% confidence that she will behave more positively toward Blacks than Whites, albeit only in a setting with the characteristics of the McConnell and Leibold study.

Reanalysis 5: Predicting Behavior With and Without the IAT

Yet another perspective on the predictive utility of the IAT can be gained by considering its standard error of estimate. If the IAT is an individual-difference measure of implicit prejudice and a related propensity to discriminate, then high IAT scorers should discriminate more against Blacks than do low scorers. One way to test whether this relationship holds, and with what degree of precision, is to examine the level of discrimination predicted by IAT scores and the level of discrimination that actually occurred within a sample. This comparison can be obtained by using linear regression to generate predicted discrimination scores for participants based on IAT scores and comparing the predicted to the actual discrimination score for each participant.

The full range of observed IAT scores is shown on the *x*-axis of Figure 2, with molar ratings of the relative quality of experimenter interactions along the *y*-axis. This scatter plot replicates that in Figure 1 but adds a regression line predicting the behavior of participants from IAT scores and a line reflecting the sample mean. Although the sloped regression line captures the general trend suggested by the data (when outliers are included in the analysis), it is generally a poor predictor of behavior. Consider as just one example the lowest IAT scorer within the sample. For this individual, the IAT's prediction was off by 2.16 units (actual molar rating of −4.00 compared to predicted molar rating of −1.84).

If discrepancy scores are computed for all participants, one can calculate the average discrepancy between predicted and observed scores (or one can use instead the standard error of the estimate, which is analogous to this average discrepancy score and which we use here). For the McConnell and Leibold (2001) data, the average discrepancy was 1.24 units. Thus, if one were to use the IAT to predict a given participant's behavior, estimates would tend to be wrong by about 1.25 points on the behavior metric (which had a possible range of −8 to 8 but an actual range for this sample of −4.00 to 1.6). In comparison, if one simply predicts that each person will show a level of bias equal to the average bias in the sample, one does almost as well at predicting behavior. In the McConnell and Leibold study, participants on average favored the Black participant by 0.86 units ($M = -0.86$) according to the judges' observations, and using this mean score to predict behavior for every participants leads to an average discrepancy of 1.32 units.

Thus, a model with no predictors performed just 0.08 behavioral scale units worse than the model with the IAT as a predictor. This 0.08-unit improvement occurs for a criterion with an 18-point range (and a functional 8-point range in the current study), meaning that the IAT improved prediction by less than half of 1% of the possible range of scores. One could achieve virtually the same level of predictive accuracy achieved with the IAT as a predictor, and save the resources associated with administering and scoring the IAT, by assuming that everyone in the study will exhibit an average level of discrimination. This finding corroborates the prediction interval finding: McConnell and Leibold's (2001) data do not present a strong case for the predictive utility of the IAT.

Reanalysis of the Ziegert and Hanges (2005) Study

Description of the Study

Overview. Ziegert and Hanges (2005) sought to test whether "implicit racist attitudes interacted with a climate for racial bias to predict discrimination" (p. 553). They predicted that higher levels of implicit racism would result in discrimination against Blacks only in a climate that tolerates discrimination. Thus, they did not propose as strong a role for implicit racism as did McConnell and Leibold. Whereas McConnell and Leibold argued that IAT scores would predict racial preference even in the absence of any prompting, Ziegert and Hanges argued that a climate for bias was necessary for implicit prejudice to translate into discrimination (and compare Vanman et al., 2004, which found no relation between IAT scores and judgments about candidates for a teaching fellowship but did find a relation between these judgments and activity of muscles used in facial expressions as measured by electromyography).

Procedure. Ziegert and Hanges replicated the procedure used by Brief, Dietz, Cohen, Pugh, and Vaslow (2000), but they supplemented the design with an implicit measure of racial prejudice in the form of the race IAT. One hundred and three non-Black participants played the role of a manager who was given the task of evaluating eight job applicants in a hypothetical company for an open vice president position. The managers made their candidate evaluations on a 5-point scale ranging from 1 (*should not have been referred*) to 5 (*excellent referral*).[12] Dossiers for the hypothetical job candidates provided information about each candidate, including race information, with three candidates being Black and three White. Ziegert and Hanges used an outcome variable equivalent to the mean of the ratings for the three Black applicants minus the mean of the ratings for the three White applicants.

Ziegert and Hanges (2005) manipulated the conditions under which participants acting as managers rated the applicants. Half of the participants were randomly assigned to the *climate for equality* condition, in which they received a memo from the hypothetical president of the company indicating that candidates should be evaluated on the basis of their education and experience. The other half of the participants were assigned to the *climate for racial bias* condition, in which participants were informed that candidates should be evaluated on the basis of their education and experience but also were told

> Given that the vast majority of our workforce is White, it is essential we put a White person in the [vice president] position. I do not want to jeopardize the fine relationship we have with our people in the units. Betty (the outgoing vice president) worked long and hard to get those folks to trust us, and I do not want her replacement to have to overcome any personal barriers.

