Journal of Applied Psychology
2015, Vol. 100, No. 5, 1468–1481

0021-9010/15/$12.00 http://dx.doi.org/10.1037/a0038379

Toward a Meaningful Metric of Implicit Prejudice

Hart Blanton
University of Connecticut

James Jaccard
New York University

Erin Strauts
University of Connecticut

Gregory Mitchell
University of Virginia

Philip E. Tetlock
University of Pennsylvania

The modal distribution of the Implicit Association Test (IAT) is commonly interpreted as showing high levels of implicit prejudice among Americans. These interpretations have fueled calls for changes in organizational and legal practices, but such applications are problematic because the IAT is scored on an arbitrary psychological metric. The present research was designed to make the IAT metric less arbitrary by determining the scores on IAT measures that are associated with observable racial or ethnic bias. By reexamining data from published studies, we found evidence that the IAT metric is "right biased," such that individuals who are behaviorally neutral tend to have positive IAT scores. Current scoring conventions fail to take into account these dynamics and can lead to faulty inferences about the prevalence of implicit prejudice.

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

Since passage of landmark civil rights legislation in the 1960s, survey researchers have observed steady declines in overt racism and other forms of prejudice (Carmines, Sniderman, & Easter, 2011; Schuman, Steeh, Bobo, & Krysan, 1997). Many researchers have been skeptical of these trends and have pointed to the pervasiveness of inequality in American society as evidence that hidden forms of prejudice must be alive and well (e.g., Dovidio & Gaertner, 2010). Recently, psychologists have drawn on the distinction between explicit and implicit attitudes to develop methods of capturing sentiments that self-report-based surveys allegedly fail to detect, with several creative measures of implicit attitudes having been developed over the past 15 years (see Petty, Fazio, & Briñol, 2008; Wittenbrink & Schwarz, 2007). These measures yield response patterns that are widely interpreted as showing pervasive *implicit biases* (a term that we use to encapsulate the hidden forms of implicit prejudicial associations that ostensibly are detected by these measures). By far, the most influential of these new measures is the Implicit Association Test (IAT). Modal distributions of IAT scores—especially from data collected from millions of administrations of the test over public websites—suggest that survey researchers have substantially underestimated contemporary levels of prejudice (e.g., Nosek et al., 2007), which has led to calls for legal and organizational reforms to address the high levels of implicit racial biases revealed by this measurement advance (e.g., Ayres, 2001; Bennett, 2010; Levinson & Smith, 2012; Saujani, 2003; see Banaji & Greenwald, 2013).

Such applications of IAT test scores are problematic, however, because the IAT relies on an *arbitrary metric*: Computed IAP scores have yet to be systematically mapped onto true scores on the underlying dimension of implicit bias and, ultimately what is of policy interest, actual discrimination (Blanton & Jaccard, 2006a). As such, researchers cannot definitively link the degree of behavioral bias to any specific IAT score, and it follows that they also cannot use the distribution of IAT scores to infer either the prevalence or the average magnitude of behavioral bias in any given group. We introduce methods that can help give meaning to IAT scores and then apply these methods to published studies that have used IAT measures of racial and ethnic bias in order to predict discriminatory behavior. Our analyses offer evidence that the zero point on the IAT fails to map onto behavioral racial neutrality and that researchers' tendency to treat the zero point as valid likely leads to significant overestimation of the degree of implicit bias in populations.

This article was published Online First January 19, 2015.

Hart Blanton, Department of Psychology, University of Connecticut; James Jaccard, Silver School of Social, Work New York University; Erin Strauts, Department of Psychology, University of Connecticut; Gregory Mitchell, School of Law, University of Virginia; Philip E. Tetlock, Department of Psychology and Wharton School, University of Pennsylvania.

Gregory Mitchell and Philip E. Tetlock are consultants for LASSC, LLC, which provides services related to legal applications of social science research, including research on prejudice and discrimination. We thank David M. Amodio, Jeremy D. Heider and Jonathan C. Ziegert for providing the datasets needed for the secondary analyses and Wilhelm Hofmann for providing both data and extensive comments on the first draft of this manuscript.

Correspondence concerning this article should be addressed to Hart Blanton, Department of Psychology, Unit 1020, 406 Babbidge Road, University of Connecticut, Storrs, CT 06269-1020. E-mail: hart.blanton@uconn.edu

Arbitrary Metrics

Many psychological constructs are hypothetical in nature and cannot be directly observed. Researchers infer standing on these constructs by placing individuals at different points on observable metrics. Used here, the term *metric* refers to the numbering system that the researcher employs when describing individuals' standings on the construct of interest. In much psychological research, the metric is arbitrary: it is not known exactly where a given score locates an individual on the underlying psychological dimension, nor is it known how a one-unit change on the observed score corresponds to the magnitude of change on the underlying dimension (Blanton & Jaccard, 2006a). Stated more formally, observed scores are thought to be some function of the underlying true scores, but the exact function form typically is not known—it might be linear (which is the assumption of interval measurement) or nonlinear (such as a logit function, which is assumed by many Item Response Theory measurement models). In much applied research, linear functions between observed and true scores are assumed, but even then, the values of the slope and intercept linking the observed scores to the true scores are unknown, thereby yielding an arbitrary metric (see Edwards & Berry, 2010).

Arbitrary metrics are quite common in research. The arbitrary nature of metrics in the field of prejudice does not necessarily limit theory tests if researchers are not invested in making claims about the absolute meaning of specific scores on measures of prejudice. By assuming interval-level measurement, many forms of theory tests can be effectively implemented (Lord, 1952; Lord, Novick, & Birnbaum, 1968; Nunnally, 1967). However, if one seeks to imbue interpretation of an observed score on a scale as reflecting a certain degree or amount of the underlying construct (e.g., prejudice), then the arbitrariness of the metric is problematic and steps must be made to make the scale nonarbitrary.

The IAT and Arbitrary Metrics

The IAT is a speeded binary-classification task designed to quantify attitudes on a metric derived from sets of two measurable response latencies. Respondents are shown stimuli on a computer screen and are asked to press a key on the right or left side of the computer, depending on the category to which the stimulus belongs. In the case of the "race IAT" (which was designed to measure implicit racial preferences for Whites over Blacks), half the stimuli are photos of either Black or White faces and the other half are words that are either positive or negative in character. For half of the trials (the "compatible task"), the individual presses one key if the face is White or the word is positive and a different key if the face is Black or the word is negative. For the other half of the trials (the "incompatible task"), the individual presses one key if the face is White or the word is negative and a different key if the face is Black or the word is positive. It is assumed that an individual who is prejudiced against Blacks will be slower to respond in the latter, incompatible task than the former, compatible task. The latencies of each response on a given trial are recorded (in milliseconds), and the mean of these latencies are subtracted from one another as part of a complex scoring algorithm designed to expunge known sources of method-specific variance (Greenwald, Nosek, & Banaji, 2003; cf., Blanton, Jaccard, & Burrows, in press).

On the IAT metric, positive scores are interpreted as revealing an implicit preference for Whites over Blacks, negative scores are interpreted as revealing an implicit preference for Blacks over Whites, and scores near zero are interpreted as revealing little or no implicit bias. IAT researchers thus assume that the IAT possesses what is often termed a *rational zero point*, namely the observed value of 0 on the IAT maps onto the true zero point on the underlying bipolar dimension of prejudice that separates preferences for Whites over Blacks from preferences for Blacks over Whites. Note that our use of the term *rational zero point* is distinct from the notion of an absolute zero in Steven's classic taxonomy of measurement. As used here, a rational zero point applies only to bipolar constructs and refers to the true zero point of "neutrality" that separates the negative side of the construct from the positive side of the construct. The IAT assumes that a score of 0 on the observed metric maps directly onto this rational zero point. Despite this, there is little empirical evidence to support this supposition. The present research provides perspectives on the extent to which equal treatment of Blacks and Whites is reflected by scores of zero on the IAT.

The score of zero is not the only number on the IAT metric that has been imbued with real-world meaning without an empirical rationale. Researchers also use cut points to categorize respondents in terms of their degree of implicit bias, and, although not widely publicized, the cut values used to categorize respondents have shifted over the years. Prior to 2003, the researchers who maintained the primary IAT website based their bias categories on Cohen's (1988) criteria for categorizing small, medium, and large effect sizes based on mean differences (d scores of 0.20, 0.50 and 0.80). They measured the difference of the mean RT for the two trials measured in raw response latencies (measured in milliseconds) and converted this to a d score for the difference using Cohen's classic formula for *d* based on a group level standard deviation; they then labeled respondents as having *no*, *slight*, *moderate*, and *strong* biases depending on if they exceeded Cohen's criteria for labeling effects as *small*, *medium*, and *large* (Brian Nosek, personal communication, August, 2002). After 2003, with the adoption of a new scoring algorithm (Greenwald et al., 2003), the IAT was placed on a new metric (called a "D score" as opposed to a "d score"), and the distribution of scores on the metric changed accordingly. It turns out that if Cohen's traditional conventions had been applied to this new D score metric, the appearance of "strong" racial biases would have spiked considerably. This did not happen, however, because the architects of the IAT instead adopted new cut points for their D score, centered around the D values of 0.15 (*slight preference*), 0.35 (*moderate preference*) and 0.64 (*strong preference*). These values map roughly onto Cohen effect size (*d*) values of 0.30, 0.70, and 1.3. With this shift in the scoring algorithm and the cut values used to specify the level of bias associated with a score, the percentage of Americans labeled as holding a strong racial preference for Whites diminished considerably relative to the percentages that would have been so labeled prior to 2003 (see Blanton & Jaccard, 2008 for direct comparisons). The problem with the *d* score and D score cut-offs, of course, is that both are arbitrary in character and cannot properly be treated as indicators of "true" levels of bias. We make use of both the *d* score and D score cut-off strategies in the present research.

