Weak Criticisms and Selective Evidence: Reply to Blanton et al. (2009)

Allen R. McConnell
Miami University

Jill M. Leibold
Litigation Insights

We respond to a critique by H. Blanton et al. (2009), challenging our previous work demonstrating that an Implicit Association Test designed to assess implicit prejudice reliably predicts intergroup discrimination (A. R. McConnell & J. M. Leibold, 2001). We outline 3 flawed aspects of the critique. First, we note that claims that an outlier should be eliminated from the original data set are unfounded, and even with the elimination of this outlier, the conclusions of our original work are still strongly supported by the data. Second, we explain that concerns about interjudge reliability are specious and that considerable data support the validity of the judges’ observations reported in our original study. Third, we note that claims of a disconnect between Implicit Association Test scores (argued to show negative bias against Blacks) and behavioral measures (argued to show pro-Black bias) are inappropriate because they neglect the relativistic nature of the key measures. Implications for the relation between implicit attitudes and behavior, for the law, and for future work on implicit attitudes are discussed.

Keywords: implicit attitudes, prejudice, discrimination, Implicit Association Test

Preface

We are pleased to offer our thoughts on the Blanton et al. (2009) article. In July 2006, we were contacted by an author on that article (G. Mitchell, personal communication, July 25, 2006), who asked us for copies of the data sets for our publication, McConnell and Leibold (2001). Despite the fact that the request came 5 years after publication and 8 years after data collection, we were happy to comply. At that time, we inquired about the nature of the interest in our data (e.g., what hypotheses were being tested), but no specifics were offered. In February 2007, we received a summary of their analyses, and we informed them that it is our practice to not offer comments on work outside of the normal editorial review process. We were not provided with a copy of their manuscript until June 2008, when we were asked to write a commentary on it after it had been peer reviewed. Thus, this reply is our first opportunity to speak to a number of reservations we have about their work. Herein, we focus on three concerns that were especially striking to us, and we elaborate on the broader implications of this exchange.

Selective Focus on a Single Dependent Measure

One of the points made by Blanton et al. (2009) is that it is tenuous to make strong claims about relations between implicit attitudes and behavior based on a small amount of data. We agree that it would be preferable to see more studies demonstrate the implications of implicit attitudes for behavioral relations and, indeed, this was a primary motive behind our conducting the McConnell and Leibold (2001) study in the first place. One of the primary arguments marshaled by Blanton et al. against our work is that, in their view, our critical Implicit Association Test (IAT)—behavior demonstration hinges on 1 outlier participant, who revealed strong racial prejudice on the IAT (524-ms IAT effect, compared to the sample mean of 158 ms). Yet, it is interesting that this individual did not reveal the largest IAT effect in the entire sample (that distinction belonged to a 20-year-old woman at 578 ms). Thus, the choice of this critical outlier on statistical grounds seems questionable given that this person did not have the most extreme IAT score (a point not mentioned by Blanton et al.). Moreover, if a particular individual had an aberrant IAT score, this should make it more difficult to observe significant IAT–behavior correlations in our data set, not easier (because correlations require correspondence between two variables, and thus a deviant score on one measure would make such a relation less likely).

If this outlier did not have the most extreme IAT effect score in our sample, why do Blanton et al. (2009) claim that this individual should be discarded from the analyses? They argued her age (50 years) was problematic. We must admit that this is the first time we have encountered reservations about having a sample not composed exclusively of undergraduates aged 18–22 (there were also 2 participants in their 30s who, apparently, did not concern Blanton et al.). If one wishes to argue that nontraditional participants should be excluded from research, one must have a very compelling rationale for such a position. Blanton et al.’s argument is that older participants would have cognitive deficits (no citation or evidence is provided) and that research has shown that cognitive deficits can produce confounds on the IAT (McFarland & Crouch, 2002).

There certainly is considerable evidence that cognitive functioning declines with age (e.g., Foster & Taylor, 1920; Saltzhouse, Atkinson, & Berish, 2003), but the data also reveal that declines in normal individuals are relatively modest until around age 60, when the impact of aging becomes more pronounced (e.g., Hasher,
Stoltzfus, Zacks, & Rypma, 1991; Salthouse & Ferrer-Caja, 2003; Schroeder & Salthouse, 2004). Thus, the current critical outlier, in the absence of other evidence, is not a candidate for revealing strong cognitive deficits. Yet, Blanton et al. (2009) assert that the literature supports the position that such an individual would reveal inappropriate IAT scores. In particular, they cite McFarland and Crouch (2002), who concluded that general cognitive ability can introduce systematic method variance in the IAT, though those authors did not examine older adults in their work. But the Blanton et al. argument would require that a 50-year-old woman reveal sizable cognitive deficits, and such a premise does not comport with the existent literature (see Salthouse et al., 2003). Blanton et al. also cite an IAT study by Hummert, Garstka, O’Brien, Greenwald, and Mellott (2002) that examined older adults (mean age in their sample was 69 years) and their age-related attitudes using the IAT. Although Hummert et al. concluded that one should take general speed of responding (in the form of standardizing response latencies) into consideration when interpreting IAT results with older adults, those authors did not question the validity or utility of the IAT with such populations. Further, it is important to note that their average older participant was 19 years older than the outlier in question. Thus, it is unclear how one can argue that excluding this 50-year-old woman from the McConnell and Leibold (2001) analyses is justified. Moreover, because implicit attitudes reflect recurring paired associations between attitude objects and evaluations (e.g., Fazio, Eiser, & Shook, 2004; Jellison, McConnell, & Gabriel, 2004; Rydell & McConnell, 2006; Smith & DeCoster, 2000), it is quite probable that older individuals would, all things being equal, naturally have relatively stronger implicit prejudice against Blacks because they are more likely to have encountered Black–evaluation associative pairings during decades when cultural attitudes were even more negative toward Blacks. Consistent with this reasoning, Nosek et al. (2007) found that older individuals have the strongest racial IAT biases. Thus, rather than seeing this outlier with a relatively large bias against Blacks on the IAT as an anomaly (again, she did not have the most extreme IAT score in the sample), this logical outcome is what one would anticipate a priori.

Perhaps even more disconcerting than the questionable logic behind the argument to delete this individual from our analyses is that, upon closer examination of the data, the elimination of this participant does not really call into question the IAT–behavior finding in McConnell and Leibold (2001). It is curious that when challenging our findings, Blanton et al. (2009) focus exclusively on one single dependent measure, the judges’ molar ratings of interactions, from a pool of several measures reported in our original article that were significantly correlated with the IAT. As a brief reminder to the reader, we assessed participant behavior toward both the Black and White experimenters (separately) using a number of converging indexes borrowed from the literature (e.g., Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Fazio, Jackson, Dunton, & Williams, 1995). Specifically, the experimenters (both Black and White) evaluated the behavior of each participant using five different scales (e.g., participant friendliness, abruptness), which were combined into an overall index of experimenter perceptions. Also, trained judges watched videotaped interactions between the participant and each experimenter (Black and White) and for each made both global (molar) ratings (i.e., participant curtness, friendliness, and general comfort), and they also assessed a number of very specific behaviors associated with more positive interpersonal conduct (e.g., amount of smiling, speech errors made in interactions). Thus, there were three classes of behavioral variables (each based on difference scores between interactions involving the Black and White experimenters, such that larger scores reflected relatively more positive interactions with the White experimenter in comparison to the Black experimenter): experimenter ratings, judges’ molar ratings, and judges’ ratings of a number of specific behaviors.

