


# Implications of the Implicit Association Test D-Transformation for Psychological Assessment

Assessment  
2015, Vol. 22(4) 429–440  
© The Author(s) 2014  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1073191114551382  
asm.sagepub.com  


Hart Blanton<sup>1</sup>, James Jaccard<sup>2</sup>, and Christopher N. Burrows<sup>1</sup>

## Abstract

Psychometricians strive to eliminate random error from their psychological inventories. When random error affecting tests is diminished, tests more accurately characterize people on the psychological dimension of interest. We document an unusual property of the scoring algorithm for a measure used to assess a wide range of psychological states. The “D-score” algorithm for coding the Implicit Association Test (IAT) requires the presence of random noise in order to obtain variability. Without consequential degrees of random noise, all individuals receive extreme scores. We present results from an algebraic proof, a computer simulation, and an online survey of implicit racial attitudes to show how trial error can bias IAT assessments. We argue as a result that the D-score algorithm should not be used for formal assessment purposes, and we offer an alternative to this approach based on multiple regression. Our critique focuses primarily on the IAT designed to measure unconscious racial attitudes, but it applies to any IAT developed to provide psychological assessments within clinical, organizational, and developmental branches of psychology—and in any other field where the IAT might be used.

## Keywords

psychometrics, Implicit Association Test, general processing speed, assessment

It is likely that more people have self-administered the Implicit Association Test (IAT) than any other modern psychological inventory. In just 15 years, the IAT has been delivered millions of times via a website hosted by Harvard University (<https://implicit.harvard.edu/implicit/>). The most popular and promoted use of this test is to diagnose hidden racial, ethnic, and sexual biases (Nosek, Banaji, & Greenwald, 2002; cf., Blanton & Jaccard, 2006a, 2006b). The modal response pattern on such measures suggests that the majority of respondents harbor moderate to high levels of in-group bias. For the IAT designed to measure anti-Black attitudes, for instance, the majority of White Americans who take this test are told they have an “automatic preference” for White people over Black people, with fully 27% receiving feedback that they have a “strong” preference in this direction (Table 1).

Because such psychological feedback can promote insight into hidden psychological tendencies, the IAT is often promoted as a useful assessment tool in college textbooks and courses (Plous, 2003; Whitley & Kite, 2009), sensitivity workshops (Pendry, Driscoll, & Field, 2007), and through the news media (Vendatam, 2005) and popular press (Banaji & Greenwald, 2013; Gladwell, 2005). IAT assessments have also been promoted within clinical psychology (Roefs et al., 2011; Rüsche et al., 2011; Schreiber,

Bohn, Aderka, Stangier, & Steil, 2012), organizational psychology (Rudman, 2008), developmental psychology (Rosen, Milich, & Harris, 2007), and in the legal domain where scholars have pointed to the potential applicability of IAT assessments for legal decisions (Greenwald, 2006; Scheck, 2004). Reflecting this enthusiasm, the current IAT webpage hosts variants of the measure that provide psychological feedback related not just to intergroup biases but also to depression, alcoholism, therapeutic preferences, disordered eating, anxiety, and self-esteem.<sup>1</sup>

Given the widespread and growing use of this assessment tool, we take a closer look at the scoring algorithm used to diagnose psychological states. We put aside concerns about the validity of the measure itself (see Blanton et al., 2006) and focus only on the scoring algorithm common to all IAT measures that currently are and in the future might be used to diagnose psychological states (Greenwald, Nosek, & Banaji, 2003). We focus attention on the IAT

<sup>1</sup>University of Connecticut, Storrs, CT, USA

<sup>2</sup>New York University, New York, NY, USA

## Corresponding Author:

Hart Blanton, Department of Psychology, University of Connecticut, Storrs, CT 06269-1020 USA.

Email: [hart.blanton@uconn.edu](mailto:hart.blanton@uconn.edu)

**Table 1.** Percentage of Respondents Receiving Different Psychological Labels.

Psychological label	Proportions (%)	
Strong automatic preference for White people compared with Black people	27	} 70
Moderate automatic preference for White people compared with Black people	27	
Slight automatic preference for White people compared with Black people	16	
Little or no automatic preference between White and Black people	17	
Slight automatic preference for Black people compared with White people	6	} 12
Moderate automatic preference for Black people relative with White people	4	
Strong automatic preference for Black people compared with White people	2	

Note. Psychological labels given to respondents who take the race Implicit Association Test (IAT) at <https://implicit.harvard.edu/implicit/>. These frequencies were collected from that site in March 2014.

designed to measure hidden anti-Black racial prejudices, but our analysis applies to all IAT measures.

## Measuring Implicit Racial Attitudes

People vary in their degrees of prejudice. A White person who would not hire, trust, or even help a Black person might be viewed by most people as “strongly prejudiced,” whereas one who would do each of these acts—but after a moment’s hesitation—might be viewed as “moderately prejudiced.” A useful and valid self-report measure of racial bias should (a) sort individuals along the underlying continuum and (b) identify the amount of bias that a given individual possesses with respect to that continuum. Of course, no single item or question on a bias measure will be perfect. Random error affects the answers individuals give to each item on a questionnaire. It is out of appreciation of this psychometric reality that researchers and practitioners take steps to minimize the influence of item error on a person’s overall test score. Most typically, they do this by aggregating responses across multiple items. The logic is that item-level error for specific questions cancels across items in the computation of the total score (Rushton, Brainerd, & Pressley, 1983).

Much like a self-report measure with multiple questions, the IAT uses multiple trials that are combined to generate a single, aggregated test score. Each “test item” in an IAT is presented as a stimulus that requires a response. The task for the respondent is to correctly classify each stimulus into one of two different categories. Two blocks of stimuli trials are presented to each respondent, one block that is meant to be easy for someone who has a high standing on the implicit construct of interest and one block that is meant to be difficult

for someone who has high standing on the construct of interest. The speeds with which these classifications are made are then combined by calculating a mean response latency for each block (typically comprising between 20 and 60 trials) and a difference score is calculated between the block means.

For the typical IAT designed to measure anti-Black racial prejudice, the stimuli that respondents must categorize are images of Black versus White faces and words that are positively or negatively valenced. A “compatible response latency” (CRL) is the average latency for the block of trials designed to be easy for people who prefer Whites to Blacks; an “incompatible response latency” (IRL) is the average latency from the block of trials designed to be difficult for people who prefer Whites to Blacks. All IAT have an IRL block and a CRL block and the mean scores for these blocks are differenced,

$$IAT_{RAW} = IRL - CRL \quad (1)$$

where  $IAT_{RAW}$  is the estimate of the IAT effect computed in a raw (millisecond) response metric, IRL is the mean response across items for the incompatible trials, and CRL is the mean response across items for the compatible response trials.