Making the IAT Metric More Meaningful

Arbitrary metrics become meaningful metrics when researchers conduct studies to map specific observed scores onto outcomes, events, or behaviors that have consensual meaning in relation to the unobserved, theoretical dimensions of interest (Blanton & Jaccard, 2006a; see Kazdin, 2006; LeBel, 2011; Sechrest, McKnight, & McKnight, 1996). With a psychological inventory that measures attitudes relevant to hiring bias, for instance, one could identify the specific attitudinal score where one expects there to be no influence of applicant race on employer hiring decisions. It would be meaningful to view that observed score—whatever the specific value was on the observed metric—as an *empirical zero-point* because at that score, employers are as likely to hire Whites as to hire Blacks, all else being equal. If a group of studies shows consistent patterns in the relation of sets of scores to types of behavior, then meaningful "cut points" on the metric can be identified. For hiring bias, for instance, one could identify the specific scores on a racism inventory for which an employer will give qualified Black job applicants 10%, 20%, or 30% less likelihood of being hired relative to equally qualified White competitors. As the full range of scores take on meaning over time, consensus might emerge among scientists and practitioners about which specific values indicate meaningful shifts from "slight" to "moderate" to "strong" degrees of racial bias by an employer.

Evidence From Published Reports

Efforts to chart the IAT metric have not been systematically pursued, but results from published reports suggest that behavioral neutrality occurs at larger values than a score of zero on the IAT. Consider McConnell and Leibold (2001). This study examined whether the race IAT predicts the quality of interactions between White participants and a Black versus a White experimenter. Estimates of interaction quality were based on judges' ratings of the friendliness, abruptness, and comfort level that participants exhibited during the two types of interactions. McConnell and Leibold reported that scores on the race IAT were significantly correlated with the difference in interaction quality for White versus Black experimenters, such that the greater the implicit preference for Whites observed on the IAT, the more positive interactions were with the White relative to Black experimenter, $r = .34$.

Although this finding suggests some predictive utility of the race IAT (cf. Blanton et al., 2009; Oswald et al., 2013), a critical piece of information necessary to evaluate the metric was not reported in the original publication. In their reanalysis of the data, Blanton et al. (2009) found that the White–Black interactions were actually more positive, on average, than the White–White interactions. This suggests there was a disconnect between the IAT and behavior. Specifically, 90% (37 of 41) of McConnell and Leibold's (2001)'s participants had positive IAT scores (traditionally interpreted as revealing a pro-White bias in the sample), whereas 70% of the participants reacted more favorably to the Black than White experimenter (suggesting a pro-Black bias in the sample). Blanton et al. (2009) published the full scatter plot and plotted the regression line linking IAT scores to the behavioral difference score. From their Figure 2 (p. 575), it is evident that behavioral neutrality occurred at roughly an IAT score of 0.42 (on a metric that ranges from −2 to +2). Using the reported means, sample sizes, and standard deviations in the published reports, we found that the predicted mean IAT value was statistically significantly greater than 0 when there was behavioral neutrality, $t(40) = 20.69$ $p < .05$.[1] This suggests what we term a *right bias* in the IAT metric measuring prejudice, in that individuals who treated Blacks and Whites identically were scored on the positive side of the distribution of the IAT, the end that is typically labeled as indicative of anti-Black attitudes.

Unfortunately, researchers rarely report summary statistics associated with behavioral differences between Blacks versus Whites in IAT studies, focusing readers' attention instead on the high levels of implicit bias suggested by the IAT distributions themselves. But other published results are suggestive of a right bias in IAT scores. Ashburn-Nardo, Knowles, and Monteith (2003) had 83 Black participants rate their preferences for working with potential Black and White partners. They found that the stronger the implicit preference for Whites relative to Blacks, the stronger the stated preference for a White relative to a Black interaction partner, $r = .23$. However, inspection of their Figure 2 (p. 77) suggests that behavioral neutrality mapped onto a raw IAT score (not a D score) of about 250 milliseconds. The value of 250 milliseconds was statistically significantly larger than 0, $p < .05$, a finding traditionally interpreted as indicative of anti-Black bias.[2] In another report, Greenwald, Smith, Sriram, Bar-Anan, and Nosek (2009) studied a sample of 1,057 U.S. participants in an online survey who were over the age of 18 and who expressed a voting preference in the 2008 election. Although the sample strongly favored Barack Obama (82.4%) over John McCain (15.8%), it was also characterized by positive IAT scores measured on the D-score metric. IAT scores in this sample were significantly greater than zero in a direction suggesting an anti-Black bias ($M = 0.06$, $SD = 0.42$), $t(1056) = 4.65$, $p < .05$, even though the sample had a strong pro-Black (Obama) voting preference.

These two studies suggest that the "right bias" observed in McConnell and Leibold (2001) may be common, but informal inspections such as these of published reports have their limitations. It is possible, for instance, that the IAT metric is right-biased but that the magnitude of the bias has little or no practical consequence in terms of the characterizations one would make about the prevalence of implicit bias. Also, the right bias may not be grounded in formal metric issues but instead be an artifact of some other methodological facet of a study. For example, the Black experimenter in McConnell and Leibold (2001) might have been a more likable individual than the White experimenter. As a result,

[1] Although the McConnell and Leibold (2001) data were made available to us for Blanton et al. (2009), the lead author would not consent to our use of the data set for this new project. The results we report here were therefore taken entirely from public summaries reported in journal articles. This limited the scope of our analysis. McConnell and Leibold also collected data on 13 other dimensions that were meant to quantify the quality of the interaction with the Black and White experimenter (e.g., smiling, eye contact). Difference scores on these criteria could be used to gain additional perspectives on the degree of alignment between the rational and empirical zero points in this study. However, there was insufficient information in the published articles to generate estimates to estimate these values.

[2] The direction of the criterion metric in Figure 2 was scaled to measure higher pro-Black bias and so this estimate of the IAT effect is reversed in our report, to be consistent with the scorning norms of other papers reported here.

only those unusually high in anti-Black implicit bias would have reacted similarly to the two individuals. We sought to address these limitations by using a broader range of studies in the racial and ethnic domain and by applying formal estimation methods.

Analytic Approach

Location of an empirical zero point requires the identification of one or more observable, measureable criteria that strong theory suggests should only be observed in individuals who are located at the true theoretical zero-point of neutrality (LeBel, 2011; Sechrest et al., 1996). Such criteria can be difficult to identify, but one possibility was suggested by Greenwald, Nosek, and Sriram (2006), who utilized a criterion based on the zero point obtained from measures of explicit attitudes. Suppose we identify individuals who show no difference in their explicit ratings of attitudes toward Blacks versus Whites on separate scales (an explicit attitudinal difference score of 0 on the two scales). This observed difference score can be treated as a "gold standard" against which one tests the validity of the IAT zero point. If individuals with an IAT score of zero tend to have explicit attitude (difference) scores of zero, then this, according to Greenwald, Nosek et al. (2006), supports the zero point of the IAT. Statistically, one regresses the IAT score onto the difference in explicit evaluations of Whites versus Blacks. The intercept is the predicted mean IAT value when the explicit attitude difference is zero. The hypothesis is that the intercept in the resulting equation should equal zero. Such a test was reported by Greenwald, Nosek et al. (2006), who found evidence consistent with a zero intercept for implicit and explicit Presidential attitudes (although this result is not consistent with the pattern just reported for Greenwald, Smith et al., 2009, linking race to voting; nor was it replicated by Blanton & Jaccard, 2006b, who also focused on implicit and explicit race attitudes).

Use of explicit attitudes as the gold standard is problematic because social scientists have argued that implicit and explicit attitudes, particularly with respect to intergroup attitudes, are based on independent cognitive/affective systems that may not be strongly linked. They also have argued that self-reports of explicit attitudes may be subject to distortion because of social desirability considerations and introspection limits (e.g., Greenwald, McGhee & Schwartz, 1998; Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009). To the extent that the above is true, it is questionable to assume that a person's explicit claim of racial neutrality can be used to validate a person's implicit racial neutrality.