Interestingly, Blanton et al. (2009) only focus on the correlation between implicit prejudice (as assessed by the IAT) and the judges’ molar ratings and not on any of the other measures of bias that were significant in our original article. These other measures were all discussed extensively by McConnell and Leibold (2001), available to Blanton et al. in the data set provided to them, and were used as measures of biased interracial interactions in past published work. In Table 1, we provide a complete listing of how

<table>
<thead>
<tr>
<th>Measure of behavioral bias</th>
<th>None eliminated (N = 41; McConnell &amp; Leibold, 2001)</th>
<th>Outlier eliminated (N = 40; Blanton et al., 2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit measure of prejudice</td>
<td>.42**</td>
<td>.42**</td>
</tr>
<tr>
<td>Experimenters’ ratings</td>
<td>.39*</td>
<td>.41*</td>
</tr>
<tr>
<td><strong>Judges’ molar ratings</strong></td>
<td>.34*</td>
<td>.27*</td>
</tr>
<tr>
<td>Participant speaking time</td>
<td>.51**</td>
<td>.45**</td>
</tr>
<tr>
<td>Smiling</td>
<td>.39*</td>
<td>.40*</td>
</tr>
<tr>
<td>Speech errors</td>
<td>.42*</td>
<td>.36*</td>
</tr>
<tr>
<td>Speech hesitations</td>
<td>.35*</td>
<td>.25</td>
</tr>
<tr>
<td>Extemporaneous social comments</td>
<td>.31*</td>
<td>.35*</td>
</tr>
</tbody>
</table>

*Note. All measures were coded such that larger, positive values reflect relatively more positive attitudes and behaviors toward Whites in comparison to Blacks. The boldfaced row is the only measure discussed by Blanton et al. (2009) in their critique of McConnell and Leibold (2001). *p < .05. **p < .01. ***p < .10.
the IAT correlated to each and every behavioral measure (and also with explicit racial attitudes) that revealed a significant correlation in McConnell and Leibold (2001) as a function of whether all participants were included (left column; as in McConnell & Leibold) or whether the outlier participant was omitted (right column; as per Blanton et al.).

A few things should be noted. First, Blanton et al. (2009) focus on a different IAT index than the one we presented in our original publication. Specifically, McConnell and Leibold (2001) reported an IATFULL index that compared the average response latency for Blocks 3 and 4 to the average response latency for Blocks 6 and 7. Yet in their reanalysis, Blanton et al. focused on a different IAT index (IATTRIM) that ignored Blocks 3 and 6, and thus relied on a comparison between Blocks 4 and 7. Although the differences between these two IAT indexes are small (e.g., the correlation between judges’ molar ratings and the IATFULL is \( r = .34, p < .05 \), and the correlation between the judges’ molar ratings and the IATTRIM is \( r = .32, p < .05 \)), it is unclear why Blanton et al. would emphasize a different measure (i.e., the IATTRIM) than the one reported in the original article that they are critiquing. Thus, in the current article (see Table 1), we focus on the same IAT index reported in McConnell and Leibold.

As Table 1 reveals, when we eliminate the outlier, the once significant correlation between judges’ molar ratings and the IATFULL measure drops from \( r = .34, p < .05 \), to \( r = .27, p < .10 \). In our judgment, the elimination of this outlier bumped this particular correlation a bit over the \( p < .05 \) barrier, but we do not view this change as monumental. But more important, Table 1 reveals that if Blanton et al. (2009) had reported on the predictive utility of the IAT for the other indexes of bias reported by McConnell and Leibold (2001), the vast majority of these correlations remained strong and significant even after we eliminated the outlier. Indeed, as people revealed relatively more negativity toward Blacks on the IAT, they had relatively more negative explicit attitudes toward Blacks; the White experimenters perceived relatively more positive interactions with the participants than did the Black experimenters; and the judges viewed that the participants showed less speaking time, less smiling, more speech errors, and fewer extemporaneous social comments toward the Black experimenter than the White experimenter. Each and every one of these findings is directly at odds with the Blanton et al. conclusion that the IAT did not predict behavioral bias in interracial interactions once the outlier was dismissed. Thus, out of the seven significant correlations between the IAT and biased behavior (as assessed both by the experimenters and by the judges) reported in McConnell and Leibold, five remained significant at conventional levels following the elimination of the outlier. Moreover, three of the correlations were larger following the elimination of the outlier too. Thus, we see considerable evidence that the IAT continues to predict indicators of biased intergroup behavior even with the elimination of this outlier.

It is curious why Blanton et al. (2009) only focus on one outcome measure, especially when it was one that produced the most relatively congruent results for their claims. Blanton et al. argued that they used the judges’ ratings (instead of the experimenters’ ratings) because “these are not filtered by the experience of actually participating in the interactions” (p. XX). This is a peculiar assertion for a number of reasons. First, Blanton et al. do not report the other ratings provided by the same judges (e.g., bias in speaking time, smiling) that, based on the analyses reported in Table 1, do not support their conclusions. Second, work in the literature (e.g., Fazio et al., 1995) has relied on experimenters’ experiences in interracial interactions as the key behavioral datum. In fact, who is better qualified to know whether an interaction was stilted than those involved? It is unclear how any bias could be introduced by the experimenters because they were completely unaware of each participant’s prejudice level. Thus, the claim that their reports are inadmissible is hard to support. Indeed, one of the strengths of McConnell and Leibold (2001) was that we assessed both experimenters’ experiences and independent judges’ perceptions, and the strong correspondence between the experimenters’ and judges’ molar ratings (\( r = .41, p < .01 \)) provides strong evidence for their validity.

In sum, the pattern of results observed, even with the elimination of the outlier, do not support the Blanton et al. (2009) conclusion that the correlation between the IAT and racially biased behavior was dependent on outliers. It is unclear why Blanton et al. do not note that these other correlations remained significant even when they (questionably) eliminated the outlier from their analyses. Further, it is also puzzling why Blanton et al. focus on a different IAT index than the one reported in the original article. It is understandable how a reader might conclude that such differences and omissions could reflect a selective presentation of the results, especially given that the (still) significant correlations that Blanton et al. fail to discuss are showcased in the original McConnell and Leibold (2001) article and that these data were in the data set provided to them.

Specious Concerns About Interjudge Reliability

Another point emphasized in the Blanton et al. (2009) critique is that interjudge reliability was problematic for the two judges who coded the videotapes. Despite suggestions to the contrary, the judges were extensively trained, and they spent 4 months painstakingly coding the videos. More important, in all cases, these judges showed significant correspondence across the 16 different dimensions they assessed for each of the two interactions between experimenters and the 41 participants. Although any study could benefit from greater interrater reliability, any dissimilarities between judges’ ratings would only increase variability in their assessments, making it more (not less) difficult to observe the significant relations between the IAT and biased behaviors found in the study.

Moreover, if the judges were truly off base, their overall perceptions (i.e., judges’ molar ratings) would have been discrepant with the experimenters’ own reports, but there was reliable correspondence between them (\( r = .41, p < .01 \)). Instead, the judges always showed significant agreement with each other, and in

---

1 This exercise was the first time we had conducted analyses on these data in over 8 years. Our reanalysis replicated each and every finding reported in McConnell and Leibold (2001), with one exception. Specifically, the significant correlation between the IAT and bias in extemporaneous social comments shifted from .32 to .31 (both \( p < .05 \)). It is unclear what is responsible for this difference (e.g., rounding differences resulting from files being imported into sundry spreadsheet and statistical program packages, or differences in the software available today versus in 1999 when the original analyses were conducted).
concert their combined ratings significantly correlated with the independent reports of the experimenters. Thus, the general claim by Blanton et al. (2009) that the judges’ ratings were unreliable is unsubstantiated. Consequently, their attempt to explore patterns of correlations for a particular judge seems unwarranted and post hoc.