As with any measure, a researcher should expect to observe some degree of item-level error (or what we can term *trial error* in this case) for stimuli presented within the IRL block and in the CRL block. Here, trial error refers to the “nonsystematic error” that causes an individual to respond faster or slower to different trials within a given block, independent of the construct being assessed. Trial error might occur because a particular stimulus on a given trial is unusually attention grabbing or because of momentary distractions in the testing environment. Both IRL and CRL are computed by aggregating across individual trials, thereby negating this error through a process of cancelation. Equation (1) can be viewed as a form of averaging across all items employing reverse scoring because subtracting CRL from IRL is algebraically equivalent to reverse-scoring CRL and averaging it with IRL. As such, Equation (1) reflects a common method of dealing with the problem of item-level error, namely, aggregation. All other factors being equal, a person with a large  $IAT_{RAW}$  difference score is viewed as being more racially biased by the IAT architects.

## $IAT_{RAW}$ Assessment Strategy

On the IAT website, researchers originally used  $IAT_{RAW}$  to provide respondents with feedback about the magnitude of their implicit racial bias by using Cohen’s (1988) criteria for categorizing small, medium, and large effect sizes ( $d$  scores for  $IAT_{RAW}$  of 0.20, 0.50, and 0.80; Anthony Greenwald, personal communication, August 2002). They did this by

measuring the difference of the mean reaction time for the two types of trials, per Equation 1, then converting this score into d-score units, using Cohen's classic formula. They used a single fixed estimate of the group-level, between-subjects standard deviation of  $IAT_{RAW}$ . For instance, if data suggested that the typical standard deviation for the IAT difference score computed from Equation (1) was 200 milliseconds for a given population, then the estimate of implicit bias is computed using the formula:

$$d = IAT_{RAW} / 200 \quad (2)$$

Researchers then labeled respondents as having "no," "slight," "moderate," or "strong" implicit biases, based on whether or not they exceeded Cohen's criteria for labeling experimental effect sizes as "small," "medium," and "large."

### *IAT<sub>D</sub> Scoring Algorithm*

After 2003, IAT researchers adopted a new scoring algorithm (Greenwald et al., 2003) that placed the IAT on a new metric (called a "D score" as opposed to a "d score"). The reason for the change was to address confounds associated with traditional IAT scores. For instance, it was known that  $IAT_{RAW}$ , at times, is confounded with a person's speed at performing cognitive tasks in general (McFarland & Crouch, 2002). When this dynamic occurs, the faster a respondent's reaction time while taking the IAT, the smaller the  $IAT_{RAW}$  difference score tends to be and thus the lower the estimate of implicit bias. The artifact is reflected in a positive correlation of extremity of IAT effects with response latency (Greenwald et al., 2003, p. 200). Greenwald et al. (2003) examined eight different "candidate" algorithms to determine which one would reduce the unwanted correlation between IAT scores and average response latencies, while at the same time maximizing other associations that were desired (e.g., the magnitude of correlation with explicit measures of the same construct). The D-scoring algorithm was chosen because it tended to have lower correlations with average response latency than other algorithms considered.<sup>2</sup> It is defined as

$$IAT_D = IAT_{RAW} / SD_{WI} \quad (3)$$

where  $SD_{WI}$  is a within-individual standard deviation of response latencies calculated across the compatible and incompatible items/trials.<sup>3</sup>  $IAT_D$  scores range from -2.0 to +2.0, with more extreme scores interpreted as revealing more extreme standing on the psychological construct of interest (e.g., more extreme racial bias) and with scores closer to zero interpreted as revealing less extreme standing (e.g., racial neutrality).

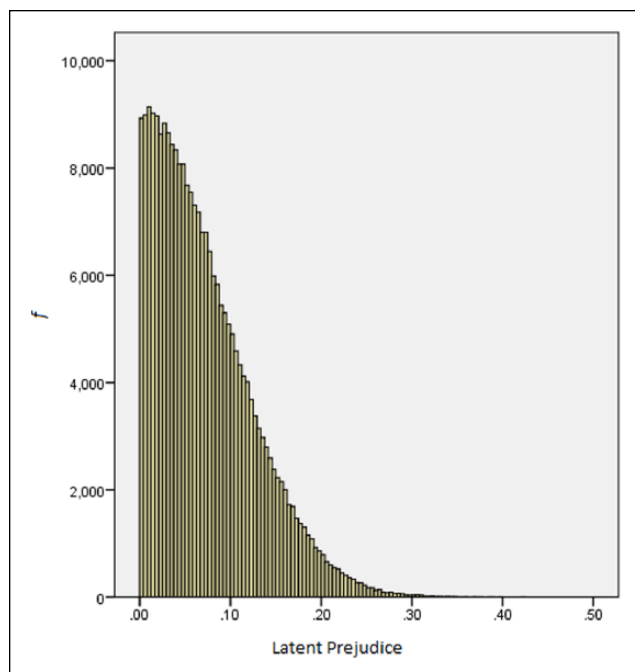
Importantly, with the new D score, IAT researchers also asserted new criteria for classifying people into categories of "no," "slight," "moderate," and "strong" automatic preferences for Blacks or Whites. Although no explicit rationale was ever published, cut-values centered around the D values of 0.15 ("slight preference"), 0.35 ("moderate preference"), and 0.64 ("strong preference") were adopted. These same cut-points are used in the other IAT measures used to assess psychological states (Anthony Greenwald, personal communication, August 2002).

### *Effect of Trial Error on IAT Assessments*

The move to Equation (3) changed the way in which trial error influences psychological assessment, as the  $SD_{WI}$  score was formally integrated into the IAT scoring algorithm. This change unwittingly produced a situation in which a reduction in trial error (typically something a researcher or practitioner would view as desirable) now results in more extreme bias scores, everything else being equal. This is true for any IAT measuring any psychological construct, if the D algorithm is used.

Consider, for example, a researcher who administers the race IAT by having people take the test online (as is currently done on the main IAT website). Online administration prevents the researcher from exerting control over the testing environment. Some respondents might take the test when they are tired (either late at night or early in the morning), some might take the test while multitasking, and some might take the test after having had a few alcoholic drinks. The result of these demands and distractions is that some respondents will exhibit higher levels of trial error, reflecting the fact that their attention is unstable from one experimental trial to the next. By contrast, another researcher administers the same test in a research laboratory. Each respondent in the study takes the test in a study cubicle, situated in a quiet research laboratory where outside distractions are reduced and variability in cognitive functioning is minimized (by ensuring that the test is only administered in the middle of the day, when respondents might be better able and motivated to concentrate). Because trial error is in the denominator of Equation (3) (within  $SD_{WI}$ ), the second sample should appear to be more racially biased than the first sample, even if there are no differences in their racial attitudes. This is because as trial error is reduced,  $IAT_D$  scores homogenize toward the extremes. Indeed, it can be shown mathematically that if trial error is reduced to zero, every individual in the study will receive an absolute implicit bias score at its theoretical extreme (2.0) and will be characterized as strongly biased (and see Appendix for algebraic proof).