[12] Ziegert and Hanges (2005) collected data on 103 participants, but 6 participants had missing values on one or more of the main variables. These participants were excluded from Ziegert and Hanges's analyses and ours as well, to maintain consistency between the data sets, with a resulting sample size of 97.

After completing the candidate evaluations, participants completed the race IAT, which served as the implicit measure of prejudice. Participants had previously completed explicit measures of prejudice as part of a mass testing packet completed by introductory psychology students.

Reported results. Ziegert and Hanges (2005) found that scores on the race IAT, but not scores on the explicit prejudice measures, correlated significantly with candidate ratings in the climate for racial bias condition, with higher IAT scores associated with more negative ratings of Black relative to White candidates. Ziegert and Hanges found no such association in the climate for equality condition. The interaction contrast comparing regression coefficients in the two conditions using the IAT score as a predictor of candidate evaluations was statistically significant. This finding supported Ziegert and Hanges's hypothesis that the impact of implicit prejudice would be detectable in the climate for racial bias but not in the climate for equality.

Methodological Comments

Three aspects of Ziegert and Hanges's (2005) methodology limit the generality of their finding.

Experimental manipulation. The instructional set in the climate for racial bias, as presented above, was blatantly racist and was delivered in a hypothetical, role-playing setting. Thus, one may question the ability of this study to speak to the everyday relationship between implicit prejudice and discriminatory behavior. Kang and Banaji (2006) acknowledged that the manipulation used by Ziegert and Hanges (2005) "seems unrealistic because such preferences are no longer written down" (p. 1063), but they argued that this artificiality was a strength of the study because "the outlandishness of the request should have worked *against* finding any behavioral correlation" (p. 1063). Of course, one could just as easily argue that this artificiality is a weakness because the heavy-handedness of the request created demand characteristics (Gaes, Kalle, & Tedeschi, 1978; Orne, 1962).

Applicant equivalence. Ziegert and Hanges (2005) did not test whether their participant population considered their six hypothetical candidates to have equivalent qualifications before assigning race to the candidates. Instead, it was assumed that Brief, Buttram, Elliott, Reisenstein, and McCline's (1995) pretesting of these materials with another sample and the random assignment of race and sex to the hypothetical candidates sufficed to equalize the candidates. That left open the possibility that the three applicants labeled Black were perceived as less qualified than those labeled White. Ziegert and Hanges stated that all six candidates were constructed to have "outstanding qualifications" and that "prior work has shown that there are no differences among these six candidates when race information is removed" (p. 556), but an examination of ratings of the candidates by Ziegert and Hanges's participants does not support this conclusion.

If the six candidates were equally qualified, one would expect no significant differences in the evaluations of these candidates holding race constant. However, there were significant differences in how the Black candidates were rated, $F(2, 96) = 31.08$, $p < .001$, such that one Black candidate was rated significantly higher ($M = 4.18$, $SD = 0.87$) than the other two ($M = 3.53$, $SD = 0.98$ vs. $M = 3.52$, $SD = 0.90$), $t(96) = 7.86$, $p < .001$, $\eta^2 = .15$. There were similar differences among the White applicants, $F(2, 93) = 14.93$, $p < .001$, such that one White applicant received significantly higher ratings ($M = 4.60$, $SD = 0.62$) than the other two ($M = 4.19$, $SD = 0.82$ vs. $M = 4.16$, $SD = 0.76$), $t(96) = 5.44$, $p < .001$, $\eta^2 = .24$. These findings contradict the assertion that, but for race, the applicant profiles were comparable. Although it is possible that qualification differences within the two groups of candidates canceled out one another across the two groups and across experimental conditions, these differences confound interpretations of Ziegert and Hanges's (2005) data as showing that implicit bias drove differences in candidate ratings within the climate for racial bias.