A better, albeit still imperfect, test focuses on indices of behaviors as opposed to attitudinal differences as the gold standard. For the race IAT, one would expect to observe an IAT score of zero for respondents whose observed treatment of Black people is comparable to their observed treatment of White people. If people act more positively toward Whites than Blacks, one would expect the IAT score to be greater than zero. If people act more positively toward Blacks than Whites, one would expect the IAT score to be negative. This logic also can be tested through regression analyses, but with an equation in which race IAT scores are regressed onto a behavioral difference score rather than an attitudinal difference score. The intercept in this equation reveals the mean IAT score one expects to observe among individuals who exhibit no behavioral preference for Whites versus Blacks.

The present research adopts this behavior-differential approach to yield empirical perspectives on zero points in IAT research. The fundamental question we address is whether people who treat Blacks and Whites roughly the same have a zero (or near zero) score on the IAT. Stated another way, does behavioral neutrality map onto IAT neutrality? One possible answer to this question is a simple affirmative, whereby individuals who treat Whites and Blacks the same evidence zero or near-zero IAT scores. A more plausible answer, to us, is that the behavioral and IAT zero points will not map onto one another because there is random "noise" (i.e., other factors) that will sometimes push behavioral bias upward or downward depending on the circumstances and context. This logic model predicts an average correspondence across studies (as the positive biases cancel the negative biases), but not necessarily correspondence for any given study. As will be seen, both of these possible answers proved to be wrong.

To introduce our analytic method, we first present our reanalysis of a set of studies reported by Hofmann, Gschwendner, Castelli, and Schmitt (2008). Their studies used IAT measures designed to assess evaluative preferences for Italians versus Africans in a sample of Italians (Study 1) and evaluative preferences for Germans versus Turks in a sample of Germans (Study 2). We examine these two studies in detail for the purposes of evaluating the properties of the metric employed.[3] We then present results obtained by applying the same method to other studies in which IAT scores were correlated with discrimination criteria.

Zero-Point Analysis of Hofmann et al. (2008)

Overview

Hofmann et al. (2008) reported two studies that employed a semistructured interview procedure to examine the impact of implicit preferences on the treatment of persons from majority and minority groups. Italian participants in Study 1 interviewed both an African and an Italian confederate, ostensibly for the purpose of becoming acquainted with and learning interview procedures for subsequent interactions. Interviews were video-taped and content coded by independent judges, allowing for direct comparisons of the treatment of members of the two groups. The two interviews were carried out in one of two experimental conditions. Participants in a *Memory Load* condition were instructed to conduct the interviews while also working to recall as many words as they could from a list of 20 words observed earlier. Participants in a *Full Resource* condition were given no such instructions. Study 2 followed much the same procedure, with minor changes to the timing of interviews, the cognitive load procedure and number of trials in the IAT. The major modification was that Study 2 utilized German participants, who interviewed either Turkish or German confederates.

IAT Measure

The IAT in Study 1 employed photos of eight Italian and eight African males, and the IAT in Study 2 used eight German and

[3] We note that the focus of their research was on whether cognitive load would moderate the relationship between IAT scores and behavior. The authors did not state any hypotheses or draw inferences regarding the prevalence of implicit bias based on IAT zero scores.

eight Turkish facial stimuli. The standard D scoring algorithm was used in this study, with IATs in each study scored such that higher scores represented more negative evaluations of the out-group. However, in the data set provided to us by the authors, IAT scores for Study 1 were also computed in the raw response latencies and the log latencies. Because these metrics also have a history of use in the field, we decided to conduct metric analyses of all three scoring procedures for Study 1.

Outcome Measures

Participant-confederate interactions in both studies produced several measures that can be used to judge relative treatment of the Italian and African candidates: (a) *global ratings* of the overall quality of the interactions were computed from observers' ratings of the interactions along five dimensions using 7-point scales (polite-impolite, relaxed-nervous, pleasant-unpleasant, talkative-quiet, and warm-cold); (b) observers rated participants for the amount of *eye contact* made with the interviewee, (c) observers rated participants for the use of *speech illustrators* (e.g., waving of hands and arms to emphasize a point); (d) observers rated participants for the use of *body adaptors* (touches of one's own body or head); (e) participants made *competence ratings* of the skills exhibited by each interviewee. The original researchers hypothesized that more eye contact and greater use of speech illustrators and body adaptors indicated a greater comfort and engagement on the part of participants. For each of these criteria, difference scores were computed, such that positive values indicated more negative treatment of the African confederate relative to the Italian confederate.

Table 1
Zero-Order Correlations Between IAT Score and Primary Outcome Measures in Hoffman et al. (2008)

	Memory load	Full resource
Study 1 (D Score)		
Speech illustrators	0.32*	−0.09
Body adapters	−0.02	0.13
Eye contact	0.17	−0.05
Global rating	0.01	0.13
Rated competence	0.08	0.11
Study 1 (Raw score)		
Speech illustrators	0.27*	−0.13
Body adapters	−0.01	0.07
Eye contact	0.12	0.07
Global rating	0.01	0.16
Rated competence	0.03	0.08
Study 1 (Log score)		
Speech illustrators	0.27*	−0.09
Body adapters	−0.02	0.00
Eye contact	0.11	0.03
Global rating	0.08	0.16
Rated competence	−0.06	0.01
Study 2 (D Score)		
Speech illustrators	0.31*	−0.06
Body adapters	0.36*	−0.15
Eye contact	0.32*	−0.07
Global rating	0.06	0.14
Rated competence	−0.06	0.06

Note. *N* ranged from 84 to 86 in Study 1 and from 76 to 77 in Study 2, depending on missing values.
* reflects parameter estimate is statistically significant from a null value of 0 at $p < .05$.

Results

Zero-order correlations. We replicated Hofmann et al.'s results with only minor, inconsequential disparities. Table 1 presents the correlations between the IAT measures used and each of the five behavioral outcomes for each of the two studies.[4]

Empirical zero-point analysis. We regressed the IAT score onto each of the five bias criteria separately for each experimental condition. Analyses were carried out on the D score for Studies 1 and 2. Given the additional data provided to us for Study 1, we repeated the procedure for the raw response latencies and log latencies from the IAT.

The intercepts for each analysis (see Table 2) reflect the expected mean IAT score when behavior toward Blacks and Whites is equally positive (i.e., when the behavioral difference score is zero). For all criteria, conditions, and scoring approaches, the intercepts were statistically significantly different from zero and positive, indicating that the IAT metrics were all "right biased." That is, even individuals who acted in an unbiased, neutral manner in the conditions created by Hofmann et al. (2008) produced positive IAT scores that are often interpreted as evidence of implicit racial or ethnic bias. In fact, based on the intercepts, participants who were not behaviorally biased typically had an IAT D score of roughly 0.61 in Study 1, which according to current coding conventions is near the upper limit of what would be labeled "moderate" bias in favor of Whites (i.e., 0.65). In Study 2, these participants had an IAT D score of roughly 0.32, which is just under the cutpoint for that same categorization. Also noteworthy, the degree of right bias was consistent across the different behavioral measures in each study. Thus, our analysis suggests that a wide range of positive IAT scores may not be associated with biased behavior and should be viewed cautiously and not simply assumed to be markers for implicit racial bias.

Empirically correcting the IAT distribution for zero-point bias. One can correct for the right bias in the IAT by subtracting the value of the intercept in the regression analysis from each individual's IAT score so that the observed value of zero on the IAT corresponds to the best estimate of the theoretical zero point where behavior is equal toward Blacks and Whites. If the predicted value of the IAT D score in Study 1 is 0.61 when behavior is racially neutral, for example, then the IAT D score is recalibrated by subtracting 0.61 from it so that it will equal 0.0 when the behavioral differential equals 0.00. We performed this recalibration for both Study 1 and 2 and then applied cut points to characterize the degree of implicit bias in the sample. For the D score, we applied the conventions currently employed for the D-score. For the log and raw scores in Study 1, we applied Cohen's criteria.

Table 3 presents the percentage of individuals in each category for the uncorrected and recalibrated IAT scores. With uncorrected

[4] The original authors focused on regression coefficients in models testing linear interactions between the IAT and the experimental condition and did not report the zero-order correlations or main effects for each outcome by condition.

Table 2
Intercept From Regressing IAT Onto Behavioral Criteria in Hoffman et al. (2008)

Outcome	Memory load	Full resource
Study 1 (D Score)		
Speech illustrators	0.63 (± .10)	0.61 (± .10)
Body adapters	0.62 (± .10)	0.60 (± .11)
Eye contact	0.60 (± .10)	0.62 (± .10)
Global rating	0.62 (± .11)	0.62 (± .11)
Rated competence	0.63 (± .10)	0.62 (± .10)
Study 1 (Raw score)		
Speech illustrators	1.41 (± .37)	1.32 (± .38)
Body adapters	1.39 (± .39)	1.34 (± .39)
Eye contact	1.32 (± .39)	1.38 (± .38)
Global rating	1.42 (± .39)	1.39 (± .38)
Rated competence	1.37 (± .39)	1.37 (± .38)
Study 1 (Log score)		
Speech illustrators	0.18 (± .04)	0.17 (± .04)
Body adapters	0.18 (± .04)	0.17 (± .04)
Eye contact	0.17 (± .04)	0.18 (± .04)
Global rating	0.18 (± .04)	0.18 (± .04)
Rated competence	0.18 (± .04)	0.18 (± .04)
Study 2 (D Score)		
Speech illustrators	0.35 (± .06)	0.33 (± .07)
Body adapters	0.32 (± .06)	0.33 (± .06)
Eye contact	0.33 (± .06)	0.33 (± .07)
Global rating	0.34 (± .07)	0.33 (± .06)
Rated competence	0.33 (± .06)	0.33 (± .06)

Note. Values in parentheses are for 95% margin of error.