Of course, trial error of zero should not be expected in practice, and so extreme scores will not be generated by this



**Figure 1.** Distribution of true bias in simulated population.

factor alone. Nevertheless, the built-in dynamic of reduced trial error leading to increased characterizations of bias is psychometrically unsatisfying. To document the effects of incorporating  $SD_{wi}$  into the  $IAT_D$  score across differing degrees of trial error, we performed a computer simulation that varied the amount of trial error.

## Computer Simulation

Using the same  $-2$  to  $+2$  scale for the true scores used for the  $IAT_D$  score, we simulated a population of individuals using the upper half of a normal distribution in which 90% of the scores were between 0 and 0.15 (which would be declared as lacking implicit bias by  $IAT_D$  standards) and where 10% of the scores were greater than 0.15 (which would be declared as having at least “slight bias” by  $IAT_D$  standards). Less than 1% of the population had scores greater than 0.35 (which would be declared as having at least “moderate bias” by  $IAT_D$  standards). Figure 1 presents a histogram of the distribution of true bias scores for the simulated population ( $N = 250,000$ ). Given this distribution, characterizations by the observed  $IAT_D$  scores that we create for this population should predominately indicate lack of implicit bias. We represent the true score values by  $T$  and the observed scores, generated using the methods described below, as  $X$ .

## Method

We first simulated 20 compatible and 20 incompatible IAT trials for each individual in the population. For each trial

and for each individual, we first generated a response latency (in milliseconds) for the compatible trial using the equation

$$CRL_{ij} = 800 + -400T_i$$

where  $CRL_{ij}$  is the compatible response latency for individual  $i$  on trial  $j$  and  $T_i$  is the individual's true implicit bias score. For the incompatible trials, we used a comparable equation but changed the sign of the slope:

$$IRL_{ij} = 800 + 400T_i$$

where  $IRL_{ij}$  is the incompatible latency for individual  $i$  on trial  $j$ . With no random error affecting any of the scores at this juncture, all 20 compatible trials for an individual had the identical score, and this was also true for the 20 incompatible trials (although the incompatible trial scores were different from those for the compatible trials by virtue of the different slope values in the above equations). The mean of the incompatible trials for an individual minus the mean of the compatible trials for that individual (i.e., the  $IAT_{RAW}$  computed via Equation 1) had a correlation of 1.0 with the true implicit bias scores across individuals, again because no random error for within-trial variability had yet to be introduced. The grand mean of the response latencies across all trials (both compatible and incompatible) and across all individuals is 800 milliseconds (which is dictated by the above chosen intercepts and slopes). If one regresses the “observed” IAT difference scores onto the true bias scores at this juncture (per Equation 1), the intercept is 0, the unstandardized slope is 800, and the reliability is 1.00 (i.e., there is no error variance). All these values are expected, given the above equations.

We next introduced within-block random error for each trial by adding a random error score to each trial latency, where the error score was randomly selected from a normal distribution with a mean of zero and a standard deviation  $\sigma_{TE}$ . We investigated 10 different values of  $\sigma_{TE}$ . At the lowest level of trial error,  $\sigma_{TE}$  was set to a value that was 1% of the grand mean latency (i.e., 8 milliseconds given a grand mean latency of 800 milliseconds). At the highest level of noise,  $\sigma_{TE}$  was set to 70% of the grand mean latency (i.e., 560 milliseconds given a grand mean latency of 800 milliseconds). Independent error values were used for each trial.

For this simulation, the function relating the observed latency differences relative to true bias is linear. The only psychometric complication that is introduced in the simulation is the presence of within-individual, within-block random error. Without trial error, there is perfect correspondence between true prejudice and observed prejudice as reflected by  $IAT_{RAW}$ . The simulation allows us to test the ability of the  $IAT_D$  score to accurately characterize implicit bias within a population where the true bias levels of

**Table 2.** True and Observed Distributions for Implicit Prejudice for 10 Levels of Random Trial Error.

		Trial error									
		1%	5%	7%	10%	13%	16%	20%	30%	50%	70%
Intraclass correlation		0.96	0.48	0.32	0.21	0.15	0.12	0.10	0.07	0.06	0.05
Reliability for IAT <sub>RAW</sub>		0.99	0.92	0.88	0.75	0.64	0.54	0.43	0.25	0.11	0.06
Reliability for IAT <sub>D</sub>		0.39	0.73	0.73	0.66	0.58	0.50	0.40	0.24	0.11	0.06
Within-block trial correlation		0.88	0.23	0.13	0.07	0.04	0.03	0.02	0.01	0.002	0.001
Bias category	True %	Observed percentage									
Strong pro-Black	0.0	0.0	0.1	0.1	0.2	0.3	0.3	0.4	0.5	0.8	1.0
Moderate pro-Black	0.0	0.3	0.8	1.2	1.7	2.2	2.6	3.2	4.5	6.2	7.4
Slight pro-Black	0.0	0.4	1.8	2.6	3.7	4.7	5.6	6.9	9.1	12.2	13.7
No bias	90.0	1.3	6.3	8.7	12.2	15.3	18.0	21.0	26.5	31.3	33.2
Slight pro-White	9.8	1.4	6.7	9.1	12.4	15.0	17.1	19.1	21.7	22.5	21.8
Moderate pro-White	0.10	2.6	12.6	16.9	21.5	24.5	26.3	27.1	25.4	20.9	18.4
Strong pro-White	0.0	94.1	71.5	61.3	48.3	38.0	30.0	22.4	12.4	6.1	4.4

Note. Trial error sets the within-block standard deviation to the shown percent of grand mean latency across all trials and all individuals. Intraclass correlation is the proportion of total variation due to between-subject variation in latencies across compatible and incompatible trials. Reliability estimates for IAT<sub>RAW</sub> and IAT<sub>D</sub> reflect the proportion of variation in each IAT estimate that reflects true prejudice. Within-block trial correlation is the average within-block correlation between individual trials across subjects. True percentage values represent the true percentage of individuals in each bias category and Observed percentage represents the percentage of individuals categorized in each of these same categories using current IAT<sub>D</sub> scoring conventions, with variability occurring across columns as a function of differences in trial error.

participants are known and where differing degrees of trial error are systematically manipulated.

## Results

For both the IAT<sub>D</sub> and IAT<sub>RAW</sub>, Table 2 presents for each level of trial error modeled the reliability of the observed measure, the proportion of total variation in the latencies that is due to between-subject variation in block mean latencies (i.e., the intraclass correlation) and the average within-block correlation between individual trials across subjects (which is analogous to the average item correlation on traditional scales). Note, as one would expect, the reliability of IAT<sub>RAW</sub> increases as trial error decreases. This also tends to be true of the IAT<sub>D</sub> scores, but this is only up to a point. At about 7% trial error, the reliability begins to level off and at trial errors less than 7% (i.e., where there is *less* noise), the reliability of the IAT<sub>D</sub> decreases. At very low levels of trial error (1%), the reliability is no longer at levels acceptable for psychological assessments (0.39). This reduction in reliability occurs because as trial error approaches zero (the lower extreme of reliability), the observed values of IAT<sub>D</sub> approach the limit of the possible range of score; that is, a value of 2.0 (see the appendix for mathematical details). This points to another unintended effect of the scoring algorithm: If researchers are highly successful at reducing nonsystematic error at the level of trials, the IAT<sub>D</sub> will become unreliable.