Scoring the IAT. Ziegert and Hanges (2005) scored the IAT in a way that, to our knowledge, has never been used in another published study. They recorded response latencies for the two IAT tasks (typically referred to as the *compatible* and *incompatible* tasks) and recorded the error rates for these two tasks (i.e., how often individuals made classification errors). They then standardized both the reaction time scores and the error scores and averaged these two indices together, before computing the IAT difference score. They did not transform their reaction times using standard methods to address known IAT confounds (see Greenwald, Nosek, & Banaji, 2003), nor did they report the correlation between latencies and error scores. Since there are no published psychometric analyses of the reliability or validity of Ziegert and Hanges's unique scoring procedure, one cannot assume that the same IAT scores or overall results would have been obtained if a traditional IAT scoring algorithm had been used. Ziegert and Hanges's unique scoring method also puts the IAT on a new metric, which prevented us from investigating zero-point implications for attitudinal versus behavioral bias, as we did in the McConnell and Leibold (2001) study.[13]

Reanalysis 1: Replication of Results and the Influence of Outliers

Ziegert and Hanges (2005) used hierarchical linear modeling to test the hypothesized interaction between context and the IAT as a predictor of applicant evaluations and reported that the interaction was statistically significant ($p < .05$). In our reanalysis, we could not replicate this interaction effect using traditional standard errors ($p < .07$) or robust standard errors ($p < .06$), but Ziegert and Hanges did not provide details of their analytic strategy and thus it is possible our parameterization differed from theirs. Nevertheless, this difference in the results, though not large, does suggest a

[13] The unavailability of the raw IAT data in the Ziegert and Hanges (2005) study creates a problem for investigators who seek a deeper understanding of the individual-difference variables that might be assessed by as complex an assessment tool as the IAT. Consider, for instance, the possible predictive power not of the compatible-and-incompatible-trial, response-time differentials but of the response-time differentials that arise within the compatible and incompatible trials—and that serve as the denominator term in the latest IAT scoring algorithm. Because Ziegert and Hanges could not provide us with this information, we could not test the plausible alternative hypothesis that high IAT scorers were simply those with faster general reaction times—perhaps because they were simply more attentive or alert test takers who were also more attentive to the demand-characteristic cues linked to the racist–organizational climate manipulation.

fragility in Ziegert and Hanges's key finding from the perspective of traditional alpha levels.

We examined the scatter plots for the IAT–applicant evaluation relationship for each of the separate conditions to determine visually if there were apparent outliers whose presence might have influenced the trend of the data within conditions. As shown in Figure 3, the plots for the two groups look similar except for three outliers at the top of the scatter plot for the racial bias climate condition. Removal of these individuals from the model yielded interaction contrast p values of .252 for the traditional standard error analysis and .155 for the robust standard error analysis. More importantly, application of an outlier-resistant method for comparing regression lines in two groups, based on Theil-Senn regression methods (see Wilcox, 2005), produced a statistically nonsignificant interaction contrast ($p < .13$).

We also analyzed the simple slopes within each condition. These tests were not reported by Ziegert and Hanges (2005), but they presented a figure showing the two relevant regression lines. Using the same behavioral criterion and analytic strategy of Ziegert and Hanges but applying them separately to the conditions, we found that the slope in the climate for equality condition was statistically nonsignificant (unstandardized regression coefficient = 0.03, *ns*), but the slope in the climate for racial bias condition was statistically significant in the predicted direction (unstandardized regression coefficient = –0.20, $t[44] = 2.49$, $p < .05$). Furthermore, the percentage bend correlation between IAT scores and relative candidate evaluations in the climate for racial bias remained statistically significant (absolute $r_{pb} = .38$, $p < .02$), indicating that outliers did not affect this result. Thus, if one focuses on the IAT–judgment correlations within each condition, Ziegert and Hanges's finding is resistant to the influence of outliers.

Reanalysis 2: Decomposing Effects for White Versus Black Applicants

As with the reanalysis of McConnell and Leibold's (2001) data, we sought to decompose the effects of implicit prejudice to determine whether the tendency for the IAT to predict racial preferences was evident for both White and Black candidates considered separately. Table 4 presents the results of regression analyses predicting the evaluation of each applicant using the IAT as a predictor. The IAT was a statistically significant predictor in 2 of the 12 regression analyses that correspond to ratings of two of the three Black applicants in the racial-bias climate. This finding supports Ziegert and Hanges's (2005) predictions, in that the IAT predicted bias with respect to the treatment of two of the three Black applicants in the climate for racial bias condition.

Reanalysis 3: The IAT as a Diagnostic Tool: Prediction Intervals and the Standard Error of Estimate

As with the McConnell and Leibold (2001) study, we calculated how well one could predict discriminatory evaluations using 95% prediction intervals, focusing on individuals in the racially biased context because this was the only condition in which the IAT had predictive utility.[14] We calculated prediction intervals for every individual based on each one's IAT score and, as in the McConnell and Leibold study, the interval spanned negative to positive values and included the value of zero for every participant. For example, for the highest IAT score observed (4.30), the 95% prediction interval was –0.2 to 3.7, and for the lowest observed IAT value (–3.90), the prediction interval was –4.6 to 0.9. These data suggest the predictive utility of the IAT is limited even when individuals are directed to discriminate.