IAT scores, all three scoring approaches suggest a large percentage of individuals showing bias favoring Whites in Study 1 (84% with the D-score, 78% with raw scores, and 81% with log scores) and Study 2 (79% with the D-score). A different picture emerges when scores are adjusted for right bias. To provide for the most conservative correction, we subtracted the smallest estimate of the intercept obtained from the analyses in Table 2. Even with this conservative approach, the estimates of pro-White bias decreased dramatically in Study 1 (to 42% with the D-score, 58% with raw scores, and 58% with log scores) and Study 2 (to 29% with the D-score).

Discussion

Our treatment of Hofmann et al. (2008) describes formal methods for mapping the empirical zero point of the IAT onto observable forms of racial and ethnic bias. Our analyses reinforce earlier concerns regarding disparities between the measured zero point of the IAT and the empirical zero point defined by external behavioral estimates of racial and ethnic bias, and they suggest these issues may generalize across the major scoring algorithms in the IAT literature. When IAT scores on these different algorithms are recalibrated so that the zero points are correspondent, conclusions about the extent of implicit bias change considerably. This result, however, derives from a single published report, making broader conclusions inappropriate.

Calibrating IAT Measures of Racial Bias Across Studies

To examine the robustness of the result found with the Hofmann et al. (2008) data sets, we sought to apply the analytic method just introduced to other published studies linking the IAT to meaningful criteria of discrimination. We were able to obtain raw data from three additional publications.

Study Overviews

Amodio and Devine (2006), Study 2. This study examined whether implicit attitudes and stereotypes predicted respondent's ($n = 32$) evaluations of a Black person's writing performance.

Heider and Skowronski (2007), Study 1. These researchers examined whether the IAT predicted the choices of White participants ($n = 140$) in a Prisoner's Dilemma Game (PDG) when they believed they were competing with a White versus Black partner.

Heider and Skowronski (2007), Study 2. This study examined whether the IAT predicted the friendliness of participants ($n = 55$) in interactions with Black and White experimenters.

Table 3
Frequency Distribution for IAT Categories in Hofmann et al. (2008), Before and After Correction

Study 1 (D Score) Category	Standard Scoring	Corrected Scoring (0.60)
Strong in-group bias	49%	7%
Moderate in-group bias	25%	17%
Slight in-group bias	10%	18%
No bias	5%	21%
Slight out-group bias	8%	16%
Moderate out-group bias	3%	7%
Strong out-group bias	—	13%
Study 1 (Raw Score) Category	**Standard Scoring**	**Corrected Scoring (1.32)**
Strong in-group bias	26%	14%
Moderate in-group bias	22%	17%
Slight in-group bias	30%	27%
No bias	14%	15%
Slight out-group bias	5%	13%
Moderate out-group bias	1%	5%
Strong out-group bias	1%	9%
Study 1 (Log Score) Category	**Standard Scoring**	**Corrected Scoring (0.17)**
Strong in-group bias	34%	14%
Moderate in-group bias	19%	17%
Slight in-group bias	28%	27%
No bias	13%	18%
Slight out-group bias	4%	10%
Moderate out-group bias	6%	5%
Strong out-group bias	2%	10%
Study 2 (D Score) Category	**Standard Scoring**	**Corrected Scoring (0.32)**
Strong in-group bias	10%	1%
Moderate in-group bias	27%	8%
Slight in-group bias	42%	20%
No bias	16%	48%
Slight out-group bias	4%	16%
Moderate out-group bias	—	8%
Strong out-group bias	—	—

Note. Generalized margin of error for predictions in Study 1 are ± 11% and for Study 2 are ±12%.

Ziegert & Hanges (2005). These researchers examined whether the IAT predicted evaluations of hypothetical White and Black job applicants. In one condition, participants ($n = 52$) were encouraged to select the best applicant (the climate of equality condition); in a second condition, participants ($n = 48$) were told that, "[g]iven that the vast majority of our workforce is White, it is essential we put a White person in the [vice president] position" (the climate for racial bias condition).

IAT Measures

Amodio and Devine (2006), Study 2 examined two race IATs: the standard race IAT that measures evaluative associations for Blacks relative to Whites, and a stereotype IAT in which participants viewed two classes of words associated with the characteristics of "intelligence" or "athleticism/rhythmicity" were presented along with pictures of Black and White faces. D scoring was used to score both IATs.

Heider and Skowronski (2007), Study 1 used a race IAT that measured preference for Whites relative to Blacks, using names as the racial stimuli. Scores were recorded using the raw metric in milliseconds.

Heider and Skowronski (2007), Study *2* measured preference for Whites relative to Blacks, using photos of Black and White faces as racial stimuli. Scores were recorded using the raw metric in milliseconds.

Ziegert & Hanges (2005) measured preference for Whites relative to Blacks, using photos of Black and White faces as racial stimuli. However, data from this IAT were scored in an untraditional way. The researchers recorded response latencies for the two (compatible and incompatible) IAT tasks and recorded the error rates for these two tasks (i.e., how often individuals made classification errors). The researchers then standardized both the RT scores and the error scores and averaged these indices together before computing the IAT difference score. To our knowledge, this scoring of the IAT has not been used in any other studies.

Outcome Measures

Amodio and Devine (2006), Study 2 had participants rate the attributes of a (fictitious) Black essay writer. No mention was made in this study of a White essay writer, nor were such data provided, and so a comparison of the treatment of Black and White individual is not possible. However, the researchers quantified discrimination by comparing how IAT correlated with differences in stereotypic ratings of the essay writer "on a list of adjectives known to be highly associated with the Black stereotype (*lazy*, *dishonest*, *unintelligent*, and *trustworthy*; Devine & Elliot, 1995) intermixed with filler traits that were relatively neutral and not typically associated with the stereotype (*modest*, *assertive*, and *thoughtful*)" (p. 656). Ratings were made on 10-point scales and combined in such a way that the computed difference score provided a measure of the degree to which ratings of the Black essay writer were more stereotypic than nonstereotypic. We used this difference score as our primary criterion from the study.[5] We augmented this with a criterion based on explicit attitudes. Amodio and Devine obtained "feeling thermometer" ratings (Nelson, 2008) for Black and White people in general and the difference between these two ratings provided an alternative standard to a behavioral difference.[6]

Heider and Skowronski (2007), Study 1 indexed racial discrimination in terms of cooperation versus competition with one's partner in a Prisoner's Dilemma game. The variable was scored as the proportion of trials in which the participant was cooperative.[7] In addition, it was possible to examine differences in explicit evaluations of Whites versus Blacks, using answers to thePro-Black/Anti-Black Attitudes Questionnaire (PAAQ; Katz & Hass, 1988) and a set of semantic differential ratings.

Heider and Skowronski (2007), Study 2 measured discrimination by judging the quality of two 3-min, semistructured interactions—one with a White confederate and another with a Black confederate. The quality of these interactions was assessed by two sets of judges who either viewed muted videotapes of the conversations and rated nonverbal friendliness or listened to an audio-only recording of the conversation to rate verbal friendliness. In addition, this study used the same two explicit attitude ratings as in Study 1.

Ziegert & Hanges (2005) had participants play the role of managers in a company and completed an "in-basket exercise" in which they were supposed to make a number of managerial decisions, including evaluations of the dossiers of eight job applicants on a scale from 1 (*should not have been referred*) to 5 (*excellent referral*). The race of the applicants was randomly assigned to the dossiers, with three qualified applicants being Black and three matched candidates being White (see Blanton et al., 2009; Ziegert & Hanges, 2009 for additional perspectives on the data collected in the original study).

Results

Replications. We were able to replicate key findings except that, as discussed in Blanton and Mitchell (2011), we were not able to replicate the results reported for Study 2 of Heider and Skowronski (2007). By comparing the original raw data with the data in the final report and through correspondence with the authors, we were able to confirm that some of the data used in the final report had been fabricated, in a manner consistent with hypotheses. We thus conducted the current analyses on the raw data that were used in the analyses reported by Blanton and Mitchell (2011).

Zero-point analyses. Table 4 presents all zero-order correlations for all key outcomes in each of the studies and presents the

[5] Devine and Elliot (1995) used the 84 adjectives originally developed by Katz and Braly (1935), with an additional nine attributes that they added to the list. Although all of the stereotype words or close synonyms from Katz and Braly were used, none of the "nonstereotypic words" were used, so one cannot be entirely sure that the desired scoring yields what was intended. This is a limitation, but we included the study nonetheless to be thorough.