Inappropriate Claims About a Disconnect Between IAT Scores and Behavioral Data

In their critique, Blanton et al. (2009) assert that there is a disconnect between participants’ IAT scores (argued to be anti-Black) and behaviors (argued to be pro-Black). Specifically, Blanton et al. claim that because 37 of the 41 participants in the McConnell and Leibold (2001) study had positive IAT scores, these participants were biased against Blacks. However, viewing IAT = 0 as some sort of meaningful criterion is at odds with the understood relative nature of the IAT (see Greenwald, McGhee, & Schwartz, 1998; McConnell, Rydell, Strain, & Mackie, 2008; Nosek, Greenwald, & Banaji, 2006). What matters with the IAT is not an absolute score, but rather how one’s relative standing on the measure predicts other outcomes, which themselves are relative. Also, Blanton et al. argue that participants in McConnell and Leibold were friendlier toward Blacks than toward Whites (based on examining judges’ ratings for participants’ interactions with the Black and White experimenters separately). Yet, such a characterization is misleading. For example, perhaps our Black experimenters were actually nicer people than our White experimenters and thus elicited more positive interactions. There is no way to equate, in an absolute sense, different people with different interactions, especially with subjective rating scales (see Biernat & Manis, 1994). The point is not whether one set of ratings (e.g., those for Black experimenters) is greater than another set of ratings (e.g., those for White experimenters). Instead, it is the relative difference between these sets of ratings (i.e., how friendly were participants toward members of one race relative to another) that is predicted by the IAT, which itself is a relativistic measure.

So What Is the Bottom Line of Blanton et al. Anyway?

In our last few sections, we address broader issues related to the Blanton et al. (2009) critique. First and foremost, the point of their criticism is lost upon us. They try to argue that implicit racial attitudes do not predict interracial behaviors. In their quest to establish this claim, they attack McConnell and Leibold (2001). And in the above sections, we explain why these criticisms are unfounded. However, even if one could successfully impugn McConnell and Leibold, other published studies show that implicit racial attitudes predict interracial behavior (e.g., Dovidio, Kawakami, & Gaertner, 2002; Fazio et al., 1995). That is, a number of diverse studies using a variety of measures to assess implicit racial attitudes come to the same bottom line (i.e., automatically activated attitudes predict more spontaneous forms of behavior). Thus, several studies support the same conclusions that McConnell and Leibold advanced, yet these well-established studies are not challenged (or even discussed) by Blanton et al.

Perhaps the intended message of the Blanton et al. (2009) critique is that a specific method (i.e., the IAT) is not a suitable candidate for predicting behavior, whereas other techniques (e.g., Dovidio et al., 1997; Fazio et al., 1995) are acceptable. If this is the case, however, then other findings (also not discussed by Blanton et al.) in the literature demonstrating that the IAT predicts behavior are problematic for such a position. For instance, work such as Rydell and McConnell (2006) built implicit attitudes toward novel individuals and showed that implicit attitudes (as assessed by the IAT) but not explicit attitudes (as assessed by feeling thermometers and semantic differentials, just like McConnell & Leibold, 2001) predict spontaneous behaviors (e.g., greater seating distance from someone for whom a relatively negative bias was exhibited on an IAT). As a result, even if one questions McConnell and Leibold (2001), there is parallel evidence that the IAT itself can uniquely predict spontaneous behaviors. Thus, the Blanton et al. critique cannot successfully challenge the predictive utility of implicit attitudes in general or of the IAT in particular.

A Clarification on What We Mean by Behavior

One thing we must emphasize is that we have never claimed in our work (e.g., McConnell & Leibold, 2001; Rydell & McConnell, 2006; Rydell, McConnell, & Mackie, 2008) that implicit attitudes (as assessed by the IAT or by any other measure) should predict all forms of behavior. In fact, we have underscored in past work that nonconscious associations (measured by a variety of approaches, including the IAT) are more likely to predict spontaneous behaviors than deliberative actions (see McConnell et al., 2008; Rydell & McConnell, 2006). Blanton et al. (2009) contend that the application of implicit attitudes (e.g., nonconscious racial prejudice) is problematic for the legal system. In our own work (including McConnell & Leibold, 2001), we have never claimed that implicit attitudes will predict effortful, deliberative courtroom judgments. In fact, that is why we have focused on measures of subtle behavior, such as nonverbal displays (e.g., McConnell & Leibold, 2001) and seating distance (e.g., Rydell & McConnell, 2006). Similar claims have been advanced by others (most notably Dovidio et al., 2002). That being said, we do believe that nonconscious biases (including implicit prejudice) can impact behavior and judgments in meaningful ways because like most dual-process systems (Chaiken & Trope, 1999), nonconscious processes typically have an automatic and unavoidable impact on perception and action that, later, more effortful processes may moderate. However, we anticipate that implicit prejudice would have a greater impact on interracial behaviors that are more spontaneous, less verbal, and less deliberative (see McConnell et al., 2008; Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006). However, the correspondence between implicit prejudice and relatively more deliberative interracial judgments and behaviors should increase considerably when one has reduced cognitive resources, an incorrect understanding of their own biases, or less motivation to render effortful decisions.

Implications for the Law

We contend that an appreciation of implicit attitudes and their implications for courtroom behavior is very important in understanding the outcomes of the legal system. Throughout a trial, jurors view evidence and testimony through the lens of their attitudes and beliefs. Processes such as voir dire, for example, are critical for identifying and removing prejudiced jurors. Yet during voir dire, jurors are asked to report their attitudes explicitly, which
can only reveal a subset of their feelings related to their behaviors and decisions (Dovidio et al., 2002; Rydell & McConnell, 2006; Wilson, 2002). Further, at this early stage in the legal process, jurors are prohibited from making any decisions and are often discouraged from expressing a bias, thus increasing the spontaneity and subtlety of their reactions to a particular case or party. Under idealized conditions, jury decisions should be quite effortful and thus less subject to the influences of nonconscious processes. Unfortunately, the courtroom is far from a perfect environment. For example, to the extent that it is a cognitively demanding place that taxes jurors’ cognitive resources, implicit attitudes are more likely to predict guilt-related judgments (Strain & McConnell, 2008). Moreover, instructions to suppress thoughts (e.g., judge’s instructions to ignore information or put attitudes aside) can, ironically, make those thoughts hyperaccessible because of rebound effects (Wegener, 1989), increasing the application of one’s stereotypes and prejudices (e.g., Macrae, Bodenhausen, Milne, & Jetten, 1994). And more generally, jurors who are relatively un-motivated may just not engage in effortful judgment and decision making in the first place, such as in a recent mistrial in Sydney, Australia, where jurors hearing a case involving a possible life sentence played Sudoku instead of paying attention to the proceedings because they were bored (ABC News, 2008).

These insights are especially important because decades after the Civil Rights Act of 1964 and the Age Discrimination in Employment Act of 1967, sex, race, religion, and other employment discrimination claims are still being filed, with age and family responsibility discrimination cases, in particular, being filed in ever-increasing numbers (Hodge, 2004; Still, 2006; United States Equal Opportunity Employment Commission, 2007). In fact, recently the Supreme Court shifted the burden of proof onto employers in age discrimination cases to show that their actions toward an employee were nondiscriminatory and taken for reasons other than age (Meacham v. Knolls Atomic Power Lab, 2008). A burden shift such as this is a key reflection of the level of discrimination that the highest court sees in the United States today. Yet, these cases are notoriously difficult to prove because the discrimination occurs over time and often in subtle ways when employers have discretion in subjective ratings of employee performance (Meacham v. Knolls Atomic Power Lab, 2008). Employers may pay women less or lay off older workers, but they probably do not often explicitly rationalize these decisions as a belief that “it’s a man’s job,” “women will not put in the time and effort it takes,” or “older employees are ineffective” (Fiske, 1998; Walker v. Fred Neshit Distrib. Co., 2004). Thus, subtle and nonconscious biases can have important implications in the workplace as well as in the courts that hear worker discrimination cases. Although we would not contend that implicit attitudes and prejudices are the sole determinant of biased behaviors in the workplace and courtroom, we do believe they play a role, and in some circumstances a very consequential role, in determining critical outcomes. Because bias in the courtroom is especially insidious given the basic precepts of jurisprudence in our society, minimizing its effects among empanelled jurors is especially important. But sadly, it is well documented that jurors have difficulties in ignoring specific features of cases or in discounting their own biases (Sunstein, Hastie, Payne, Schkade, & Viscusi, 2003), especially in situations where their bias is nonconscious and thus jurors would not have a starting point from which to consider how to correct any biases they possess (Wegener & Petty, 1997). In sum, it is clear that the relations between implicit prejudice and discriminatory behavior are nontrivial, and real implications exist for the legal system.