To compare the model that included the IAT with a model that had no predictors, we calculated the standard error of estimate for the IAT for only individuals in the racially biased context, which yielded an average error in prediction of 0.76. For the no-predictor model, the corresponding error rate was 0.78. Thus, the IAT reduced error by only 0.02 units on a scale that could range from –4.00 to 4.00 (and did range from –1.00 to 4.00). This small degree of error reduction is not surprising given that almost 95% of the variance in the behavioral measure was due to factors not measured by the IAT.

Discussion

Our reanalysis of two key studies on the IAT–behavior linkage failed to find a robust relationship between IAT scores and discriminatory behavior. The result that Ziegert and Hanges (2005) reported for the interactive effect of implicit prejudice and racist climate on discriminatory behavior was not robust to the influence of outliers or to specification changes in the regression analysis, but when we examined IAT–behavior correlations within each climate condition, Ziegert and Hanges's finding was resistant to the influence of outliers. We also found that the IAT is not informative as a diagnostic tool in the way that would be most natural in legal settings because no individual's discriminatory behavior could be reliably predicted from his or her IAT score. Given the unique methods employed in Ziegert and Hanges and the limited predictive validity of the IAT, the results of Ziegert and Hanges do not suggest a robust relationship between implicit bias, as measured by the race IAT, and discriminatory ratings of Black candidates.

Many interactions in social science turn out to depend on the power of independent variables to move small numbers of respondents to polarized locations on dependent measures. Thus, the conditionality of Ziegert and Hanges's results on extreme scorers in an artificial climate is not inherently troubling from a theoretical perspective. However, from an applied perspective, the Ziegert and Hanges data provide little support to those who assert broadly that the typical distribution of IAT scores shows a robust tendency for most people in most groups to favor Whites relative to Blacks or that IAT scores are good or reliable predictors of discrimination against protected groups (e.g., Kang, 2005). The IAT, as uniquely scored by Ziegert and Hanges, added very little predictive power

[14] We performed this analysis using traditional ordinary least squares regression where the dependent variable was defined as the difference between the mean evaluation of the White applicants minus the mean evaluation of the Black applicants. This is equivalent to the hierarchical linear modeling approach used by Ziegert and Hanges (2005), except the repeated measure factor of race of the applicant is treated as a fixed effect rather than a random effect. The fixed effect analysis yields lower standard errors and more narrow confidence intervals than the random effects approach, so our analyses make it more likely the prediction intervals will not contain the value of zero.

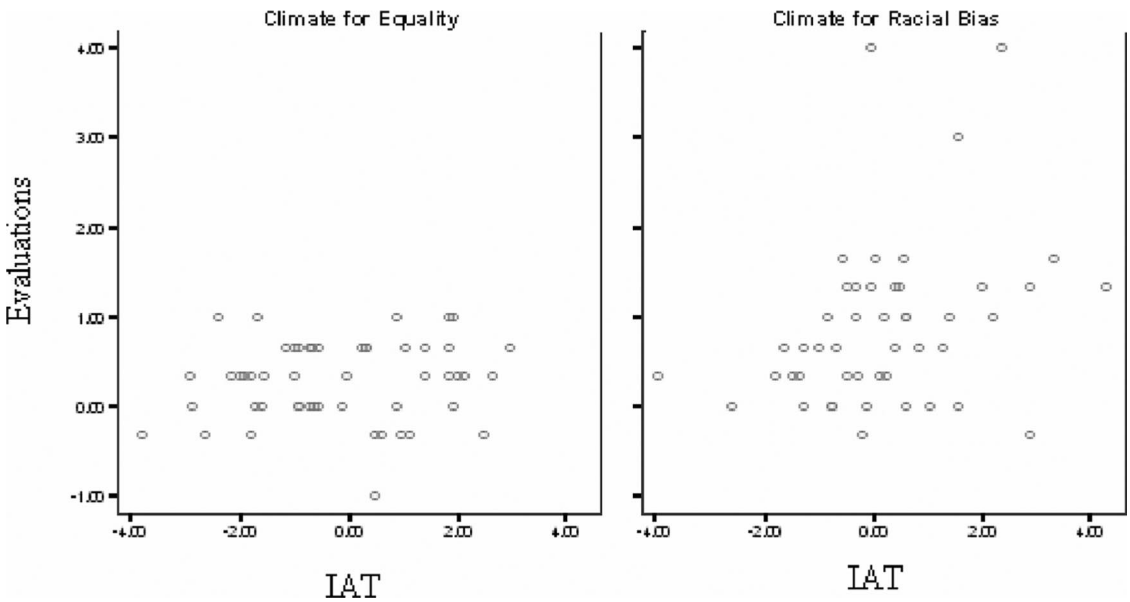

Figure 3. Scatter plot of candidate evaluations from separate evaluation climates in Ziegert and Hanges (2005). IAT = Implicit Association Test.

beyond that of a simple base rate model and could not delimit the range of likely discrimination in any meaningful way.