[6] The primary hypothesis in this paper was for a stronger association between the stereotype IAT and explicit stereotype ratings of the writer and the evaluative "race" IAT and explicit attitudes. Amodio and Devine predicted that the stereotyping IAT should not predict affect-based outcome measures. To be fully inclusive, we used each of the two difference criteria as empirical standards that might provide insight into each of the two IAT metrics.

[7] The primary outcome variable in the published study was the proportion of trials of cooperation when one's partner was thought to be Black. The data provided to us also included data when one's partner was thought to be White, making possible zero-point calibration (see the reanalysis in Blanton & Mitchell, 2011).

intercepts obtained for each of the studies when the IAT was regressed onto the behavioral differential. Data from Heider and Skowronski (2007) and Amodio and Devine (2006) indicate the presence of the right bias found in Hofmann et al. (2008). For Ziegert and Hanges (2005), which used a unique IAT scoring approach that incorporated both error rates and latencies after standardization, the intercepts trended in a negative direction, although neither differed statistically significantly from the value of zero.

Empirically correcting the IAT distribution for zero-point bias. Table 5 presents the classification of individuals into degree of bias categories using both the original and recalibrated IAT scores in each study. In cases where we had multiple correction factors to choose from (based on analyses of different behavioral criteria), we chose the smallest intercept value to provide the most conservative recalibration of the IAT. Table 5 shows that even with this conservative approach, recalibration leads to substantial shifts in inferred degree of implicit bias in the samples.

Discussion

We found evidence for a right bias in the standard scoring algorithms of the IAT, such that a large number of individuals with positive IAT scores acted equitably or even more favorably toward Blacks than Whites. The one exception to this general trend is found in Ziegert and Hanges (2005), a study that employed a novel scoring of the IAT. Because Ziegert and Hanges' approach incorporated error scores into the computation and applied across-individual standardization, the mean structure of the data varied dramatically from that found with standard IAT algorithms. As shown in Table 5, the uncorrected distribution for this metric is notable for clustering individuals at either extreme of being strongly biased in favor of Whites or Blacks. With the standard scoring algorithms in all the other studies, respondents were clustered only in the direction of favoring Whites, though our analyses suggest these distributions are poorly calibrated with observable bias.

Table 4
Race IAT Zero-Order Correlations and Intercepts

D-Score metric	*r*	Intercept	MOE
Amodio & Devine (2006), Study 2			
Race IAT			
Stereotyping	−0.10	0.30	0.11
Explicit attitude	0.14	0.30	0.08
Stereotype IAT			
Stereotyping	0.22	0.36	0.14
Explicit attitude	0.10	0.28	0.11
Raw score metric	***r***	**Intercept**	**MOE**
Heider & Skowronski (2007), Study 1			
Cooperation in PDG	0.03	274	33
Explicit attitude criterion (PAAQ)	0.04	272	33
Explicit attitude criterion (Sem Diff)	0.02	274	34
Heider & Skowronski (2007), Study 2			
Nonverbal friendliness	0.16	162	53
Verbal friendliness	0.02	172	56
Explicit attitude criterion (PAAQ)	0.11	176	49
Explicit attitude criterion (SD)	0.15	164	53
Standardized error and latency metric	***r***	**Intercept**	**MOE**
Zeigert & Hanges (2005)			
Climate for equality			
Evaluation	0.11	−0.30	0.55
Climate for racial bias			
Evaluation	0.33*	−0.28	0.61

Note. MOE is margin of errors based on 95% confidence intervals.
* significant at $p < .05$.

Although the general trend across studies reinforces the argument for a right bias in the IAT measure of prejudice, caution again is warranted. No single behavior is likely to capture the many ways a psychological construct might express itself, and no single study can inform all the contexts in which implicit bias might influence behavior. We had hoped to obtain a larger number of data sets to pursue our analyses, but we were not able to do so (Blanton et al., 2009 documents the reasons why we were unable to obtain more datasets). We thus sought to address this limitation by performing secondary data analyses of published data sets.

Archival Analysis

The logic of the zero-point estimation strategy we applied to raw data can also be applied to published summaries of data, provided that researchers report (a) the mean and standard deviation for the IAT, (b) the mean and standard deviation for a relevant measure of bias they sought to predict with the IAT, and (c) the zero-order correlation between the two measures. With this information, one can estimate the intercept, its standard error, and its statistical significance in order to determine if a study suggests statistically significant right or left bias. However, without the actual raw data it is not possible to determine if a correction for such bias has consequential effects on the recategorization of study participants.

The largest challenge to this approach is the difficulty finding studies that report sufficient data to estimate an empirical zero point. As noted, it is rare to find descriptive statistics on discrimination criteria in IAT studies. We sought any instance in which this occurred by examining all studies of racial and ethnic bias included in a recent meta-analysis of the IAT (Oswald et al., 2013). We then searched for more recent papers by performing a search on PsycInfo and PLOS One for studies pairing the term *Implicit Association Test* or *Implicit Attitude* with terms *prejudice*, *discrimination*, or *bias*. We also inspected all articles reported in Hofmann, Gawronski, Gschwendner, Le, and Schmitt (2005), a meta-analysis of the correlation between IAT and explicit self-report measures. This provided us just three published reports that had a behavioral or judgmental criterion that we could examine.[8] These were as follows:

Biernat, Collins, Katzarska-Miller, and Thompson (2009), Study 1

Participants were 86 individuals who judged 44 Black and White targets in terms of academic ability on subjective or objec-

[8] Analyses by Ashburn-Nardo, Knowles, and Moneith (2003) that were described earlier were not reported in sufficient detail to include in this more formal analysis of slope intercepts.

Table 5
Frequency Distribution for Racial Bias Categories

Category	Standard	Corrected (0.30)
Amodio & Devine (2006), Study 2 (race IAT)		
Strong bias for Whites	—	—
Moderate bias for Whites	48%	—
Slight bias for Whites	39%	26%
None	10%	61%
Slight bias for Blacks	3%	10%
Moderate bias for Blacks	—	3%
Strong bias for Blacks	—	—
Generalized MOE = 18%		

Category	Standard	Corrected (0.28)
Amodio & Devine (2006), Study 2 (stereotype IAT)		
Strong stereotype bias for Whites	7%	—
Moderate stereotype bias for Whites	19%	3.2%
Slight stereotype bias for Whites	55%	12.9%
None	16%	61%
Slight stereotype bias for Blacks	3%	16.1%
Moderate stereotype bias for Blacks	—	3.2%
Strong stereotype bias for Blacks	—	3.2%
Generalized MOE = 18%		

Category	Standard	Corrected (274)
Heider & Skowronski (2007), Study 1		
Strong bias for Whites	33%	6%
Moderate bias for Whites	37%	9%
Slight bias for Whites	16%	14%
None	12%	35%
Slight bias for Blacks	1%	20%
Moderate bias for Blacks	—	12%
Strong bias for Blacks	—	4%
Generalized MOE = 8%		

Category	Standard	Corrected (162)
Heider & Skowronski (2007), Study 2		
Strong bias for Whites	38%	9%
Moderate bias for Whites	20%	16%
Slight bias for Whites	11%	15%
None	24%	18%
Slight bias for Blacks	6%	15%
Moderate bias for Blacks	2%	24%
Strong bias for Blacks	—	4%
Generalized MOE = 8%		

Category	Standard	Corrected (−0.30)
Ziegert & Hanges (2009; Climate for equality)		
Strong bias for Whites	23%	28%
Moderate bias for Whites	9%	8%
Slight bias for Whites	8%	8%
None	9%	9%
Slight bias for Blacks	13%	19%
Moderate bias for Blacks	11%	4%
Strong bias for Blacks	26%	25%
Generalized MOE = 7%		

Category	Standard	Corrected (−0.28)
Ziegert & Hanges (2005; Climate for racial bias)		
Strong bias for Whites	23%	25%
Moderate bias for Whites	4%	10%
Slight bias for Whites	15%	15%
None	21%	23%
Slight bias for Blacks	15%	10%
Moderate bias for Blacks	8%	8%
Strong bias for Blacks	15%	8%
Generalized MOE = 7%		

Note. Values in parentheses indicate the correction factor applied to the standard distribution. Generalized MOE = Margin of error based on a percent of 50 using 95% confidence intervals.

tive scales. The difference between objective and subjective evaluations was used to quantify stereotyping of group members as lower in academic competence, and racial prejudice was estimated by taking a Black minus White difference between shifting standards. A standard race IAT was used, with photos of White and Black faces as racial stimuli and with response measured on the D-score metric.

Greenwald, Smith et al. (2009)

As discussed earlier, these researchers studied an online sample of visitors to the Project Implicit website (https://implicit.harvard.edu) who expressed a preference for one of the two major Presidential candidates in the 2008 election. Participants completed a Brief IAT, where preferences for White versus Black faces was measured using the D-score metric. The criterion of interest here was a stated preference for Obama versus McCain. This study also included differences in explicit Likert and thermometer evaluations of White and Blacks.