Implications for Future Progress in Implicit Attitudes and Prejudice

We were pleased that many years after the publication of McConnell and Leibold (2001), additional analyses (by us and by others) bore out consistent findings reaffirming that implicit attitudes, as assessed by the IAT, reliably predicts intergroup behaviors in ways that are detectable by both those involved in the interactions and by third-party observers alike. These findings are consistent with other intergroup relations studies (e.g., Dovidio et al., 2002; Fazio et al., 1995) and with others where implicit attitudes were built in the lab and assessed with the IAT (e.g., Rydell & McConnell, 2006). The current work demonstrates that despite the concerns expressed by Blanton et al. (2009), the conclusions offered by McConnell and Leibold are quite robust even when under considerable scrutiny.

That said, we support those who challenge existent work, especially in the face of a relatively small literature on IAT–behavior relations. It was in this spirit that we freely provided all of our data upon request to Blanton et al. (2009) because we believe that doing so is important for providing a transparent account of one’s work. Indeed, scientists have an obligation to conduct research with the utmost of integrity and respect for the scientific enterprise and to subject themselves to its scrutiny. Further, it is clear that Blanton et al. experienced difficulty in securing some relevant data sets, and we can understand how frustration can result from this. With this in mind, we offer some observations that might prove useful in encouraging the free exchange of ideas and data in the future.

We believe that trust is the cornerstone of scientific openness and that trust is a two-way street. First, those who wish to critique findings must be able to trust that data will be available to them in a timely and unfettered fashion. Not only is this consistent with American Psychological Association guidelines and procedures but, more importantly, by the recognition that data (and not reports of data) are the basic building blocks of our field. The passage of time may lead to reinterpretations of findings, but in the end, data are the foundation of researchers’ theories and future work. However, scientists must know that fellow researchers will treat them fairly and respectfully. For example, when skeptics can articulate a priori hypotheses regarding their interest in others’ data, everyone benefits. Those skeptics will have a guiding framework for their inquiries, researchers whose data are reviewed will feel that the likelihood of a “fishing” or “cherry picking” expedition is minimal, and both parties can be maximally helpful to each other in identifying important facets of the work (e.g., auxiliary data, unreported observations) that can provide a more complete picture of the work conducted. And more important, strong conclusions cannot be drawn simply from reexaminations of data. Although such inquiries may identify different interpretations for existent data, it is then incumbent upon researchers to conduct primary research to establish such claims conclusively (e.g., experimentally manipulate the hypothesized confounds). When considering the critique offered by Blanton et al. (2009), perhaps those who are older might reveal stronger IAT–behavior correspondence. But
rather than viewing this as a problem for the IAT, this insight can be parlayed into testable hypotheses with a strong theoretical foundation. For instance, to the extent that automatically activated attitudes are more likely to guide behavior when cognitive expenditures are reduced (e.g., Fazio & Towles-Schwen, 1999), perhaps those who are older are less likely to inhibit their deliberately chosen behaviors (Hasher & Zacks, 1988) and thus respond more in line with their implicit prejudices in intergroup interactions. Maybe such speculations are incorrect, but they are certainly testable. Importantly, such ideas can lead to new research that pushes theory on implicit prejudice forward in compelling and sophisticated ways.

Without trust, progress in any domain of science will be stifled. And to the extent that particular researchers on either side of an intellectual divide repeatedly arouse mistrust and suspicion, whether real or perceived, walls to an open and transparent scientific enterprise will be erected and reinforced. We hope that, in addition to defending the credibility of our past work, that the current episode illustrates the benefits of an open scientific exchange of ideas that can push researchers’ understanding of important psychological phenomena and principles forward.

References


Received July 4, 2008
Revision received October 1, 2008
Accepted October 23, 2008

/589

WEAK CRITICISMS AND SELECTIVE EVIDENCE
Strong Rebuttal for Weak Criticisms: Reply to Blanton et al. (2009)

Jonathan C. Ziegert  
Drexel University

Paul J. Hanges  
University of Maryland

We respond to H. Blanton et al.’s (2009) critique of J. C. Ziegert & P. J. Hanges (2005) and illustrate that their criticisms regarding our study are based on very weak evidence. In this rebuttal, we highlight the appropriateness of our manipulation as well as present reanalyses that illustrate the predictive validity of the Implicit Association Test. In particular, we illustrate that the Implicit Association Test is related to discriminatory behavior (a) after one eliminates potential outliers using appropriate multivariate methods, (b) across conditions after one controls for the manipulation, (c) when one just examines Black applicants, and (d) after one controls for explicit attitudes. Thus, in this rebuttal, we not only reaffirm the original findings of J. C. Ziegert and P. J. Hanges (2005) but we extend them to illustrate the broader contexts in which the Implicit Association Test is related to discriminatory behavior.

Keywords: stereotypes, discrimination, Implicit Association Test

Since its development, the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) has generated a substantial amount of attention; both positive (e.g., Banaji, Nosek, & Greenwald, 2004; Gladwell, 2005; Greenwald, Nosek, & Sriram, 2006; Greenwald, Poehlman, Uhlmann, & Banaji, in press) and negative (e.g., Arkes & Tetlock, 2004; Blanton & Jaccard, 2006). Not only have social scientists been debating the merits of the IAT, but these arguments have even occurred among legal scholars (e.g., Bagens, 2007; Greenwald & Krieger, 2006; Kang & Banaji, 2006; Mitchell & Tetlock, 2006). In their article, Blanton et al. (2009) continue this debate by reanalyzing the data of two influential studies (McConnell & Leibold, 2001; Ziegert & Hanges, 2005) that have been used as evidence that “antidiscrimination law must be remade to address implicit biases” (p. 569). As a result of their reanalysis, Blanton et al. conclude that there is little support for the original findings of these studies regarding the predictive validity of the IAT.

In this article, we focus on Blanton et al.’s (2009) criticisms of the Ziegert and Hanges (2005) study. Specifically, Blanton et al. criticize both methodological aspects (e.g., climate for bias manipulation, inequivalence of applicant resumes within race, and the scoring of the IAT) as well as the analytical results (e.g., the presence of apparent outliers in the data, decomposing the effects, and the utility of the IAT as a diagnostic tool). As a result of these critiques, Blanton et al. make very strong statements regarding our study. After a careful review of their article, we believe that each of their points can be easily countered. To put it a different way, we believe that these strong criticisms by Blanton et al. are based on questionable arguments and very weak, and often contradictory, evidence.

These critiques of our study by Blanton et al. (2009) are not a surprise given the similarities of them to ones that Mitchell and Tetlock (2006) published in a legal journal. The difference now is that the editor of the Journal of Applied Psychology has given us an opportunity to correct the misperceptions and inaccuracies caused by these articles (Blanton et al., 2009; Mitchell & Tetlock, 2006). We appreciate this opportunity to present a rebuttal to these articles in order to illustrate the weakness of their criticisms.