The results of our reanalysis of the McConnell and Leibold (2001) data were more far reaching than the results of our reanalysis of the Ziegert and Hanges (2005) data (and more significant given the prominence of McConnell and Leibold as a source of support for the claim that the IAT predicts discriminatory behavior, even outside of the kind of racist climate studied by Ziegert and Hanges). Although we could replicate (roughly) the basic finding from Ziegert and Hanges, we found their result to be fragile. When we unpacked the McConnell and Leibold data, we could not validate a basic conclusion often drawn from McConnell and Leibold's study: that judges observed greater discrimination toward the Black experimenter by White participants who were scored as more implicitly biased on the IAT. A data transformation by McConnell and Leibold obscured the nature of the judges' ratings. The untransformed data suggested that participants acted more positively toward the Black experimenter than the White experimenter and, importantly, that higher race IAT scores (which are commonly interpreted as reflecting higher levels of implicit prejudice against Blacks) were actually linked to pro-Black rather than to anti-Black behavior. In addition, the McConnell and Leibold data exhibited instability in that (a) deletion of as few as one outlier from the statistical tests altered the reported significance levels and (b) the judges relied on by McConnell and Leibold reached different conclusions about the nature of the participant–experimenter interactions, leading to poor interjudge agreement and to one judge's ratings of participants' interactions with the White experimenter having excessive influence on the results.

We only made these discoveries by asking whether it was appropriate to aggregate these data in light of the low interjudge reliabilities in McConnell and Leibold (2001) and then examining the judges' raw ratings from McConnell and Leibold. The most common justification for aggregation occurs in a psychometric context where one attempts to increase the reliability of a measure of a construct by summing (or averaging) scores across items (from either self-reports or from observer reports). If the items all reflect the same construct to the same degree, then averaging causes random noise across items to cancel out, yielding a more reliable total index (see, e.g., Cronbach, Rajaratnam, & Gleser, 1963). Summation is problematic, however, if the items reflect different constructs. As noted by Carver (1989), analyses of com-

Table 4
Regression Analyses for Individual Applicants' Evaluations as a Function of the Implicit Association Test (IAT) From Ziegert and Hanges (2005)

Applicant evaluated	*B* for IAT	R^2
Equality condition		
White applicants		
Applicant 1	.008	<.001
Applicant 2	–.063	.016
Applicant 3	.014	<.001
Black applicants		
Applicant 1	.006	<.001
Applicant 2	−.085	.025
Applicant 3	−.056	.015
Racial bias condition		
White applicants		
Applicant 1	.011	.001
Applicant 2	−.025	.003
Applicant 3	−.051	.017
Black applicants		
Applicant 1	$-.194^{*}$	$.089^{*}$
Applicant 2	$-.195^{\dagger}$	.079
Applicant 3	$-.264^{*}$	$.149^{*}$

Note. *B* = unstandardized regression coefficient.
$^{*}p < .05$. $^{\dagger}p < .10$.

binations of distinct components may cause one to infer incorrectly that effects on the aggregate replicate across the distinct components. If in an observational study a result on an aggregate is driven by a single judge, then that calls into question the generalizability of the effect and the wisdom of aggregating across judges. Indeed, if an effect occurs for some but not the majority of stimuli, or for some but not the majority of behaviors, it is possible the observed effect reflects little more than a chance result.

Although we have subjected the McConnell and Leibold (2001) and Ziegert and Hanges (2005) studies to considerable scrutiny, our concern is not the defensibility of the claims made by those specific researchers or of their methods and data-analytic techniques in a general sense. Indeed, those researchers should be commended for engaging in the scientific process by sharing their data and submitting their studies to heightened scrutiny. For a number of studies, data were not retained for relatively short time periods, not to mention for the 5 years required under American Psychological Association standards and editorial policies. Thus, arguably our most disconcerting finding was corroboration of a point made by Wicherts, Borsboom, Katz, and Molenaar (2006): Social psychologists are doing a poor job complying with the scientific norm of replication.

Our primary concern is, however, with the way a small number of studies are being used to make strong claims in applied settings, including courtrooms (see Feuss & Sosna, 2007). Although one might quibble over a particular reanalysis or the implications of a specific outlier, the broad picture that emerges from our reanalysis is that the published results are likely to be conditional and fragile and do not permit broad conclusions about the prevalence of discriminatory tendencies in American society. Given the paucity of studies showing strong links between IAT scores and behavior, given our inability to gain access to published data sets, and given the weakness of the data that we did obtain, psychologists and legal scholars do not have evidentiary warrant to claim that the race IAT can accurately or reliably diagnose anyone's likelihood of engaging in discriminatory behavior, less still that there is substantial evidence of such linkages (contra Greenwald & Krieger, 2006).