Webb (2011)

Participants were 101White British undergraduate students who completed an IAT designed to measure their implicit preference for Whites over Asians by assessing the speed with which they could associate Asian versus Scottish names with pleasant versus unpleasant terms. The behavioral estimate of bias was based on whether individuals sought "advice" from an individual depicted on the computer as White versus Asian, with the assumption that Whites who are biased against Asians will be more interested and trusting of the advice given by a White relative to Asian advice giver.

Additional Studies With Explicit Attitude Data

In addition to these three studies that focused on behavior, we located seven papers that linked IAT measures of racial/ethnic bias to corresponding explicit difference scores (Blair, Judd, Havranek, & Steiner, 2010; Blair et al., 2013; Greenwald et al., 1998, Study 3; Perugini, O'Gorman, & Prestwich, 2007, Study 4; Rudman &

Ashmore, 2007, Studies 1 and 2; Sabin, Nosek, Greenwald, & Rivara, 2009; Uhlmann, Dasgupta, Elgueta, Greenwald, & Swanson, 2002). We included these papers in our analyses to be thorough, not to argue in favor of use of a zero explicit difference score as an unambiguous estimate of attitudinal neutrality.

Results

Table 6 presents zero-order correlations between the IAT and key outcomes in each of the studies, along with the relevant intercepts and margins of error. This shows that the intercepts were positive in all analyses examined. For the analyses that oriented around behavioral/judgment criteria, the intercept was statistically greater than zero in all three analyses, $p < .05$. For analyses that oriented around explicit difference scores, the intercept was positive in every instance and statistically greater zero $p < .05$, in 29 out of 31 analyses.

We performed a formal meta-analysis of the intercepts across all data sets that used the D scoring algorithm to document the average intercept value across the studies as well the variation in the intercept values. We used the random-effects robust estimation method by Hedges, Tipton, and Johnson (2010), which accommodates complex dependency structures for cases where multiple parameter estimates derived from the same sample of individuals are combined with estimates derived from independent samples. The average intercept was 0.23 ($z = 0.5.50$, $p < .001$, MOE = $\pm$ 0.09) and the estimated tau (standard deviation across studies) was 0.14, again yielding evidence of a right bias in IAT measures of prejudice.

General Discussion

The present research found evidence that the zero point of IAT measures designed to assess racial and ethnic prejudices does not map onto empirical estimates of behavioral neutrality. The evidence across the wide range of studies suggests that IAT measures of racial and ethnic prejudice are "right biased" in the sense that, on average, they place scores to the right of a behavioral zero point in a way that implies overestimation of biases in a population. Further, if one invokes the cut points for bias typically used in IAT research (and used on the Project Implicit website), they also tend to consistently overestimate the magnitude of behavioral bias.

Arbitrary Versus Meaningful Metrics

The discrepancies we found between the IAT zero point and behavioral zero points highlight the arbitrary nature of the metric generated by conventional IAT scoring algorithms. It thereby questions many of the suggested real-world, applied implications of the IAT distribution (Banaji & Greenwald, 2013). Based on this distribution, researchers have argued for the high prevalence of implicit racial bias in American society, with some arguing that modal distributions of IAT scores point to a need for reforms to legal codes (Levinson & Smith, 2012), judicial practices (e.g., Bennett, 2010; Saujani, 2003), and organizational policies (Ayres, 2001). Indeed, there is the prospect that organizations may in the future seek to use IATs as screening tools on the assumption that positive IAT scores are meaningful indicators of a propensity to discriminate (Tetlock & Mitchell, 2009; Tetlock, Mitchell, & Anastasopoulos, 2013). Our findings indicate the need for greater caution and for more applied research designed to determine the true meaning of the IAT metrics that are now widely employed in efforts to gauge implicit biases. Individuals who treat Blacks and Whites equally are likely to have positive, nonzero IAT scores that current practice would suggest implies the presence of behavioral bias in favor of Whites. This suggests caution in interpreting IAT distributions. This will be particularly true when using the IAT to make individual-level diagnoses as the margins of errors in the estimates in the table will be smaller than the standard error of individual prediction (see Blanton et al., 2009).

Of course, many other psychological metrics are arbitrary in character and this is not a problem unique to the IAT. Arbitrariness is usually not an issue in psychological research oriented around deductive theory tests. Psychological theories typically make (often complex) predictions about how one or more variables covary. Such predictions usually can be readily evaluated even with arbitrary metrics, as long as the function linking the true scores to the observed scores is linear (a common assumption made in psychological research). However, something more is needed if the psychological metric is used to make statements about prevalence or magnitude of a particular attribute (e.g., the degrees of bias in a given population) or if it is used to draw inferences about an individual's standing on that same attribute. If a psychological measure is used to screen for such qualities as workplace competencies, clinical disorders, or academic proficiency, the practitioner employing such scales needs to know how scores on the test map onto observable behaviors and the cut points that delineate meaningful shifts in expected behavior in applied contexts (see Kazdin, 2006). This knowledge typically comes after research critically examines the realities predicted for individuals with different scores. Such research is needed for the IAT.

In a powerful assessment of the social consequences of psychometric conventions, Messick (1995) argued that validity of an instrument must be measured not just in terms of how it is interpreted but also in terms of the effects these interpretations might have:

> As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion. . . . Indeed, validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use. . . . Therefore, it is fundamental that score validation is an empirical evaluation of the meaning and consequences of measurement (Messick, 1995, p. 742).

It is in this spirit that we have sought to better understand the meaning of the IAT metric. Many scholars have examined the distribution of scores on this inventory and seen a social problem that requires strong remedies. Our analyses suggest a more complicated story—one requiring greater caution. Resources for fighting discrimination are limited and must be deployed wisely. At least in the minds of courts and remedial agendas of legislatures, when psychological accounts of discrimination wax, societal-structural accounts are likely to wane (Banks & Ford, 2009). Our analysis suggests that before public priorities are shifted as a result of the IAT distribution, greater attention should be given to the real-world meaning of the metric on which these scores are located (and see Landy, 2008).

Table 6
Empirical Zero (Intercept) Estimates From Published Studies

	IAT	Criterion	*n*	*r*	Intercept	MOE
Behavioral/Judgment						
Webb (2011)	Scottish/Asian attitude (D Metric)	Advice Seeking	101	0.20	0.31*	0.21
Biernat, Collins, Katzarska-Miller, & Thompson (2009)	White/Black attitude (D Metric)	Shifting Standards	86	0.14	0.32*	0.05
Greenwald et al. (2009)	White/Black attitude (D Metric)	Obama/McCain Preference	1057	0.17	0.03*	0.01
Explicit evaluation						
Raw metric						
Uhlmann et al. (2002)	Blanco Latino/Moreno Latino attitude	Thermometer	59	0.30	120.54*	102.48
Log metric						
Greenwald, McGhee, & Schwartz (1998, Study 3)	White/Black attitude (Male names)	Thermometer	26	0.19	0.17*	0.05
Greenwald, McGhee, & Schwartz (1998, Study 3)	White/Black attitude (Male mames)	Likert	26	0.30	0.18*	0.04
Greenwald, McGhee, & Schwartz (1998, Study 3)	White/Black attitude (Female names)	Thermometer	26	0.19	0.14	0.08
Greenwald, McGhee, & Schwartz (1998, Study 3)	White/Black attitude (Female names)	Likert	26	0.30	0.15*	0.06
D Metric						
Blair et al. (2010)	White/Black attitude	Thermometer	203	0.23	0.22*	0.08
Blair et al. (2010)	White/Black attitude	Semantic differential	232	0.24	0.22*	0.10
Blair et al. (2010)	White/Latino attitude	Thermometer	214	0.26	0.20*	0.11
Blair et al. (2010)	White/Latino attitude	Semantic differential	233	0.19	0.21*	0.14
Blair et al. (2013)	White/Black attitude	Thermometer	345	0.28	0.24*	0.03
Blair et al. (2013)	White/Black attitude	Semantic differential	388	0.12	0.26*	0.06
Blair et al. (2013)	White/Latino attitude	Thermometer	343	0.27	0.30*	0.04
Blair et al. (2013)	White/Latino attitude	Semantic differential	388	0.13	0.33*	0.05
Greenwald et al. (2009)	White/Black attitude	Thermometer	1057	0.36	0.03*	0.01
Greenwald et al. (2009)	White/Black attitude	Likert	1057	0.25	0.01*	0.01
Rudman & Ashmore (2007, Study 1)	White/Black attitude	Thermometer	64	0.25	0.03*	0.01
Rudman & Ashmore (2007, Study 1)	White/Black stereotype	Thermometer	64	0.16	0.42*	0.07
Rudman & Ashmore (2007, Study 2)	Christian/Jewish stereotype	Thermometer	89	0.53	0.27*	0.05
Rudman & Ashmore (2007, Study 2)	White/Asian attitude	Thermometer	89	0.43	0.13*	0.04
Rudman & Ashmore (2007, Study 2)	White/Asian stereotype	Thermometer	89	0.28	0.11*	0.06
Rudman & Ashmore (2007, Study 2)	White/Black attitude	Thermometer	126	0.42	0.11	0.05
Rudman & Ashmore (2007, Study 2)	White/Black stereotype	Thermometer	126	0.27	0.31*	0.04
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	Chinese/Afro-Caribbean attitude	Personal competence	18	0.13	0.16*	0.03
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	Chinese/Afro-Caribbean attitude	Personal warmth	18	0.18	0.17	0.09
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	Chinese/Afro-Caribbean attitude	Social competence	18	0.14	0.19*	0.09
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	Chinese/Afro-Caribbean attitude	Social warmth	18	0.33	0.16	0.09
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	Chinese/Afro-Caribbean attitude	Personal competence	20	0.16	0.14	0.08
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	Chinese/Afro-Caribbean attitude	Personal warmth	20	0.18	0.28*	0.10
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	Chinese/Afro-Caribbean attitude	Social competence	20	0.18	0.31*	0.09
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	Chinese/Afro-Caribbean attitude	Social warmth	20	0.16	0.25*	0.11
Sabin et al. (2009, White Subsample)	White/Black attitude	Likert	1651	0.21	0.38*	0.01
Sabin et al. (2009, Black Subsample)	White/Black attitude	Likert	202	0.20	0.14*	0.03
Sabin et al. (2009, Hispanic Subsample)	White/Black attitude	Likert	114	0.24	0.35*	0.04
Sabin et al. (2009, Asian Subsample)	White/Black attitude	Likert	287	0.22	0.37*	0.03