Table 2 also presents the percentage of individuals occurring in the bias categories defined by the IAT<sub>D</sub> criteria used by the IAT website for each of the levels of trial error. These

can be contrasted with the classification of the true bias scores, also shown in Table 2. In every case, the percentage of people characterized as being biased in the population is substantially misrepresented by the IAT<sub>D</sub>. For example, in the 10% trial error condition (where the reliability of IAT<sub>D</sub> is 0.66), almost half of the individuals are characterized as having a strong implicit bias favoring Whites over Blacks when, in fact, no one in the population has this level of bias; about 70% of the population is characterized as having moderate or strong implicit bias favoring Whites over Blacks when the percentage of people for whom this truly is the case is less than 10%. These estimates offer new perspectives on the bias characterizations currently provided by the IAT website. The reliability of the race IAT in extant literature tends to be in the 0.50 range (Blanton & Jaccard, 2006a, 2006b), a value that the current simulation suggests can cause about 30% of a dominantly nonbiased population to be mischaracterized as having extreme pro-White bias and 43% to be characterized as having slight or moderate bias (see Table 2).

## Discussion

The conclusion from the simulation results is clear: As greater degrees of random noise are introduced into a participant's performance on IAT trials, the result is lowered estimates of implicit bias, everything else being equal. At reliability levels of .70 or greater, which is a minimum standard that many researchers set for adequately reliable measurement, the distortions of the IAT<sub>D</sub> are considerable when

applied to a fundamentally unprejudiced population. Interestingly, when the normally distributed trial error dominates responses for a population that is fundamentally unbiased, the resulting observed distribution of implicit scores also is normally distributed because the observed scores mainly reflect the error scores (see, e.g., the 70% trial error condition in Table 2). In this case, about 45% of the population is characterized as being biased in favor of Whites.

The simulation suggests that the incorporation of  $SD_{WI}$  into the scoring algorithm can lead to grossly inaccurate estimates of extreme implicit bias when trial error is low. In contrast, when trial error is large, implicit bias estimates are essentially distributed in accord with the random noise distribution (in this case, a normal distribution). The simulation thus documents that, the more care a researcher takes to ensure that extraneous random factors do not affect a person's performance on a given trial in the IAT (e.g., by minimizing room noise, interruptions, distractions, or causes of distress and anxiety, or by conducting pilot testing of experimental stimuli to ensure they are equally good exemplars for use in a test), the greater the degree of bias that will be mistakenly attributed. By the same token, the more careless a researcher becomes, the more random noise will dominate the observed scores, which can produce the appearance of extreme implicit bias in a nonbiased population. The simulation thus suggests that trial error is a nontrivial factor in determining the extent to which an individual is characterized as being biased.

## Online Study

The IAT website, which is used to provide feedback about implicit bias or implicit preferences to millions of people, uses a relatively "noisy" environment to administer the IAT, in that it relies on people taking the IAT online. Online IAT data also have been used in numerous IAT publications to document the prevalence of different implicit biases in the United States and other countries (e.g., Nosek et al., 2007; Schmidt & Nosek, 2010). We therefore conducted an online race IAT assessment to better document the dynamics of the  $SD_{WI}$  parameter and the  $IAT_D$  score in such situations. The study examined the relationship between  $IAT_{RAW}$  and  $IAT_D$ , general processing speed when assessed by independent methods (see Note 2), and the results of classifying people into the categories of racial bias using the different criteria adopted by the IAT website over the years.

## Participants

Respondents were recruited via one of two methods. One group was recruited for \$1 via Amazon's Mechanical Turk, an Internet crowd-sourcing site that is used to connect interested participants with psychology researchers for pay.

Evidence suggests that experiments conducted on Mechanical Turk replicate experimental effects obtained in laboratory settings, but they do so with samples that are more diverse than college samples (Buhrmester, Kwang, & Gosling, 2011). The second group was a group of volunteers who participated in response to a request to complete an assessment of implicit racial attitudes. This recruitment method bears similarity to the IAT website, where individuals come to the website because they are intrinsically interested in implicit bias. Consistent with past research, both groups tended to show an "IAT effect" toward automatic bias for Whites (i.e., quicker response times responding to the compatible as opposed to the incompatible trials), and there were no observed differences in the magnitude of this IAT effect across these two groups.

The total sample was 658 participants, after removing 62 respondents who initiated but did not complete the IAT task, and 31 who had more than 10% of trials that were excessively fast (<300 milliseconds) or slow (>3,000 milliseconds), suggesting the IAT task was not taken seriously. The sample ranged in age from 18 to 66 years ( $M = 30$  years,  $SD = 9.87$ ), with the majority being male (67%) and White (69%).

## Method

Participants were told the purpose of the study was to collect data on their explicit and implicit racial attitudes (as is the case with the IAT website). They first completed an explicit measure designed to assess racial prejudice against African Americans based on the modern racism scale (McConahay, 1986). Modern racism is characterized by beliefs that racism is not a continuing problem, that African Americans should put forth effort to overcome their situation in society without special assistance, and that African Americans are too demanding and have received more than they deserve. Respondents then were administered the IAT designed to measure preferences for people who are Black versus White (Greenwald, McGhee & Schwartz, 1998). This task began by having participants learn to correctly categorize a set of words chosen to represent "Good" (e.g., wonderful, marvelous) versus "Bad" (e.g., terrible, tragic) via one of two key presses on the computer keyboard. Participants next performed the same task to learn to correctly discriminate a set of Black versus White faces. These trials are hereafter referred to as learning trials. After this learning period, participants were presented with a set of faces and words for 24 trials of the "compatible task" (in which they categorized stimuli as "Bad or Black" or "Good and White") and 24 trials of the "incompatible task" (in which they categorized stimuli as "Good or Black" or "Bad and White"), following the standard ordering of blocks (Greenwald et al., 1998).