Of course, the facts may change. Indeed, we hope that our reanalyses of the McConnell and Leibold (2001) and Ziegert and Hanges (2005) studies will prompt efforts to show that the IAT predicts discriminatory behavior in both nonverbal and macrolevel behaviors and to examine more carefully the IAT metric so that researchers might get a better sense of which scores—if any—typically and reliably are indicative of anti-Black bias. For we agree with McConnell and Leibold that, at least in the domain of social psychology, "any psychological tool is only as good as its ability to predict human behavior" (p. 440). However, if the results of McConnell and Leibold and Ziegert and Hanges turn out to be representative of the (weak and qualified) relationship that exists between race IAT scores and criterion behaviors and if future demonstrations are restricted to controlled and contrived laboratory settings, then claims about behavioral implications of implicit prejudice should reflect the modesty of the record.[15]

Our reanalyses suggest some directions for future studies on the relation of IAT scores to discriminatory behavior. First, researchers should make greater use of robust statistical techniques to guard against the influence of outliers (Wilcox, 2005). In each of the studies considered here, robust analyses and scatter plots raised concerns about the replicability and generality of the results.

Second, researchers should move beyond simple zero-order correlational tests of (implicit) attitude–behavior relations. We pursued one strategy—based on data disaggregation—for understanding these relations, but the disaggregation strategy should be taken much further. For instance, we have noted elsewhere that the IAT is a limited tool for testing psychological theories because researchers typically use the composite IAT score to predict composite discrimination criteria. Consider that the IAT used in both of the studies we examined focused on a single IAT score that was influenced by the tendency to associate (a) Whites with positive concepts, (b) Whites with negative concepts, (c) Blacks with positive concepts, and (d) Blacks with negative concepts. Researchers in both studies then used this composite score to predict criteria that were themselves composite indicators of (a) evaluations of Black people and (b) evaluations of White people. These were computed as difference scores, which are notoriously misleading (Edwards, 2001). Many distinct psychological trends thus could have been driving the patterns McConnell and Leibold (2001) and Ziegert and Hanges (2005) documented but nonetheless led to the appearance of comparable influences at the level of the aggregate.

Third, researchers should examine the role of implicit attitudes after controlling for relevant explicitly measured beliefs and attitudes. Most studies investigating an IAT's predictive validity focus simply on zero-order relations between IAT scores and criteria, with no controls for explicit evaluations or constructs (see Greenwald et al., in press). One positive feature of McConnell and Leibold's (2001) study was that these researchers did collect explicit attitude data and controlled for these attitudes in some of their analyses. However, their explicit measures were brief and unusually crude for an investigation that sought to predict a complex behavior pattern. This approach to explicit measures is typical of IAT research. Studies that focus attention on the IAT often assess explicit attitudes using a few distal evaluations. Typical measures include broad semantic differentials (e.g., rating Blacks on the pleasant–unpleasant dimension) and feeling thermometers (see Greenwald et al., in press, for other examples). Researchers have long known that such measures are poor predictors of specific behaviors (Ajzen & Fishbein, 1977) and represent outdated and long-ago rejected representations of viable attitude–behavior models. Reliance on such explicit measures can thus lead to inflated estimates of the importance of implicit constructs if there is a mismatch between the types of attitudes assessed and types of behaviors examined (Jaccard & Blanton, 2006). In other applied domains, such as in health psychology, researchers who try to advance new attitudinal constructs typically are held to the standard of showing improved criterion prediction after known deter-

[15] We believe that the ideal solution to the implicit bias debates will take the form of adversarial collaborations that require both proponents and skeptics to agree on research designs for testing the rival predictions about the pervasiveness and potency of unconscious prejudice under various well-defined boundary conditions (Greenwald, 2004; Mellers, Hertwig, & Kahneman, 2001; Tetlock & Mitchell, in press). Nonetheless, replication studies are an important part of the scientific process (e.g., King, 1995) and providing a critique with data reanalysis is much more informative than one without such reanalysis.

minants of behavior—properly measured and conceptualized—have been controlled (e.g., using constructs from the theory of reasoned action and planned behavior; Ajzen, 1991). Until investigations of this type are performed, we think it premature to argue that a new form of attitude is being observed, one that cannot be assessed using traditional measurement techniques and that policy makers and other practitioners must grapple with (e.g., Kang & Banaji, 2006).