In this article, we present counterarguments to the points raised by Blanton et al. (2009). We illustrate that their criticisms are unfounded and their conclusions are ultimately not supported. Further, we show that the relationship between the IAT and discriminatory behavior reported in our 2005 article (Ziegert & Hanges, 2005) is actually quite robust to outliers, condition effects, and explicit attitudes. Indeed, we believe that our reanalysis, prompted by Blanton et al.’s critique, actually provides stronger evidence for the IAT’s predictive validity than the original analyses reported in Ziegert and Hanges (2005).

Before beginning with our specific arguments, we note that this article is not written to provide a general defense of the IAT. Other scholars have already provided such a defense, and a recent meta-analysis provides more comprehensive support for the predictive validity of the IAT across a variety of dependent variables (Greenwald et al., in press). Instead, our purpose in this article is to respond to Blanton et al.’s (2009) specific critiques of our study to prevent misinterpretation of the Ziegert and Hanges (2005) study in the future.

Description of Study

As noted by Blanton et al. (2009), our 2005 study was a constructive replication of Brief, Dietz, Cohen, Pugh, and Vaslow (2000). That is, we attempted to replicate and extend Brief et al.’s work in our study. Specifically, we extended Brief et al. by using the IAT to measure implicit racial attitudes as well as to use hierarchical linear modeling (HLM) to create a measure of racial bias. By using HLM, we not only created a new measure of bias but also incorporated the nested structure of our data into our analyses.
More specifically, we created a dummy-coded applicant race variable that was used as a level 1 (i.e., within-individual) predictor of participants’ ratings of the degree to which the applicants were exceptional referrals. The Level 1 slope for the race variable was our racial discrimination measure (a negative slope was indicative of bias against Black applicants, a zero slope was indicative of no racial bias, and a positive slope was indicative of a pro-Black applicant bias). We then used between-persons, or Level 2, variables (e.g., IAT, condition, etc.) to predict the magnitude of the Level 1 slope. We are unsure why Blanton et al. (2009) report that they could not exactly replicate the p values from our original interaction findings. Indeed, when we performed our HLM analyses again based on this approach from the original study, we were able to replicate and obtain the same statistical values and significant interaction (i.e., \( p < .05 \)) as reported in our 2005 article (Ziegert & Hanges, 2005).

**Methodological Considerations**

**Artificiality of the Manipulation**

One of Blanton et al.’s (2009) primary critiques of our study surrounds the manipulation that we used. Again, as our 2005 article (Ziegert & Hanges, 2005) was a constructive replication, we used the same in-basket exercise and manipulations that Brief et al. (2000) used. In their critique, Blanton et al. contend that this manipulation was artificial given that the “heavy-handedness of the request created demand characteristics” (p. 578).

Regarding this critique of the inappropriateness of the manipulation, as discussed in Ziegert and Hanges (2005), norms and climate have long been shown to have a strong impact on behavior in organizations (Schneider, 1972). Further, leaders (such as presidents of companies) can and do create social climates. Indeed, the classic works by Lewin, Lippitt, and White (1939) as well as Litwin and Stringer (1968) illustrated how leaders create social climates that impact the behavior of individuals within these settings. Linking the climate of an organization to potential discriminatory practices, Gelfand, Nishii, Raver, and Schneider (2005) noted that “organizational cultures are not neutral with regard to what types of employees fit and which employee behaviors are valued” (p. 96). Discriminatory climates can be manifest in a number of ways including “[Human Resources] practices,” “directly through derogatory language,” or “business justifications for discriminating” (Gelfand et al., 2005, p. 96).

Following this influence of leaders in creating discriminatory climates, we believe similarities exist between our manipulation and organizational settings. One example is Texaco Inc. and the documents that senior executives destroyed in August 1994 even though these documents were part of discovery in a federal discrimination lawsuit against the company. The presence of a discriminatory climate in Texaco at this time is further revealed by a tape recording of these senior executives joking about the company’s minorities and using racial epithets as they destroyed these documents (Eichenwald, 1996). While we may never know exactly what these documents said, the fact that the senior executives felt that they were sufficiently damaging that they needed to be destroyed strongly suggests that the way that we manipulated discriminatory climate in our study may not be so unlikely. We do not mean to imply that such memos are typical in the real world, just that memos documenting discriminatory climates can and have happened.

Indeed, discriminatory climates can be created by all sorts of creative ways. Employees can be “instructed either explicitly or implicitly by their organizational superiors to do the wrong thing” (Brief, 1998, p. 145) as evidenced in further examples at Shoney’s and Avon in terms of advocating racial matching (Brief et al., 2000). Further, a quick review of the Equal Employment Opportunity Commission’s Website (http://www.eeoc.gov/litigation/settlements/index.html) reveals numerous recent settlement cases in which organizations have created discriminatory climates. For example, in the 2006 *EEOC v. Morton Buildings* settlement case, a Western Regional manager directly told his subordinate not to hire any females for a commercial real estate construction and sales position. As another example, in the *EEOC v. Nine West Footwear* settlement case, women employed at Nine West’s Esprit division headquarters were exposed to a hostile work environment based on sex and national origin (Hispanic) discrimination. Not only did the top manager proposition female employees for sex and groped women’s bodies, two men also made derogatory remarks about the intelligence and sexual prowess of Hispanic women. The Human Resource department discouraged any complaints by telling the women to “get used to” and to “work around” the harassment. As these settlement cases illustrate, our manipulation of a discriminatory climate in the 2005 study (Ziegert & Hanges, 2005) was actually fairly subtle compared to how such climates can actually be created in the real world.

Blanton et al.’s (2009) critique of the “artificiality” of our manipulation raises questions about the external validity of our laboratory design. The comparison of laboratory and field studies has a long history in social sciences (e.g., Hovland, 1959; Mook, 1983; Locke, 1986). Blanton et al. note that our manipulation was delivered in a “hypothetical, role-playing setting” and therefore they question the external validity of the results to “everyday” relationships (p. 578). Blanton et al. appear to be arguing that all aspects of a situation are needed in an experiment for it to provide generalizable results. However, this extreme viewpoint is not consistent with the current view in the literature. As discussed by Waller and Zimbelman (2003), mundane realism (i.e., the extent to which elements of the real world are incorporated in an experiment) is a continuum. Our 2005 experiment (Ziegert & Hanges, 2005) fell in the middle of this continuum and this allows statements of internal and external validity (Waller & Zimbelman, 2003).

Even if our manipulation (i.e., a letter from the president of the company) does not happen in everyday situations, a key aspect of laboratory research is to illustrate the “generalization of theoretical relations among conceptual independent and dependent variables, not the specific instantiations of them” (Anderson, Lindsay, & Bushman, 1999, p. 4). As such, while the specific manipulation we utilized (i.e., a letter from the president) may not exist often in everyday settings, the fact that leaders can and do create specific social climates that may be biased (through a variety of means) remains (Gelfand et al., 2005). Thus, we contend that the theoretical relations that implicit stereotypes are related to discriminatory behaviors in a climate for racial bias demonstrated in our study does generalize to the real world.

A review of some of the previous work by several of the authors of the Blanton et al. (2009) article reveals that they understand this issue concerning the generalizability of theoretical relationships. For example, Mitchell, Tetlock, Mellers, and Ordoñez (1993) conducted a study in which participants read descriptions of hy-
pothetical societies called Alpha in which hard work was rewarded 10%, 50%, or 90% of the time depending on the condition. Participants were instructed that the government was evaluating ways to restructure this society and were then asked to imagine themselves as a member of this society before answering several questions. While this manipulation by Mitchell et al. (1993) was obviously artificial in many respects, we agree with the conclusion of these authors that such a manipulation provides meaningful information to the real world because it allows for clear understanding of underlying theoretical relationships of the constructs of interest, just as our study allowed for better understanding of how implicit attitudes may be related to discrimination in different social climates.