Indices of the following constructs were obtained from the data: (a) (explicit) modern racism, (b) implicit racism

**Table 3.** Correlation Matrix for Online Study.

	IAT <sub>RAW</sub>	IAT <sub>D</sub>	MR	GPS	SD <sub>WI</sub>
IAT <sub>D</sub>	0.95**				
MR	0.14**	0.14**			
GPS	0.03	-0.04	0.01		
SD <sub>WI</sub>	0.32**	0.14**	0.08*	0.38**	
M	136.56	0.33	3.18	851.26	404.16
SD	161.90	0.36	1.29	149.65	102.33

Note. IAT<sub>RAW</sub> and IAT<sub>D</sub> are estimates of implicit racial bias, obtained using Equations 1 and 3, respectively. IAT = Implicit Association Test. MR = modern racism. GPS = average response latency for all learning trials of the IAT (general processing speed). SD<sub>WI</sub> = trial standard deviation across all trials measured within individuals.

operationalized as IAT<sub>RAW</sub> using Equation (1), (c) SD<sub>WI</sub>, which is the estimate of trial error required to implement the newer D algorithm, and (d) implicit racism operationalized as IAT<sub>D</sub> using Equation (3). We also generated an estimate of each individual's overall speed at performing categorization tasks, computed as the average response for the learning trials (see Blanton, Jaccard, Gonzales, & Christie, 2006). We refer to this construct as general processing speed (GPS). A primary reason that Greenwald et al. (2003) adopted IAT<sub>D</sub> scoring as an alternative to IAT<sub>RAW</sub> was to minimize the correlation between IAT estimates and GPS (and see Note 2).

## Results

**Trial Error in the IAT.** For the IAT task, there were 48 trials (half incompatible and half compatible). We used standard multilevel modeling (Raudenbush & Bryke, 2002) to estimate how much of the total variability across all items and all respondents was due to between-subject differences in (mean) trial type, analogous to the intraclass correlations reported in the simulation study. The intraclass correlation was 0.21, indicating that 79% of the variation was trial error and 21% reflected between-subject differences in block means across trials. The coefficient alpha for the compatible trials was 0.87, and for the incompatible trials it was 0.82. The grand mean across all trials was 960 and the standard deviation was 450, yielding a coefficient of variation, expressed as a percentage of 46.9%. The median within-block correlation between trial latencies was 0.31 for the compatible trials and 0.26 for the incompatible trials.

**Descriptive Statistics.** The means, standards deviations, and intercorrelations for each of the key variables are shown in Table 3. There are a number of noteworthy associations. First, IAT<sub>RAW</sub> and IAT<sub>D</sub> are so strongly associated with one another that they can be considered redundant in terms of ordering individuals on their respective metrics,  $r(656) = 0.95$ ,  $p < .001$ . So strong is this association, in fact, that

despite claims that the new algorithm might promote more valid inferences when used to test psychological theories (Greenwald et al., 2003), the data suggest that one typically should expect the two scores to be interchangeable in theory tests that have a goal of using an implicit attitude score to predict criteria or where the IAT value is an outcome. In support of this interpretation, it is worth noting that two recent meta-analyses examining the predictive utility of the IAT found no differences in the correlation between IAT and criteria, based on the algorithm used (Greenwald, Poehlman, Uhlmann & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Perhaps reflective of this dynamic, the two measures had virtually the same correlation with the estimate of explicit race attitudes,  $rs(656) = 0.14$ ,  $p < .001$ .

IAT<sub>RAW</sub> was relatively uncorrelated with GPS,  $r(656) = 0.03$ ,  $ns$ , suggesting that for the current study population, GPS is not meaningfully confounded with IAT<sub>RAW</sub> (see Note 2).<sup>4</sup> Nevertheless, the observed correlation between SD<sub>WI</sub> and GPS ( $r = 0.38$ ,  $p < .01$ ) provides perspective on why division of IAT<sub>RAW</sub> by SD<sub>WI</sub> can produce an IAT<sub>D</sub> estimate that is less strongly associated with GPS in instances in which the correlation is higher than observed here. The correlation between GPS and SD<sub>WI</sub> was 0.38, suggesting that SD<sub>WI</sub> weakly reflects GPS (i.e., SD<sub>WI</sub> accounts for  $0.38^2 = 0.14$  or 14% of the variation in overall processing speed). If one divides IAT<sub>RAW</sub> by SD<sub>WI</sub>, some of the variance SD<sub>WI</sub> shares with GPS will be removed from the resulting (IAT<sub>D</sub>) value. This approach, however, is a rather crude strategy for removing GPS confounds. If one truly seeks to statistically control for a potential confound, a better and commonly used method is to obtain an independent estimate of GPS (as we have done here) and then use this measure as a statistical covariate in analyses oriented around the IAT<sub>RAW</sub> score (see Blanton, Jaccard, Gonzales, & Christie, 2006). Having said that, the focus of the current article is not on identifying the best method of controlling for GPS and other IAT confounds, but rather on how psychological assessment might be affected by a new IAT<sub>D</sub> scoring algorithm that was adopted in response to concerns over GPS and other forms of confounding. We now turn to this issue.

## Assessment

We examined three approaches to placing respondents into different bias categories. First, we examined the approach used by the IAT website prior to 2003. With this approach, the implicit attitude score for an individual was defined as IAT<sub>RAW</sub>, which was then divided by the estimate of the between-subjects standard deviation for IAT<sub>RAW</sub> (which in the present data was  $SD = 161.90$ ). This procedure converted each "raw" score into a Cohen's effect size estimate, from which respondents were categorized as having "strong," "moderate," or "slight" racial preferences using cut-points based on Cohen's criteria of 0.20, 0.50, and 0.80

**Table 4.** Percentage Distribution of Bias Categories Based on Three Assessment Methods.

Bias category	IAT <sub>RAW</sub> Cohen	IAT <sub>D</sub> Cohen	IAT <sub>D</sub> Current
Strong pro-White	50.4	57.2	19.4
Moderate pro-White	14.5	11.0	30.7
Slight pro-White	9.3	7.1	19.9
No bias	11.7	9.6	19.0
Slight pro-Black	5.9	5.8	7.4
Moderate pro-Black	3.9	4.4	3.0
Strong pro-Black	4.3	4.9	0.7

Note. IAT<sub>RAW</sub> Cohen lists percentages of respondents that would be placed in each bias category, using the conventions utilized by Implicit Association Test (IAT) demonstration website prior to 2003 (i.e., applying Cohen's criteria of 0.20, 0.50, and 0.80 after dividing IAT<sub>RAW</sub> by an estimate of the between-subject standard deviation). IAT<sub>D</sub> Cohen lists percentages that would be obtained if Cohen's criteria were applied to the new IAT<sub>D</sub> scoring algorithm. IAT<sub>D</sub> Current lists percentages that would be obtained using the new IAT<sub>D</sub> scoring algorithm and the criteria of 0.15, 0.25 and 0.65 that are currently in use on the IAT demonstration website.

for defining small, moderate, or large effect sizes. This same approach was next applied to IAT<sub>D</sub>, this time using an  $SD = 0.36$  and again applying the standard Cohen cut-values. The third approach also used the IAT<sub>D</sub> scores, but it instead categorized individuals based on the new cut-values used by the primary IAT website, namely scores of 0.15, 0.25, and 0.65.