Finally, even if future IAT studies reveal robust criterion–prediction relations after researchers control for explicit attitudes and beliefs, it will be critical for researchers to develop meaningful external criteria that can be used to validate the labels given to IAT respondents and the strong inferences made in the applied literature about the meaning of locations on the IAT metric. A simplifying assumption, embraced widely in the IAT literature, is that the IAT distribution corresponds to the distribution of people's true implicit attitudes (Banaji et al., 2004). Thus, respondents with positive race IAT scores are viewed as biased in one direction (anti-Black), respondents with negative IAT scores are viewed as biased in another (anti-White), and respondents with IAT scores near zero are viewed as nonbiased. However, no research has validated the zero point of the race IAT, and our reanalyses showed that one might make a great number of errors if one tries to make even directional predictions about behavior from someone's race IAT score, much less precise estimates of the amount of racial discrimination an individual's IAT score implies.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50,* 179–211.

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888–918.

Arkes, H., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15,* 257–278.

Bagenstos, S. R. (2007). Implicit bias, "science," and antidiscrimination law. *Harvard Law and Policy Review.*

Banaji, M. R., Nosek, B. A., & Greenwald. A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15,* 279–289.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects on trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230–244.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology, 34,* 277–297.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42,* 192–212.

Blasi, G., & Jost, J. (2006). System justification theory and research: Implications for law, legal advocacy, and social justice. *California Law Review, 94,* 1119–1168.

Bodvarsson, O. B. (2001). Where have the social sciences come in understanding racial discrimination? An economist's viewpoint. *The Social Sciences Journal, 38,* 637–650.

Brief, A. P., Jr., Buttram, R. T., Elliott, J. D., Reisenstein, R. M., & McCline, R. L. (1995). Releasing the beast: A study of compliance with orders to use race as a selection criteria. *Journal of Social Issues, 51,* 171–193.

Brief, A. P., Jr., Dietz, J., Cohen, R. R., Pugh, S. D., & Vaslow, J. B. (2000). Just doing business: Modern racism and obedience to authority as explanations for employment discrimination. *Organizational Behavior and Human Decision Processes, 81,* 72–97.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology, 56,* 577–585.

Charter, R. (2003). A breakdown of reliability coefficients by test type and reliability method and the clinical implications of low reliability. *Journal of General Psychology, 130,* 290–304.

Chugh, D. (2004). Societal and managerial implications of implicit social cognition: Why milliseconds matter. *Social Justice Research, 17,* 203–222.

Civil Rights Act of 1964, Pub. L. No. 88–352, 78 Stat. 241 (codified as amended at 42 U.S.C. § 2000e-2(a) (2007)).

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Dasgupta, N., Greenwald, A. G., & Banaji, M. R. (2003). The first ontological challenge to the IAT: Attitude or mere familiarity? *Psychological Inquiry, 14,* 238–243.

Dovidio, J. F. (2001). On the nature of contemporary prejudice: The third wave. *Journal of Social Issues, 57,* 829–849.

Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology, 59,* 815–846.

Edwards, R. R. (2001). Ten difference score myths. *Organizational Research Methods, 4,* 265–287.

Feuss, C. E., & Sosna, J. D. (2007). Courts are warming to claims of "unconscious bias." *National Law Journal, 30,* S2.

Fiss, O. M. (1974). The fate of an idea whose time has come: Antidiscrimination law in the second decade after Brown v. Board of Education. *University of Chicago Law Review, 41,* 742–773.

Franks, P., Muennig, P., Lubetkin, E., & Jia, J. (2006). The burden of disease associated with being African-American in the United States and the contribution of socio-economic status. *Social Science & Medicine, 62,* 2469–2478.

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30,* 1611–1624.

Gaes, G. G., Kalle, R. J., & Tedeschi, J. T. (1978). Impression management in the forced compliance situation. *Journal of Experimental Social Psychology, 14,* 493–510.

Gladwell, M. (2005). *Blink.* New York: Little, Brown.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients. *Journal of General Internal Medicine, 22,* 1231–1238.

Greenwald, A. G. (2004). The resting parrot, the dessert stomach, and other perfectly defensible theories. In J. Jost, M. R. Banaji, & D. A. Prentice (Eds.), *The yin and yang of social cognition: Perspectives on the social psychology of thought systems* (pp. 275–285). Washington, DC: American Psychological Association.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94,* 945–967.

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

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on the article by Blanton and Jaccard. *American Psychologist, 61,* 56–61.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (in press). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*.

Greenwald, A. G., Rudman, L. A., Nosek, B. A., & Zayas, V. (2006). Why so little faith? A reply to Blanton and Jaccard's (2006) skeptical view of testing pure multiplicative theories. *Psychological Review, 113,* 170–180.