Results Across the Manipulation

Even if one subscribes to Blanton et al.’s (2009) critique of our manipulation and discounts our arguments above, one way to minimize any demand characteristics of our manipulation would be to examine the relationship of the IAT to discriminatory behaviors after statistically removing the experimental manipulation. We did not perform this test in our original study because we were focused on the moderating effect of social climates on the IAT–discriminatory behavior relationship.

The IAT showed a significant relationship to the racial bias slope ($\gamma = -0.13$, $t_{[95]} = -3.00$, $p < .01$). This significant relationship corresponds to a correlation of $-26$ when we estimated the racial bias measure by subtracting the mean ratings for the three White applicants from the mean rating for the three Black applicants. Further, the IAT–bias slope relationship was still significant after we controlled for experimental condition ($\gamma = -0.11$, $t_{[94]} = -2.72, p < .01$). This significant relationship corresponds to a partial correlation of $-0.23$.

In summary, this new analysis speaks directly to Blanton et al.’s (2009) claims regarding the manipulation in our 2005 study (Ziegert & Hanges, 2005) and the lack of IAT predictive validity. The IAT had a significant zero-order relationship with discriminatory behavior as well as a significant partial relationship after we controlled for experimental condition. Contrary to Blanton et al.’s claims, these new analyses demonstrate that the IAT is related to discriminatory behavior independent of experimental condition.

Applicant Equivalence

Blanton et al. (2009) indicate that there were within-race differences in the resumes given to participants. We relied on the prior work by Brief et al. (2000) that demonstrated that there were no rating differences among these six candidates when race information was not included in the six resumes. Thus, we did not reevaluate this pilot work because, as indicated above, we were replicating the Brief et al. study.

While in our data there were differences between one Black applicant compared to the other two Black applicants as well as one White applicant compared to the other two White applicants, the magnitude of these within-race differences were relatively equal. It is interesting to note that the higher rated Black applicant and the higher rated White applicant were the first two applicants who participants reviewed. This suggests that there might be an order effect in our experiment. We relied on the Brief et al. (2000) pilot study with regard to the equivalence of these resumes when race information was absent.

However, to be consistent with Blanton et al. (2009), we explored the within-race differences among the resumes in our data. In addition, we also provide a 95% confidence interval for these differences. Consistent with Blanton et al., the first Black applicant received significantly higher ratings than the other two Black applicants. The difference between the first and second Black applicants was 0.65 (95% confidence interval of .46 to .84) and the difference between the first and third Black applicants was 0.66 (95% confidence interval of .47 to .85). The difference between the first and second White applicants was 0.43 (95% confidence interval of .26 to .60) and the difference between the first and third White applicants was 0.41 (95% confidence interval of .23 to .60). The overlap in the confidence intervals suggests that within-race magnitude differences are approximately similar in magnitude. In other words, the variability in resumes within race was comparable. Given that (a) prior pretesting of these resumes demonstrated that they are equal when race information was not present, (b) the apparent order effect in the ratings of the resumes, and (c) similar differences among the resumes within race, the criticism of lack of applicant equivalence is reduced.

IAT Scoring

Blanton et al. (2009) correctly point out that our scoring of the IAT differs from the improved scoring algorithm described by Greenwald, Nosek, and Banaji (2003). The reason for this discrepancy is that our manuscript was originally submitted in 2002, before the revised scoring algorithm was published by Greenwald et al. in 2003. Our scoring procedure combined both latency and error results into a $z$ score composite. As we noted in the 2005 article (Ziegert & Hanges, 2005), we followed classic psychometric procedures to justify this scale by first conducting an exploratory factor analysis that demonstrated that both latency and error information load on one factor. We transformed these two measures into their $z$ scores before creating the IAT composite score because the latency and error measures were scaled in different units.

It is very interesting to note that while the traditional scoring algorithm by Greenwald et al. (1998) focused exclusively on latencies, the revised and improved algorithm by Greenwald et al. (2003) incorporates error penalties into the IAT scoring. Thus, our unique scoring of the IAT is consistent with the improved algorithm by Greenwald et al. (2003). Both procedures use latency and error information when scoring the IAT.$^1$

Analytical Considerations

Influence of Outliers

Blanton et al. (2009) claim that our results were “not robust to the influence of outliers” (p. 578). They base this claim on their identification and removal of three outliers through their visual

---

$^1$ The raw IAT data were lost due to a computer hard drive crash. Due to the size of this file, it was stored on the original nonnetworked laboratory computer used to collect the IAT data. Fortunately, the composite IAT data and all other raw data from the 2005 study were transferred to a networked computer and were available to G. Mitchell based on his request for our data.
inspection of the scatterplots. However, visual inspection to identify outliers is a notoriously flawed approach for detecting outliers. Indeed, Inglewicz and Hoaglin (1993) noted that “visual inspection alone is not a reliable way to identify potential outliers” (p. 9). Further, Cohen, Cohen, West, and Aiken (2003) stated that researchers should “supplement such visual inspection with the use of specialized statistics known as regression diagnostics which can greatly aid in the detection of outliers” (p. 394).

Instead of this visual approach, we conducted outlier analyses using a multiple regression framework advocated by a number of scholars (e.g., Cohen et al., 2003; Rousseeuw & Leroy, 1987). This multiple regression approach provides multiple diagnostics for identifying outliers. These specific diagnostics discussed by Cohen et al. (2003) include centered leverage, externally studentized residuals, DFFITS, and Cook’s D. The centered leverage diagnostic highlights extreme values on the independent variables. The externally studentized residuals diagnostic highlights deviations of the actual dependent variable value for each observation to the predicted value along the regression line. The DFFITS diagnostic examines influence or the change in the predicted outcome if the particular case is not included in the estimate. The Cook’s D diagnostic also examines influence and is indexed by the change in the set of regression coefficients if the case is omitted. We utilized these four statistical diagnostics in our reanalysis to identify any possible outliers in our 2005 data set (Ziegert & Hanges, 2005).

According to Cohen et al. (2003), the proposed cutoff for the centered leverage values is $3kn$ (where $k$ is the number of predictors in the equation and $n$ is the sample size) for small samples. As there are three independent variables in the relationship of interest (IAT, condition, and the IAT × Condition interaction), the centered leverage cutoff value is .0927. An examination of the data identifies four observations with higher centered leverage values than this cutoff. When these observations were removed, however, the interaction between corporate climate and the IAT was still statistically significant ($\gamma = -2.15, p < .05$).

In terms of the externally studentized residuals, Cohen et al. (2003) recommends a cutoff of $\pm t_\alpha$ for a small sample. An examination of the data reveals that four observations exceeded this cutoff. With these observations removed, the interaction between corporate climate and the IAT was still statistically significant ($\gamma = -2.17, p < .05$).

According to Cohen et al. (2003), the proposed cutoff for the DFFITS diagnostic is $\pm 1.0$ for a small sample. An examination of the data reveals that no DFFITS values exceed $\pm 1.0$, which suggest that no outliers exist based on this influence statistic.

Finally, the proposed cutoff for Cook’s D according to Cohen et al. (2003) is $1.0$. Again, an examination of the data revealed that no values exceed this cutoff, which also suggests that no outliers exist in the relationship of interest based on this influence diagnostic.