The percentages of individuals in each category for the three approaches are presented in Table 4. For the two approaches that incorporated cut-values based on Cohen's criteria, the method that used IAT<sub>D</sub> scores yielded more extreme characterizations of pro-White bias than the method that utilized IAT<sub>RAW</sub> scores. For example, 50.4% of respondents were classified as strongly pro-White using IAT<sub>RAW</sub>, whereas 57.2% were so classified with a comparable method using IAT<sub>D</sub>. However, when the Cohen cutoff values for the IAT<sub>D</sub> were replaced by those currently used by the IAT website, the distribution systematically shifted toward reduced prejudice (see Table 4). Moreover, the distribution of scores for this one data set closely resembles the distribution of scores now reported by the IAT webpage. (Compare estimates in the third column of Table 4 to those reported in Table 1.) With this distribution, only 19.4% of respondents were classified as strongly pro-White. The large reduction in strongly pro-White IAT scores is due to the fact that the current IAT<sub>D</sub> cutoffs (0.15, 0.35 and 0.65) are more extreme than what were originally applied to the IAT<sub>RAW</sub>, mapping onto Cohen's effect sizes of 0.42, 0.97, and 1.81 as compared with 0.20, 0.50, and 0.80 in the other algorithm. We could locate no explanation for this shift in standards in the literature and the dynamics we observed underscore the arbitrary character of the cutoff criteria.<sup>5</sup>

Inspection of the contingency table between IAT<sub>RAW</sub> categorizations and the current IAT<sub>D</sub> categories showed that 46% of respondents would receive the same category feedback whether the IAT<sub>RAW</sub> algorithm is used or the current IAT<sub>D</sub> algorithm is used. However, 54% of respondents would receive new category feedback, with less biased categorizations being the norm when the current IAT<sub>D</sub> approach is used.

Another way to show the role that increased trial error plays in shifting bias characterizations is to compare  $SD_{WI}$  scores for the  $n = 228$  respondents who received the same categorization for both the IAT<sub>RAW</sub> and the current IAT<sub>D</sub> methods, with the  $n = 364$  respondents who had their categorization shift downward in the direction of more racial neutrality with IAT<sub>D</sub> scoring and the current cutoffs. Those who shifted toward more neutrality had higher  $SD_{WI}$  scores ( $M = 428.88$ ,  $SD = 87.21$ ) than those whose diagnoses remained constant ( $M = 379.85$ ,  $SD = 110.58$ ),  $t(590) = 6.00$ ,  $p < .001$ . Those who shifted also had longer average response latencies ( $M = 938.04$ ,  $SD = 156.45$ ) than those who did not shift ( $M = 889.04$ ,  $SD = 165.97$ ),  $t(590) = 3.62$ ,  $p < .001$ . Both results suggest that those who shifted toward neutrality might have had lower motivation or ability to attend to the computer task, causing them to receive less extreme attitudinal diagnosis. In contrast, only two individuals in the data set shifted to a more extreme racial bias estimate with the new scoring algorithm and scoring convention (with one shifting from "slight" to "moderate" pro-White categorization and another shifting from a neutral categorization to a slight pro-Black preference). These two individuals had the two smallest trial errors in the data set, each scoring over 2.3 standard deviations below the sample mean  $SD_{WI}$ . The effect of their being highly consistent in responding, despite the potentially distracting online assessment procedure, was that each appeared to be more racially biased when scored using the new algorithm.

## Discussion

These data add perspectives to the results of the computer simulation. The simulation demonstrated that, if a researcher is effective at reducing trial error, then the result will be more extreme diagnoses of implicit bias—regardless of the psychological construct being assessed. The current data show that, for an online sample taking an IAT designed to measure implicit racial preferences, the levels of trial error one can expect will tend to be sufficiently large, such that when combined with the cut-points used by IAT researchers, many individuals will be given less extreme feedback than would otherwise be the case. The results of the online study underscore the somewhat arbitrary character of the cut-points used for IAT<sub>D</sub>, especially when they are mapped onto Cohen effect size standards. Prior to 2003, the cut-points mapped onto Cohen effect sizes of 0.20, 0.50, and 0.80. Based on our online study, they now map onto Cohen



effect sizes of about 0.42, 0.97, and 1.81. What are the cut-points that should be adopted? What score truly reflects a slight degree of bias, a moderate degree of bias, and strong bias relative to implicit constructs? In our opinion, the field has mostly shirked such questions and simply accepted the cut-points provided by the architects of the IAT.

Unfortunately, we have no external criteria to know if inferences of degree of bias are more accurate when generated via the  $IAT_{RAW}$  approach or the  $IAT_D$  approach. Is the true implicit bias in the most extreme category favoring Whites 59% (per the second column of Table 4) or is it closer to 19% (per the third column of Table 4). One simply does not know. What is known is that on the much used IAT demonstration website, characterizations of the degree of pro-White bias in American society shifted dramatically in 2003, when the  $IAT_D$  replaced the  $IAT_{RAW}$  and when cut-values based on Cohen effect sizes were replaced a new set of scores. The lack of any external bases for knowing what base rates should be observed with valid instrument designed to assess implicit racial bias needs to be addressed in future research.

## General Discussion

The introduction of the new scoring algorithm for the IAT represented an attempt to address a range of artifacts known to affect  $IAT_{RAW}$ . In addition to potential confounding by GPS, known confounds that have concerned researchers include such individual difference factors as the respondent's more general skills on cognitive tasks (e.g., Rothermund & Wentura, 2004), their prior experience taking the IAT (Nosek et al., 2002), social desirability/faking (e.g., Steffens, 2004), and task-switching skills, intelligence, and cognitive flexibility (e.g., Mierke & Klauer, 2001; Rothermund & Wentura, 2004). It is important to address confounding factors such as these, but it was probably not realistic to think that the mere act of dividing  $IAT_{RAW}$  by  $SD_{WI}$  would eliminate the effects of these known artifacts, much less those unknown artifacts that might be awaiting discovery.

The data from the online study suggest that the switch to the new algorithm typically will have little empirical consequence when the researcher's goal is to use IAT scores to predict psychological criteria. This is because of the high degree of association between  $IAT_{RAW}$  and  $IAT_D$  ( $r = 0.95$ ). Such strong associations between the two approaches suggests that both known and unknown artifacts that afflict  $IAT_{RAW}$  are probably not being corrected to a substantial degree by dividing raw IAT scores by  $SD_{WI}$ .<sup>6</sup> In some respects, this result is reassuring because it suggests that the sizeable literature that has focused on correlational tests using  $IAT_{RAW}$  scoring would yield comparable inferences had  $IAT_D$  instead been used (and see Oswald et al., 2013 for empirical support of this point).

## Assessment Implications

It is only with assessment strategies whose goal is not simply ordering individuals along the dimension of interest but also in making statements about the absolute standing of people on the underlying dimension of implicit attitude, that consequential differences between the scoring algorithms are observed. Our results suggest the  $IAT_D$  is problematic in ways the  $IAT_{RAW}$  is not. Simple inspection of the algebra of the scoring algorithm reveals that, when trial error exerts little influence on test scores, all individuals are given the same extreme IAT score (see the appendix). The simulation added to this mathematical fact by showing that the  $IAT_D$  can grossly misrepresent the degree of bias in a nonbiased population as a result, and this was true for all levels of trial error. The online study reaffirmed this consequential role of trial error, and it further underscored the arbitrary nature of the cut-values currently associated with  $IAT_D$  assessments.