Haines, E. L., & Sumner, K. E. (2006). Implicit measurement of attitudes, stereotypes, and self-concepts in organizations: Teaching old dogmas new tricks. *Organizational Research Methods, 9,* 536–553.

Heider, J. D., & Skowronski, J. J. (2007). Improving the predictive validity of the Implicit Association Test. *North American Journal of Psychology, 9,* 53–76.

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging, 17,* 482–495.

IAT Corporation. (n.d.). *FAQs* [Question 16]. Retrieved December 15, 2008, from https://implicit.harvard.edu/implicit/demo/background/faqs.html#faq20

Jaccard, J., & Blanton, H. (2006). A theory of implicit reasoned action: The role of implicit and explicit attitudes in the prediction of behavior. In I. Ajzen, D. Albarracin, & J. Hornik (Eds.), *Prediction and change of health behavior: Applying the reasoned action approach* (pp. 69–94). Mahwah, NJ: Erlbaum.

Jolls, C., & Sunstein, C. R. (2006). The law of implicit bias. *California Law Review, 94,* 969–996.

Juhn, C., & Potter, S. (2006). Changes in labor force participation in the United States. *Journal of Economic Perspectives, 20,* 27–46.

Kang, J. (2005). Trojan horses of race. *Harvard Law Review, 118,* 1489–1593.

Kang, J., & Banaji, M. R. (2006). Fair measures: A behavioral realist revision of "affirmative action." *California Law Review, 94,* 1063–1118.

Katz, M. J. (2006). The fundamental incoherence of Title VII: Making sense of causation in disparate treatment law. *Georgetown Law Journal, 94,* 489–551.

King, G. (1995). Replication, replication. *PS: Political Science and Policy, 28,* 444–452.

Lane, K. A., Kang, J., & Banaji, M. R. (2007). Implicit social cognition and law. *Annual Review of Law and Social Science, 3,* 427–451.

Lodge, M., & Taber, C. S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology, 26,* 455–482.

Lucas, S. R., & Paret, M. (2005). Law, race, and education in the United States. *Annual Review of Law and Social Science, 1,* 203–231.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54,* 913–925.

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.

McFarland, S. G., & Crouch, Z. (2001). A cognitive skill confound on the Implicit Association Test. *Social Cognition, 20,* 483–510.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12,* 269–275.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85,* 1180–1192.

Mitchell, G., & Tetlock, P. E. (2006). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal, 67,* 1023–1121.

Munro, G., Leary, S. P., & Lasane, T. P. (2004). Between a rock and a hard place: Biased assimilation of scientific information in the face of commitment. *North American Journal of Psychology, 6,* 431–444.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

O'Neill, J. (1990). The role of human capital in earnings differences between Black and White men. *Journal of Economic Perspectives, 4,* 25–45.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776–783.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology, 32,* 299–328.

Richeson, J. A., & Shelton, J. N. (2005). Thin slices of racial bias. *Journal of Nonverbal Behavior, 29,* 75–86.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133,* 139–165.

Rothermund, K., Wentura, D., & De Houwer, J. (2005). Validity of the salience asymmetry account of the Implicit Association Test: Reply to Greenwald, Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General, 134,* 426–430.

Rudman, L. A. (2004). Social justice in our minds, homes, and society: The nature, causes, and consequences of implicit bias. *Social Justice Research, 17,* 129–142.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations, 10,* 359–372.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57,* 743–762.

Shelton, N., Richeson, J. A., Salvatore, J., & Trawalter, S. (2005). Ironic effects of racial bias during interracial interactions. *Psychological Science, 16,* 397–402.

Tetlock, P. E. (2005). *Expert political judgment.* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Mitchell, G. (in press). Implicit bias and accountability systems: What must organizations do to prevent discrimination? In B. M. Staw and A. Brief (Eds.), *Research in Organizational Behavior* (Vol. 30). New York: Elsevier.

Uhlmann, E. L., Brescoll, V. L., & Paluck, E. L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice. *Journal of Experimental Social Psychology, 42,* 491.

Vanman, E. J., Saltz, J. L., Nathan, L. R., & Warren, J. A. (2004). Racial discrimination by low-prejudiced Whites: Facial movements as implicit measures of attitudes related to behavior. *Psychological Science, 11,* 711–714.

Western, B., & Pettit, B. (2005). Black–White wage inequality, employment rates, and incarceration. *American Journal of Sociology, 111,* 553–578.

Wicherts, J., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726–727.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.

Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology, 90,* 553–562.

Received August 16, 2007
Revision received July 23, 2008
Accepted October 23, 2008 ■