Overall, when rigorous multivariate approaches for detecting outliers were used (Cohen et al., 2003), the interaction between the IAT and condition on discrimination remained statistically significant across four different outlier detection approaches. Thus, we conclude that the interaction between the IAT and corporate climate condition is a significant predictor of discrimination and remains so even when removing potential outliers. 3

In addition to the visual outlier approach, Blanton et al. (2009) use the Theil-Sen regression method, which they refer to as “an outlier-resistant method for comparing regression lines in two groups” (p. 578). However, there are two problems with using this approach. First, according to Wilcox (2005), the Theil-Sen approach is appropriate in the case of simple regression (i.e., one predictor). The interaction hypothesis of interest has three predictors (i.e., experimental condition, IAT measure, and interaction term). Second, in contrast to Blanton et al.’s claim that the Theil-Sen regression method is an outlier-resistant method, simulations have shown that this approach can be heavily influenced by outliers (Wilcox, 2004; Wilcox & Keselman, 2004). Indeed, Wilcox and Keselman (2004) noted that “the Theil-Sen estimator can be adversely affected by two or more properly placed outliers” (p. 363). Given this susceptibility of the Theil-Sen technique to outliers and Blanton et al.’s own assertion that outliers exist in the data, any findings from their analysis should be considered tentative at best. 4

Decomposing the Effects

The second reanalysis by Blanton et al. (2009) decomposes the effects on White versus Black applicants. That is, they consider the

---

1. According to Cohen et al. (2003), the proposed cutoff for the centered leverage values is $3kn$ (where $k$ is the number of predictors in the equation and $n$ is the sample size) for small samples. As there are three independent variables in the relationship of interest (IAT, condition, and the IAT × Condition interaction), the centered leverage cutoff value is .0927. An examination of the data identifies four observations with higher centered leverage values than this cutoff. When these observations were removed, however, the interaction between corporate climate and the IAT was still statistically significant ($\gamma = -2.15, p < .05$).

2. According to Cohen et al. (2003), the proposed cutoff for the centered leverage values is $3kn$ (where $k$ is the number of predictors in the equation and $n$ is the sample size) for small samples. As there are three independent variables in the relationship of interest (IAT, condition, and the IAT × Condition interaction), the centered leverage cutoff value is .0927. An examination of the data reveals that no DFFITS values exceed $\pm 0.406$, which suggest that no outliers exist based on this influence statistic.

3. Overall, when rigorous multivariate approaches for detecting outliers were used (Cohen et al., 2003), the interaction between the IAT and condition on discrimination remained statistically significant across four different outlier detection approaches. Thus, we conclude that the interaction between the IAT and corporate climate condition is a significant predictor of discrimination and remains so even when removing potential outliers.

4. In addition, Blanton et al.’s (2009) very own reanalysis of the simple slopes within each condition actually reaffirms our findings. That is, their analysis illustrates that within the climate for racial bias, the IAT is related to discriminatory behavior (absolute $t_{pb} = .38, p < .05$), which is what we would have predicted. Their analysis used the percentage bend correlation, which is a robust technique that is resistant to outliers (Wilcox, 2005). Thus, Blanton et al. rightly note that the IAT is related to discriminatory behavior within the climate for racial bias condition when an analytic technique that is not affected by outliers is used.
impact of the IAT separately on Black and White applicants. This analysis is a curious choice as it moves further away from a main definition of discrimination. In the Ziegert and Hanges (2005) study, we specifically adopted the disparate treatment definition of discrimination that occurs when different standards are applied to different groups through the use of our HLM analysis (with the within-individual analysis of the slopes). By focusing on ratings within each group, Blanton et al., in this reanalysis, move further away from the legal definition of discrimination, as Mitchell and Tetlock (2006) stated “under current law, disparate treatment claims can encompass acts of discrimination” (p. 1038). Thus, we believe that focusing on only ratings of Black or White ratings in isolation results in a loss of important information as to whether disparate treatment exists (as one only can evaluate the ratings of one group).

Despite the reduced utility of identifying discrimination with this approach, Blanton et al.’s (2009) reanalysis confirms our findings. As expected, the IAT was related to the ratings of Black applicants in the climate for racial bias condition (two at $p < .05$ and the third was $p < .06$ for a two-tailed test; all three would have $p < .05$ for a one-tailed directional test based on the directional hypothesis). Indeed, as Blanton et al. note “this finding supports Ziegert and Hanges’ (2005) predictions” (p. 578).

While Blanton et al. (2009) examine the relation of the IAT to each individual applicant in isolation, a more appropriate approach would have been to look at the average ratings of the Black applicants. Examining the average rating of the Black applicants is a more parsimonious method that allows for broader conclusions in terms of ratings toward an applicant class (rather than individual applicants). With this approach, the IAT is also related to the average rating of the Black applicants in the climate for racial bias condition ($r = -.22, t[44] = -2.60, p < .05$). Further, these results hold looking at the relationship of the IAT to the ratings of Black applicants across condition as well. That is, after one controls for condition, the IAT is related to the average rating of Black applicants ($r = -.12, t[94] = -2.72, p < .01$). These results reaffirm Blanton et al.’s reanalysis, which lends further support to the Ziegert and Hanges (2005) original finding illustrating the relationship between the IAT and discriminatory behavior.

**Prediction Intervals**

The final issue raised by Blanton et al. (2009) concerns determining the practical utility of the IAT. Blanton et al. address this issue by regressing a racial bias discrepancy score (i.e., $Y_{\text{White}} - Y_{\text{Black}}$) on the IAT. The 95% prediction interval for the predicted dependent variables were then examined to determine whether they included 0 (i.e., no bias) in them for particular IAT values. Blanton et al. report that the IAT in our study had limited utility because

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. (p. 578)

We agree with Blanton et al. (2009) that it is important to determine the practical utility of any test or organizational procedure. Indeed, there are many different methods proposed for generating bands (Aguinis, Cortina, & Goldberg, 1998; Cascio, Outtz, Zedeck, & Goldstein, 1991; Hanges & Gettmann, 2004; Hanges, Grojean, & Smith, 2000; Schmidt, 1995). Blanton et al.’s approach appears to be in the spirit of the Aguinis et al. (1998) approach, albeit Blanton et al. use regression to generate their prediction intervals. Despite the fact that the Blanton et al. approach appears to be conceptually close to established banding methodology, we have two issues with regard to the approach used by Blanton et al.

First, there are two different types of prediction confidence intervals that can be developed. In one approach, 95% prediction intervals can be generated around the expected value (i.e., the regression line’s predicted value) for a particular IAT value (Draper & Smith, 1981). In the second approach, prediction intervals can be generated to include some percentage (e.g., 95%) of all possible responses that individuals with a particular IAT score might produce (Draper & Smith, 1981). In other words, the first type of prediction interval (mean PI) estimates the average (i.e., expected value or true score using psychometric terminology) of the bias that would be produced by a particular IAT score, whereas the second type of prediction interval (individual PI) estimates the entire range of possible biases (i.e., true score plus error) that could be exhibited by an someone with a particular same IAT score. It is probably not surprising to learn that the mean PI is smaller than the individual PI.

Blanton et al. (2009) used the individual PI in their estimates. We believe that this is inappropriate. Rather than trying to estimate the range of the distribution of possible values for a given IAT score, we believe that it is more important to produce a prediction interval around the expected value of the dependent variable for a given IAT score. To illustrate this point, we regressed the same racial bias discrepancy score used by Blanton et al. ($Y_{\text{White}} - Y_{\text{Black}}$) on to the IAT for our entire 2005 sample (Ziegert & Hanges, 2005). We then computed the 95% mean prediction intervals for this analysis. In contrast to Blanton et al.’s results, the person with the highest IAT score had a 95% prediction interval was 0.7 to 1.59. Indeed, only 6 participants in our entire sample had mean prediction intervals that included zero. Thus, this analysis reveals that the IAT has more practical utility for differentiating among biased raters than implied by Blanton et al.