The most troubling property of  $IAT_D$  approach is the fact that with this approach to scoring, trial error exerts independent influence on psychological inferences. The current study focused on the measurement of implicit racial bias, but the logic applies to any IAT measure that relies on the new scoring algorithm. On the face of it, it seems untenable that individuals taking different forms of the IAT should be diagnosed as being more racially biased, depressed, narcissistic, or socially phobic—or that their scores on most any psychological construct of interest should be so influenced—simply because they managed to have little random error across experimental trials when reacting to the test stimuli.

In standard psychometrics, the accuracy of inferences is generally *improved* as one brings factors associated with item-level noise under control. People do not become more extreme. Rather, their scores become more closely centered on their true scores (to the extent that the test also is valid). Moreover, in standard psychometric treatments of test data, reduction of item-level noise will result in better internal reliability. The simulation study showed that, although the reliability of the  $IAT_D$  increases as trial error is brought under control, this is only true up to a point. Once trial error is diminished sufficiently, the  $IAT_D$  becomes less reliable because of the homogenization dynamics that operate. These two related effects of diminished trial error on IAT scores—the categorization of all respondents as extreme and the ultimate decrease in test reliability—should give researchers pause about using the D-score algorithm.

## Alternatives to D-Score Assessments

Given the widespread application of this assessment tool in clinical, organizational, and legal settings, and its wide dissemination to the public (to millions of individuals who have turned to the IAT demonstration website to receive

feedback on their hidden psychological attributes), better scoring procedures and assessment practices are needed in the IAT literature. A place to start is to revisit the logic underlying IAT-scoring algorithms. The stated hope of pursuing new scoring procedures like  $IAT_D$  is that the new algorithms will somehow expunge contaminants from the IAT score, leaving a new score that has been corrected for known confounds. The spirit of such an approach is problematic because the strategy (a) presumes the mathematical operations are precise enough to remove the effects of confounds when, in fact, such precision is likely not attainable from a simple algebraic correction and (b) makes strong assumptions about the universality of the correction; that is, that it will apply across populations and contexts of interest, regardless of the specific construct being measured or the stimuli used to measure these constructs.<sup>7</sup>

As an alternative, we suggest that researchers pursue more traditional methods of correcting for known confounds; methods that do not require such strong assumptions. We believe that researchers should develop independent measures of known confounds and then statistically control for these confounds using standard regression-based strategies that permit flexible modeling that is grounded in empirically derived (regression) weights. Such modeling ensures that confounds are weighted properly for the particular topic, context, and target population. In addition, the approach can be modified to accommodate both linear and nonlinear functions, with the choice of modeling informed by standard model-fit diagnostics. Such a “measure-and-model approach” to handling known confounds is straightforward. It also can be adapted to take advantage of latent variable modeling, with multiple indicators to address both random and systematic error.

### Making Assessments Less Arbitrary

Regardless of how confounds are addressed, a more significant change in practices associated with current IAT conventions is to move away from the categorization of individuals based on arbitrary cut-values and scoring criteria. As Blanton and Jaccard (2006a, 2006b) noted, both the  $IAT_D$  and the  $IAT_{RAW}$  have *arbitrary metrics*. That is, it is neither 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. Arbitrary metrics become meaningful 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; Sechrest, McKnight, & McKnight, 1996). For instance, for 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. But do IAT scores of zero truly reflect the absence of bias? IAT researchers have yet to tackle this question empirically with any vigor, and they have instead defaulted to assuming that current conventions lead to meaningful categorization of individuals. Imagine, for purposes of comparison, if the *Diagnostic and Statistics Manual of Mental Disorders* categorized individuals with mental disorders based on where their scores fell along different psychological metrics without any reference to external diagnostic criteria. To move beyond the shortcomings highlighted in the current analysis, IAT researchers need to move to classification systems that are more empirically grounded.

As one example of an empirical strategy, an organizational researcher might first empirically identify the score on the IAT (using whichever algorithm is preferred), where one observes no impact of applicant race on employer hiring decisions—the score where an employer is as likely to hire Whites as to hire Blacks, all else being equal (see Blanton & Jaccard, 2006b). This strategy suggests an empirical-based zero point of the metric (which may or may not be the observed zero point on the IAT). This information can then be used to support the development of more meaningful cut-points for creating categories, with increasingly more consequential degrees of observed hiring bias being mapped directly onto increasingly higher scores on the IAT (e.g., at a value of  $X$ , 10% more Whites are likely to be hired than Blacks). The original approach adapted by the IAT webpage was to incorporate Cohen’s criteria for all the different substantive IAT measures, although the webpage then abandoned these criteria for unknown reasons. If researchers instead pursue the empirical strategies we espouse for mapping specific IAT scores onto behavior to derive understanding of score meanings, then over time they can make a case for *empirical* cut-points that define categories based on the documented meaning of test scores. For hiring bias, for instance, one could identify the specific scores on the IAT 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 (see Blanton & Jaccard, 2006b). As the full range of scores take on meaning over time, consensus can 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. Research of this character—research that essentially makes an arbitrary metric nonarbitrary—is needed to truly imbue the IAT metrics with meaningful cut-points. It is our hope that the present article encourages such efforts.

## Appendix

### Implications of Including $SD_{WI}$ in IAT D Score

An important feature of the IAT D score is that, for any given individual, the value of  $SD_{WI}$  is calculated across both the compatible and incompatible trials rather than within each block of trials separately. It is thus influenced by two factors, (a) the amount of trial error (random noise) that is operating and (b) the difference between IRL and CRL (which is of interest in estimating implicit bias).  $SD_{WI}$  can be decomposed into these two components, namely the systematic component that represents the mean differences between blocks ( $SD_{BET}^2$ ) and the random noise component that represents within-block variability in the compatible and incompatible tasks ( $SD_{NOISE}^2$ ):

$$SD_{WI} = \left( SD_{BET}^2 + SD_{NOISE}^2 \right)^{1/2}$$

It can be shown mathematically that, assuming an equal number of compatible and incompatible trials,  $SD_{BET}$  equals one half the difference  $IRL - CRL$  and that  $SD_{NOISE}^2$  equals one half the sum ( $SD_{IC}^2 + SD_C^2$ ), where  $SD_{IC}^2$  is the within-block variance of the latency scores for the incompatible trials and  $SD_C^2$  is the within-block variance of the latency scores for the compatible trials (i.e., trial error within each block). If there is no random noise or trial error influencing the within-block trials, then  $SD_{WITHIN}^2 = 0$  and the  $IAT_D$  reduces to

$$IAT_D = (IRL - CRL) / [ (0.50 (IRL - CRL)) ]$$

where IRL and CRL are the mean response latencies for the incompatible and compatible blocks, respectively. This can be further reduced to

$$IAT_D = 2.0$$

Thus, without trial error,  $IAT_D$  will always equal the mean difference between  $\mu_{IRL}$  and  $\mu_{CRL}$  divided by half this same difference, and so its absolute value must equal 2.0.

### Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### Notes

1. <https://implicit.harvard.edu/implicit/user/pimh/selectastudy.html> (March, 2014).
2. Greenwald, Nosek, and Banaji (2003) determined bias by correlating the average response across all IAT trials (compatible

and incompatible) with scores on those same trials but combined by differencing (via Equation 1). Using the identical items for both a measure of response speed and a measure of implicit attitude introduces complex mathematical dependencies between the indices. We prefer to assess the constructs with different responses to avoid such dependencies. We therefore use the person's average response latency during the learning trials for the IAT to index response speed (i.e., trials where respondents are learning to correctly discriminate individual stimuli). These response latencies are not used in the computation of the IAT raw or D scores. In our research, we find that the average speed in these learning trials correlates with average raw response latencies for IRL and CRL (separately) at values between  $r = 0.50$  and  $0.80$ . This suggests that speed at the learning trials does capture some of the same individual difference factor that influences speed at performing the experimental trials. However, unlike the approach used by Greenwald et al. (2003), the learning-score index we use is often only slightly correlated with the  $IAT_{RAW}$  difference score (see Blanton et al., 2006, Table 1, p. 199 and Table 3 of the current report and compare these values with those found in Greenwald et al., 2003, table 1, p. 198). This finding suggests there may have been less need to pursue a new algorithm to address the confounding of  $IAT_{RAW}$  with average response latency than was originally posited. However, now that the alternative  $IAT_D$  algorithm has been adopted to address this and other potential confounds, it is important to consider the implications of the new algorithm for psychological assessment.

3. Greenwald et al. (2003) also recommend computing an  $IAT_D$  for a first set of practice trials in each block and a second set of experimental trials. This is effectively averaging two  $IAT_D$  and does not influence our analysis.
4. Using an alternative operationalization of GPS (general processing speed) that incorporates CRL and IRL, consistent with Greenwald et al. (2003), GPS was modestly correlated with  $IAT_{RAW}$ ,  $r(656) = 0.13$ ,  $p < .01$  but not with  $IAT_D$ ,  $r(656) = .02$ , *ns*.
5. The disparity between  $IAT_{RAW}$  and current  $IAT_D$  categorizations was first noted by Blanton and Jaccard (2009), who documented that prior to 2003, the IAT website estimated that 48% of the test-taking population had strong preferences for Whites relative to Blacks, but that after adoption of the new scoring algorithm, that same estimate dropped to 27% (see Table 1). The differences in percentages observed in our data mimic these shifts in estimates occurring over the history of IAT website, providing an empirical account for why there was such a dramatic shift in feedback provided to online survey respondents.
6. We have found similar levels of association in our secondary analysis of IAT data sets provided to us by other researchers.
7. These concerns can be exacerbated by some of the arbitrary data handling procedures found in the IAT literature. For instance, Greenwald et al. (2003) suggest that, prior to computing an  $IAT_D$  score, researchers might add 600 milliseconds to every IAT trial answered incorrectly because this procedure might correct for the fact that incorrect answers are often made more quickly than correct answers. This approach to data handling only increases the need to embrace strong assumptions regarding the precision and invariance of the IAT metric.

## References

- Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York, NY: Random House.
- Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist*, 61, 27-41.
- Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist*, 16, 62-71.
- Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, 34, 277-297.
- Blanton, H., & Jaccard, J. (2014). *Not so fast: Ten challenges to importing implicit attitude measures to media psychology*. Manuscript submitted for publication.
- Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Perspectives on criterion prediction. *Journal of Experimental Social Psychology*, 42, 192-212.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3-5.
- Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Gladwell, M. (2005). *Blink: The power of thinking without thinking*. New York, NY: Little, Brown.
- Greenwald, A. G. (2006, September 3). *Expert report of Anthony G. Greenwald, Satchell v. FedEx Express*, No. C 03-2659 (N.D. Cal.).
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85, 197-216.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & 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.
- Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist*, 61, 42-49.
- McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-125). Orlando, FL: Academic Press.
- McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the implicit association test. *Social Cognition*, 20, 483-510.
- Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für Experimentelle Psychologie*, 48, 107-122.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, 6, 101-115.
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., . . . Smith, C. T. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18, 36-88.
- Oswald, F., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT research. *Journal of Personality and Social Psychology*, 105, 171-192.
- Pendry, L. F., Driscoll, D. M., & Field, S. T. (2007). Diversity training: Putting theory into practice. *Journal of Occupational and Organizational Psychology*, 80, 27-50.
- Plous, S. (2003). *Understanding prejudice and discrimination*. New York, NY: McGraw-Hill.
- Roefs, A., Huijding, J., Smulders, F. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. M. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin*, 137, 149-193.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Ed.). Thousand Oaks: Sage Publications.
- Rosen, P. J., Milich, R., & Harris, M. J. (2007). Victims of their own cognitions: Implicit social cognitions, emotional distress, and peer victimization. *Journal of Applied Developmental Psychology*, 28, 211-226.
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General*, 133, 139-165.
- Rudman, L. A. (2008). The validity of the Implicit Association Test is a scientific certainty. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 426-429.
- Rüsch, N., Schulz, D., Valerius, G., Steil, R., Bohus, M., & Schmahl, C. (2011). Disgust and implicit self-concept in women with borderline personality disorder and posttraumatic stress disorder. *European Archives of Psychiatry and Clinical Neuroscience*, 261, 369-376.
- Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, 94, 18-38.
- Scheck, J. (2004, October 28). *Expert witness: Bill Bielby helped launch an industry—suing employers for unconscious bias*. Retrieved from <http://www.law.com/jsp/PubArticle.jsp?id=900005417471>
- Schmidt, K., & Nosek, B. A. (2010). Implicit (and explicit) racial attitudes barely changed during Barack Obama's presidential campaign and early presidency. *Journal of Experimental Social Psychology*, 46, 308-314.
- Schreiber, F., Bohn, C., Aderka, I. M., Stangier, U., & Steil, R. (2012). Discrepancies between implicit and explicit self-esteem among adolescents with social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, 43, 1074-1081.
- Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist*, 51, 1065-1071.
- Steffens, M. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, 51, 165-179.
- Vendatam, S. (2005, January 23). See no bias. *The Washington Post*. Retrieved from <http://www.washingtonpost.com/wp-dyn/articles/A27067-2005Jan21.html>
- Whitley, B. E. Jr., & Kite, M. E. (2009). *The psychology of prejudice and discrimination* (2nd ed.). Belmont, CA: Thomson/Wadsworth.