* significant at $p < .05$.

Across-Study Variability in the Degree of Right-Bias and Future IAT Research

As evidenced in our meta-analysis of IAT D scores, there was variability in the observed "correction factors" for right bias across studies. We believe that no one correction factor will apply across all scenarios. Just as the reliability and validity of a scale can differ across facets of measurement (a fundamental tenet of Cronbach's generalizability theory), this also is likely to be the case for other metric properties of a scale. A scale that is reliable for college students, for example, may not be reliable for poor, inner city adolescents; a scale that is reliable when administered individually may not be reliable when administered in group settings. If researchers seek to use the IAT to make statements about implicit prejudice and its behavioral consequences, we believe they can and should routinely perform the types of regression analyses used in the current research. That is, they should regress the IAT onto the behavioral difference and examine the value (and margin of error) of the intercept. A positive intercept indicates that people who appeared behaviorally neutral in the study tended to have positive IAT scores. Such a result can then be taken into consideration when making statements about implicit prejudice scores. A researcher, for example, should be reluctant to attribute meaningful prejudice in a study where the typical IAT score is 0.28 and the intercept (indicating the IAT score where subjects in the study were behaving in a clearly unbiased way) is, say, 0.30.

The Nature of Current Scoring Conventions

Analyses also reveal additional challenges for the IAT scoring conventions that are driven by arbitrary "cut points" to categorize individuals in terms of their degree of bias. Initially, researchers based such categorization by applying Cohen's *d* to the raw score metric (see Blanton & Jaccard, 2006a), although they dropped this criteria when the new scoring algorithm yielded larger effect sizes (Blanton & Jaccard, 2008). In each case, however, the embrace of specific scores as meaningful, independent of any known behavioral implications of these scores, is problematic. Such an approach ignores where on the true dimension of bias that an observed scores maps. We suggest caution in the use of these cut points for categorizing individuals until defensible empirical standards for them can be established. This requires the kind of research we pursued in our analyses to make a metric nonarbitrary. (For a detailed analysis of this issue, see Blanton, Jaccard, & Burrows, 2014).

Limitations

One potential objection to our zero-point analysis is that it included explicit attitude differences as a criterion for zero-point confirmation. We acknowledged that it has often been argued that implicit and explicit attitudes are dissociated in memory and, as such, one might expect estimates of bias obtained from these two measures to diverge. Explicit attitude measures also are said to be more subject to social desirability bias, hence one should distrust self proclamations of racial neutrality. We agree that such limitations should be given weight, but our approach has precedence in research conducted by architects of the IAT (Greenwald et al., 2006), and we felt it was important to be inclusive in our analyses. Despite these reservations, it is noteworthy that analyses using explicit attitude criteria and those oriented around behavior converged on the same conclusion—that the IAT measures of racial and ethnic prejudice are right-biased.

A second possible limitation is that the behaviors used to calibrate the zero point of the IAT in the studies we examined are influenced by multiple determinants that may also affect metric evaluation. One might argue, for instance, that with discriminatory behavior participants are motivated to present themselves in a positive light, and so it is not so much that the IAT is "right biased" as it is that the behavioral criteria were "left biased" (such that participants' behavior suggested greater tolerance than they held internally; see Dovidio & Gaertner, 2010). To this we would reply that, if the IAT truly is to be used as an estimate of discriminatory tendencies, then inferences must be grounded in observable outcomes. It does not strike us as sufficient to state that the zero point is meaningful and non-arbitrary, simply because people with positive IAT scores might exhibit discriminatory behaviors that have not yet been empirically documented. At some point, the zero-point should be linked to actual, detectable forms of behavioral bias or individuals taking the IAT can be labeled as racially biased, simply by fiat.

Further, if such a critique is accurate, it suggests that the links between IAT scores and criteria need to be adjusted, based on the broader context in which the IAT might be applied (e.g., with adjustments varying across contexts as a function of social desirability concerns). This points to yet another difficulty of relying on the seeming face validity of the IAT metric. If the estimate of the empirical zero point for a given IAT score shifts dramatically from situation to situation or from behavior to behavior or from population to population (all of which seems possible; see Oswald et al., 2013), then this further limits the inferences one can make based on a given IAT score. As with other limitations, this possibility points to the need for more focused research that further maps the meaning of IAT scores — not simply in general or in the abstract — but in the very specific applied settings and the very specific populations for which test scores might be used to draw meaningful, real-world inferences. Our argument is not that our analyses reveal a definitive answer to the meaning of the IAT metric. Our argument is that there is a need for more rigorous research examining the meaning of the IAT metric, given its history of strong application (e.g., Banaji & Greenwald, 2013; cf., Landy, 2008). Until that is done, researchers should refrain from the types of strong statements often made regarding the assumed negative consequences that pervasive positive IAT scores might have in applied settings (e.g., Ayres, 2001; Bennett, 2010; Levinson & Smith, 2012; Saujani, 2003; see Banaji & Greenwald, 2013).

Conclusions

Our findings suggest that some area on the IAT distribution might reveal tendencies toward meaningful forms of racial and/or ethnic preference. But, if such is the case, our findings also link these biases with a smaller portion of the IAT distribution than is commonly argued. To date, one factor driving interest in the IAT has derived from the assumption that the IAT metric has uncovered individuals who are predisposed to commit prejudicial acts that they would rather not commit. The present research suggests a need to revisit this assumption.

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91,* 652–661. http://dx.doi.org/10.1037/0022-3514.91.4.652

Ashburn-Nardo, L., Knowles, M. L., & Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition, 21,* 61–87. http://dx.doi.org/10.1521/soco.21.1.61.21192

Ayres, I. (2001). *Pervasive prejudice? Unconventional evidence of race and gender discrimination*. Chicago, IL: University of Chicago Press.

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York, NY: Random House.

Banks, R. R., & Ford, R. T. (2009). (How) does unconscious bias matter?: Law, politics, and racial inequality. *Emory Law Journal, 58,* 1053–1122.

Bennett, M. W. (2010). Unraveling the Gordian knot of implicit bias in jury selection: The problems of judge-dominated voir dire, the failed promise of Batson, and proposed solutions. *Harvard Law & Policy Review, 4,* 149–171.

Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin, 35,* 16–28. http://dx.doi.org/10.1177/0146167208325195

Blair, I. V., Havranek, E. P., Price, D. W., Hanratty, R., Fairclough, D. L., Farley, T., . . . Steiner, J. F. (2013). Assessment of biases against Latinos and African Americans among primary care providers and community members. *American Journal of Public Health, 103,* 92–98. http://dx.doi.org/10.2105/AJPH.2012.300812

Blair, I. V., Judd, C. M., Havranek, E. P., & Steiner, J. F. (2010). Using community data to test the discriminant validity of ethnic/racial group IATs. *Zeitschrift Für Psychologie, 218,* 36–43.

Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41. http://dx.doi.org/10.1037/0003-066X.61.1.27

Blanton, H., & Jaccard, J. (2006b). Tests of multiplicative models in psychology: A case study using the unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 113,* 155–166. http://dx.doi.org/10.1037/0033-295X.113.1.155

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology, 34,* 277–297. http://dx.doi.org/10.1146/annurev.soc.33.040406.131632

Blanton, H., Jaccard, J., & Burrows, C. N. (2014). Implications of IAT D-transformation for psychological assessment. *Assessment*. Advance online publication.

Blanton, H., Jaccard, J., & Burrows, C. (in press). Implications of the IAT D-transformation for psychological assessment. *Assessment*. http://dx.doi.org/10.1177/1073191114551382

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal Of Applied Psychology, 94,* 567–582.