Our second problem with the Blanton et al. procedure is that it ignored the substantial literature in the applied sciences that has long discussed issues surrounding practical utility (e.g., Brogden, 1946, 1949; Cronbach & Gleser, 1965; Taylor & Russell, 1939). This literature clearly shows that there are many factors, in addition to the validity coefficient, that influence the practical utility of a test. Practical utility is a function of the base rate of the predicted phenomenon (e.g., biased ratings), the costs (both hidden and observable) associated with the predicted phenomenon (e.g., lawyer costs to defend the company; costs associated with negative publicity; costs associated with failing to hire qualified workers), the number of people affected by the test, and so forth. All these factors combine to influence the practical utility of a procedure. Indeed, when these factors are considered, sometimes a procedure with limited predictive validity can still have substantial practical impact. Blanton et al. simply ignored these important additional

---

5 Blanton et al.’s (2009) racial bias discrepancy score is actually the reverse of our original slope measure (i.e., $Y_{\text{Black}} - Y_{\text{White}}$).
factors when they attempted to evaluate the practical utility of the IAT with our sample.

Supplemental Analyses

In their discussion, Blanton et al. (2009) call for researchers to "examine the role of implicit attitudes after controlling for relevant explicitly measured beliefs and attitudes" (p. 580). They go on to criticize explicit measures in typical IAT research and state that "until investigations of this type are performed, we think it premature to argue that a new form of attitude is being observed" (p. 581). Curiously, Blanton et al. chose to ignore the measures of explicit attitudes in our data. Our 2005 study (Ziegert & Hanges, 2005) included two established explicit measures of stereotypes: the Attitudes Towards Blacks Scale (ATB; Brigham, 1993) and the Modern Racism Scale (MRS; McConahay, Hardee, & Batts, 1981). Because Blanton et al. indicate that such analyses are important but yet do not report any results of these reanalyses, we have chosen to do so.

We reran our HLM analyses controlling for explicit racial attitude using the ATB and MRS. First, we reanalyzed Hypothesis 3 from Ziegert and Hanges (2005) with MRS and ATB included in the model. The interaction of the IAT and climate manipulation on the racial bias slope remained significant ($\gamma = -.18$, $t(91) = -2.09$, $p < .05$). Next, given Blanton et al.'s (2009) criticism of our manipulation, we looked at the effect of the IAT across condition when controlling for MRS and ATB. When we controlled for explicit racism and condition, the IAT was still related to discrimination ($\gamma = -.11$, $t(92) = -2.60$, $p < .05$). Given Blanton et al.'s concern for outliers, we also reran our outlier analyses after controlling for the MRS and ATB. After deleting the four cases with high leverage values, we included the MRS and ATB in the HLM analyses. Results were still significant such that the interaction of the IAT and condition was related to our measure of discriminatory behavior after including explicit attitudes in the HLM equation ($\gamma = -.22$, $t(87) = -2.18$, $p < .05$). After removing the four values with extreme studentized residuals and then including MRS and ATB in the HLM analyses, we found that the interaction remained significant ($\gamma = -.15$, $t(87) = -2.19$, $p < .05$). Finally, in terms of decomposition of effects, the IAT predicts the average ratings of the three Black applicants within the climate for racial bias condition after one controls for the MRS and ATB ($\gamma = -.22$, $t(42) = -2.56$, $p < .05$).

Overall, these findings illustrate that the IAT predicts discriminatory behavior over and above explicit attitudes. This finding was consistent along several reanalyses: (a) replication of the original finding from Ziegert and Hanges (2005), (b) when one controls for condition, (c) after one eliminates potential outliers, and (d) when one decomposes the effects to examine just Black applicants. These findings directly address Blanton et al.'s (2009) recommendation that researchers should control for explicit attitudes and illustrate that the IAT assesses attitudes distinct from traditional explicit measures and these implicit attitudes are predictive of behavior above and beyond traditional measurement techniques.

Conclusion

In their article, Blanton et al. (2009) make very strong claims against the conclusions of our 2005 study (Ziegert & Hanges, 2005). We hope that the reader will agree, however, that we have provided further evidence in this article that our original conclusions were appropriate. Blanton et al. claim that they "failed to find a robust relationship between IAT scores and discriminatory behavior" (p. 578). In contrast to their conclusion, we provided a number of additional analyses that repeatedly demonstrated the predictive validity of the IAT. Contrary to Blanton et al.'s claims, we found that the IAT predicted discriminatory behavior across experimental conditions, it predicted discriminatory behavior even after outliers are eliminated, it predicted discriminatory behavior even after we dropped the White applicants and performing our analyses on only the Black applicants, and it predicted discriminatory behavior after we controlled for the explicit racial bias measures. In summary, we found that the predictive validity of the IAT was quite robust.

Blanton et al. (2009) also claim that "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" (p. 578). Again, our reanalyses reported in this article demonstrate that this criticism is baseless. Blanton et al. only used visual inspection to detect outliers, a procedure that is notoriously imprecise. We reanalyzed our data using four different methods for detecting statistical outlier and influential observations. We found that the IAT–climate manipulation interaction was still significant and robust. Indeed, one appropriate conclusion of our results is actually provided by Blanton et al. themselves. As Blanton et al. state, "when we examined IAT–behavior correlations within each climate condition, Ziegert and Hanges’s finding was resistant to the influence of outliers" (p. 578).

Blanton et al. (2009) also raise several methodological issues and they note that "given the unique methods employed in Ziegert and Hanges [2005] . . . the results of Ziegert and Hanges do not suggest a robust relationship between implicit bias, as measured by the IAT, and discriminatory ratings of Black candidates" (p. 578). Again, we hope that we have convinced the reader that this claim is not well supported either. Even if one were to agree with Blanton et al. regarding the unrealistic nature of our climate manipulation, our reanalyses illustrate that the IAT predicts discriminatory behavior across conditions. That is, we have a significant main effect between the IAT and discriminatory behaviors even after controlling for our climate manipulation. In other words, the reanalyses reported in this article suggest that the IAT’s predictive validity is more robust than we initially suggested in the original Ziegert and Hanges (2005) article.

Further, Blanton et al. (2009) claim that the IAT “could not delimit the range of likely discrimination in any meaningful way” (p. 578). Our reanalyses illustrate that the relationship between the IAT and discriminatory behavior is moderate in size. Further, when we created 95% mean prediction intervals, the vast majority of our participants had prediction intervals that did not overlap with zero. All other factors remaining constant, the IAT appears to have some practical utility in our 2005 study (Ziegert & Hanges, 2005).

Finally, we illustrated that the IAT significantly predicted biased ratings even after we controlled for explicit racial attitudes. Blanton et al. (2009) indicate that this analysis is very important to establish the predicting validity of the IAT. Interestingly, our reanalysis provide evidence that the IAT demonstrated “improved
criterion prediction after known determinants of behavior—properly measured and conceptualized—have been controlled” (p. 580–581). Thus, based on Blanton et al.’s own suggestion, this finding supports the increased utility and predictive validity of the IAT.

In sum, as Blanton et al. (2009) note strong claims require strong evidence. We have shown that Blanton et al.’s claims, however, are based on weak evidence. Overall, we believe that our arguments and reanalyses have provided a strong rebuttal in support of the original Ziegert and Hanges (2005) findings as well as illustrating the broader predictive validity of the IAT on discriminatory behavior.

References


Received July 14, 2008
Revision received September 30, 2008
Accepted October 23, 2008

---

**Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted**

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write APA Journals at Reviewers@apa.org. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.

- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.

- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, “social psychology” is not sufficient—you would need to specify “social cognition” or “attitude change” as well.

- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.