Blanton, H., & Mitchell, G. (2011). Reassessing the predictive validity of the IAT II: Reassessing the predictive validity of Heider & Skowronski (2007). *North American Journal of Psychology, 13,* 99–106.

Carmines, E. G., Sniderman, P. M., & Easter, B. C. (2011). On the meaning, measurement, and implications of racial resentment. *The Annals of the American Academy of Political and Social Science, 634,* 98–116. http://dx.doi.org/10.1177/0002716210387499

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

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton Trilogy revisited. *Personality and Social Psychology Bulletin, 21,* 1139–1150. http://dx.doi.org/10.1177/0146167295211002

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 2, pp. 1084–1121). Hoboken, NJ: Wiley.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668–689. http://dx.doi.org/10.1177/1094428110380467

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

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. http://dx.doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the implicit association test: Comment on Blanton and Jaccard (2006). *American Psychologist, 61,* 56–61. http://dx.doi.org/10.1037/0003-066X.61.1.56

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41. http://dx.doi.org/10.1037/a0015575

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy, 9,* 241–253. http://dx.doi.org/10.1111/j.1530-2415.2009.01195.x

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(5), 39–65.

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

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31,* 1369–1385. http://dx.doi.org/10.1177/0146167205275613

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations, 11,* 69–87. http://dx.doi.org/10.1177/1368430207084847

Katz, D., & Braly, K. W. (1935). Racial prejudice and racial stereotypes. *The Journal of Abnormal and Social Psychology,* Vol. 30, 175–193.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structure. *Journal of Personality and Social Psychology, 55,* 893–905. http://dx.doi.org/10.1037/0022-3514.55.6.893

Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61,* 42–49. http://dx.doi.org/10.1037/0003-066X.61.1.42

Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 379–392. http://dx.doi.org/10.1111/j.1754-9434.2008.00071.x

LeBel, E. (2011). The utility and feasibility of metric calibration for basic psychological research. (Doctoral dissertation). University of Western Ontario–Electronic Thesis and Dissertation Repository, Paper 174. Retrieved from http://ir.lib.uwo.ca/etd/174

Levinson, J. D., & Smith, R. J. (Eds.). (2012). *Implicit racial bias across the law*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511820595

Lord, F. M. (1952). *A theory of test scores* (Psychometric monograph No. 7). Richmond, VA: Psychometric Corporation.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, UK: Addison Wesley.

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. http://dx.doi.org/10.1006/jesp.2000.1470

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Nelson, S. (2008). Feeling Thermometer. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (p. 277). Thousand Oaks, CA: SAGE. http://dx.doi.org/10.4135/9781412963947.n183

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., . . . Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18,* 36–88.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171–192.

Perugini, M., O'Gorman, R., & Prestwich, A. (2007). An ontological test of the IAT: Self-activation can increase predictive validity. *Experimental Psychology, 54,* 134–147. http://dx.doi.org/10.1027/1618-3169.54.2.134

Petty, E. E., Fazio, R. H., & Briñol, P. (2008). *Attitudes: Insights from the new implicit measures*. New York, NY: Psychology Press.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations, 10,* 359–372. http://dx.doi.org/10.1177/1368430207078696

Sabin, J. A., Nosek, B. A., Greenwald, A. G., & Rivara, F. P. (2009). Physicians' implicit and explicit attitudes about race by MD race, ethnicity, and gender. *Journal of Health Care for the Poor and Underserved, 20,* 896–913. http://dx.doi.org/10.1353/hpu.0.0185

Saujani, R. M. (2003). The implicit association test: A measure of unconscious racism in legislative decision-making. *Michigan Journal of Race and Law, 8,* 395–423.

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

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51,* 1065–1071. http://dx.doi.org/10.1037/0003-066X.51.10.1065

Tetlock, P. E., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? In B. M. Staw & A. Brief (Eds.), *Research in Organizational Behavior, 29,* 3–38. http://dx.doi.org/10.1016/j.riob.2009.10.002

Tetlock, P. E., Mitchell, G., & Anastasopoulos, L. J. (2013). Detecting and punishing unconscious bias. *The Journal of Legal Studies, 42,* 83–110.

Uhlmann, E., Dasgupta, N., Elgueta, A., Greenwald, A. G., & Swanson, J. (2002). Subgroup prejudice based on skin color among Hispanics in the United States and Latin America. *Social Cognition, 20,* 198–226. http://dx.doi.org/10.1521/soco.20.3.198.21104

Webb, T. L. (2011). Advice-taking as an unobtrusive measure of prejudice. *Behavior Research Methods, 43,* 953–963. http://dx.doi.org/10.3758/s13428-011-0122-8

Wittenbrink, B., & Schwarz, N. (2007). *Implicit measures of attitudes*. New York, NY: Guilford 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. http://dx.doi.org/10.1037/0021-9010.90.3.553

Ziegert, J. C., & Hanges, P. J. (2009). Strong rebuttal for weak criticisms: Reply to Blanton et al. (2009). *Journal of Applied Psychology, 94,* 590–597. http://dx.doi.org/10.1037/a0014661

Received September 22, 2013
Revision received September 29, 2014
Accepted October 3, 2014 ■

Correction to Blanton et al. (2015)

In the article "Toward a Meaningful Metric of Implicit Prejudice," by Hart Blanton, James Jaccard, Erin Strauts, Gregory Mitchell, and Philip E. Tetlock (*Journal of Applied Psychology*, Advance online publication. January 19, 2015. http://dx.doi.org/10.1037/a0038379), there are errors in some of the values listed in Table 6 that do not alter any of the conclusions or substantive statements in the original article. The corrected portion of Table 6 follows. The positive intercepts in this table represent the estimated IAT score when the criterion has a value of zero (suggesting attitudinal neutrality), except in the equation examining voter preference in Greenwald et al. (2009), where the intercept estimated the IAT score of Obama voters.

Table 6
Empirical Zero (Intercept) Estimates From Published Studies

	n	*r*	Intercept	MOE
Behavioral/Judgment				
Webb (2011)	101	0.20	0.31*	0.07
Biernat, Collins, Katzarska-Miller, & Thompson (2009)	86	0.14	0.32*	0.11
Greenwald et al. (2009)	1057	0.17	0.03*	0.03
Explicit evaluation				
Raw metric				
Uhlmann et al. (2002)	62	0.30	120.54*	44.29
Log metric				
Greenwald, McGhee, & Schwartz (1998, Study 3)	26	0.19	0.17*	0.07
Greenwald, McGhee, & Schwartz (1998, Study 3)	26	0.30	0.18*	0.05
Greenwald, McGhee, & Schwartz (1998, Study 3)	26	0.07	0.14*	0.07
Greenwald, McGhee, & Schwartz (1998, Study 3)	26	0.11	0.15*	0.06
D Metric				
Blair et al. (2010)	203	0.23	0.22*	0.05
Blair et al. (2010)	232	0.24	0.22*	0.04
Blair et al. (2010)	214	0.26	0.20*	0.06
Blair et al. (2010)	233	0.19	0.21*	0.06
Blair et al. (2013)	345	0.28	0.24*	0.04
Blair et al. (2013)	388	0.12	0.26*	0.04
Blair et al. (2013)	343	0.27	0.30*	0.04
Blair et al. (2013)	388	0.13	0.33*	0.04
Greenwald et al. (2009)	1057	0.36	0.03*	0.02
Greenwald et al. (2009)	1057	0.30	0.01	0.03
Rudman & Ashmore (2007, Study 1)	64	0.25	0.42*	0.14
Rudman & Ashmore (2007, Study 1)	64	0.16	0.27*	0.09
Rudman & Ashmore (2007, Study 2)	89	0.53	0.10	0.10
Rudman & Ashmore (2007, Study 2)	89	0.43	0.11	0.15
Rudman & Ashmore (2007, Study 2)	89	0.28	0.11*	0.11
Rudman & Ashmore (2007, Study 2)	126	0.42	0.31*	0.08
Rudman & Ashmore (2007, Study 2)	126	0.27	0.16*	0.08
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	18	0.13	0.17	0.21
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	18	0.08	0.19	0.19
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	18	0.14	0.16	0.23
Perugini, O'Gorman, & Prestwich (2007, Study 4 Control)	18	0.33	0.14	0.18
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	20	0.16	0.28*	0.21
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	20	0.18	0.31*	0.18
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	20	0.18	0.25	0.26
Perugini, O'Gorman, & Prestwich (2007, Study 4 Experimental)	20	0.13	0.29*	0.22
Sabin et al. (2009, White Subsample)	1651	0.21	0.38*	0.02
Sabin et al. (2009, Black Subsample)	202	0.20	0.14*	0.09
Sabin et al. (2009, Hispanic Subsample)	114	0.24	0.36*	0.09
Sabin et al. (2009, Asian Subsample)	287	0.22	0.35*	0.05

* significant at $p < .05$.

http://dx.doi.org/10.1037/a0039215