

Rapid #: -27751344

CROSS REF ID: **4303528**

LENDER: **BTS (Bates College) :: Ladd Library**

BORROWER: **TXA (Texas A&M University) :: Main Library**

TYPE: Book Chapter

BOOK TITLE: The Cambridge Handbook of Implicit Bias and Racism

USER BOOK TITLE: The Cambridge Handbook of Implicit Bias and Racism

CHAPTER TITLE: Listening to Measurement Error: Lessons from the Implicit Association Test

BOOK AUTHOR: Blanton

EDITION:

VOLUME:

PUBLISHER: Cambridge University Press

YEAR: 2024

PAGES: 337 - 373

ISBN: 9781108840309

LCCN:

OCLC #:

Processed by RapidX: 10/21/2026 8:01:02 AM

This material may be protected by copyright law (Title 17 U.S. Code)

12 Listening to Measurement Error: Lessons from the Implicit Association Test

Hart Blanton and James Jaccard

The implicit attitude construct has captured the imaginations of researchers across multiple disciplines and shifted research priorities in the attitude field. A major stimulus for this movement was the hope that new technology could reveal underlying attitudes that had eluded attitude researchers for decades. Interest in implicit measures has extended across research domains but it is most strongly rooted in the study of prejudice and discrimination, where *implicit bias* has become an organizing concept. Implicit bias has been advanced as a potent source of discriminatory behavior and judgment, and so new implicit bias measures have been advanced as tools that can be used to predict and explain these actions. That's been the hope, and researchers have had well over twenty years now to produce the goods.

Healthy Skepticism

The two of us entered the implicit attitude domain about two decades ago, as believers in the foundational premises undergirding implicit bias research. Much of what makes social psychology fascinating to us are the errors and biases in judgments that operate outside awareness, often with consequential effects. The implicit bias construct has brought important perspectives to the study of prejudice and discrimination, where greater awareness of bias in society at large has the potential to promote social justice and social welfare. Like most of our research, we approached this new assessment method with a degree of

healthy skepticism. Unfortunately, our skepticism has grown over the course of our research program.

We focused most of our research attention on one measurement technique, the Implicit Association Test (or *IAT*; Greenwald et al., 1998). We made this decision for three primary reasons. First, researchers using this assessment tool were making what we viewed as dubious claims that lacked empirical support, not only in their professional publications but also in their interactions with the popular press, particularly in their promotion of a demonstration webpage. Second, through strong interpretations of what was essentially a flawed psychometric inventory, IAT advocates began introducing questionable analytic and measurement approaches. We feared the result could be an erosion of scientific norms in this field of study, which might influence broader norms in research psychology. Third, we were concerned that a science of racism based on weak measurement might lead to the misdirection of intellectual and other resources in the fight against bias and discrimination. As such, beginning in 2002, we started a program of research to evaluate this new implicit attitude construct and measures of it in light of modern attitude theory and psychometric theory, with our first publications appearing in 2006 (Blanton & Jaccard, 2006c; 2006d; Blanton et al., 2006). These papers resulted in open debates with IAT architects, who have consistently defended their practices against our critiques. In this chapter, we review our work in

this area and then discuss recent attempts to address the concerns we have introduced.

Back to the Future

Prior to the 1998 introduction of the IAT, there were many unsuccessful attempts to measure “true” or “hidden” attitudes, independent of what people self-report. One effort, known as the bogus pipeline technique (Jones & Sigall, 1971), sought to trick people into thinking that psychologists could measure how they really felt about a topic, the implication being that it was futile to try to mislead researchers by reporting false attitudes. This ruse was accomplished by having a person connected to what was essentially a fake polygraph and then having them answer questions about their attitudes and opinions. The idea was that people would not want to be “second-guessed” by a machine and would therefore respond truthfully. Alternatively, individuals were asked to guess the response of the machine, again, on the assumption that doing so would produce truthful responding. Some twenty years after its introduction, Roese and Jamieson (1993) provided a critical evaluation of the method and concluded that the method does, indeed, tend to produce measurable differences relative to traditional self-reports. Despite an initial surge in use, however, the technique was largely abandoned by social psychologists, because of perceptions of its ineffectiveness, its impracticality, its dubious theoretical base, its sole focus on encouraging truthful responding, and its questionable ethics. Roese and Jamieson (1993) explained its “rise and fall” based on Kenneth Ring’s (1967) analysis of faddishness in social psychology in which “a new (or seemingly new) territory is discovered, explored for a while, and then usually abandoned when the going gets rough or uninteresting” (p. 120).

Another early effort used the physiological tools of the day, such as the Galvanic Skin Response, the Palmer Sweat Index, and measures of pupillary dilation or heart rate as an index of attitude that people could not falsify. For example, an early study by Rankin and Campbell (1955) used the galvanic skin response (GSR), an index of skin conductance thought to assess activity of the “fight-or-flight” instinct controlled by the sympathetic nervous system. They found higher GSR levels in Whites who had incidental hand contact with a Black experimenter as compared to a White experimenter, a result that escaped replication in a later study by Porier and Lott (1967). Hess (1965) argued and provided evidence that pupil dilation and constriction occurs not only because of changes in light intensity but also because of ongoing mental activity focused on interest, emotions, thought processes, and attitudes. He argued for the use of pupil dilation and constriction as an indicator of people’s attitudes when they are exposed to an attitude object. Similarly, Buckhout (1966) found a relationship between changes in heart rate that accompanied exposure to counter-attitudinal communications and subsequent attitude change. He then argued that changes in heart rate might act as a physiological marker when attitude change occurs.

Although each of these tools generated initial excitement as potential indices of an individual’s “true” attitudes, they eventually were rejected by the scientific community, in part, because they were affected by so many contextual factors other than attitudes (e.g., moods, hunger, light, distractors). Moreover, even minor shifts in the response metric could have dramatic effects on the determination of one’s attitudes, which was deemed problematic given high levels of random error in many of the measures (as is true when response latencies are used to

measure implicit attitudes – an issue we explore in our research program).

A third early effort to measuring “true” or “hidden” attitudes independent of self-reports focused on unobtrusive measures, elaborated in the classic book by Webb, Campbell, Schwartz, and Sechrest (1966), entitled *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. These researchers analyzed the behavioral traces that individuals leave behind, traces that supposedly indicate their attitudes. Examples of these traces included the relative wear and tear on floor tiles in museums, used to measure attitudes toward different paintings, or the choice of radio stations in cars being serviced at local auto repair shops, used to identify community listening preferences. Although still used today, these methods also fell into disfavor because they often failed to yield individualized indices of attitudes that had sufficient precision.

It was against this backdrop that an early implicit measure was introduced to social psychology as yet another attempt to provide a potential “pipeline” that could unobtrusively detect attitudes (Fazio et al., 1995). And it was against social psychology’s soul searching relative to faddism that we read the early work on the IAT. This new measure felt familiar to us in that its early intent was to tap hidden attitudes. Unfamiliar to us was the decisions the IAT architects made to engage the public through media representations of the fidelity of this new assessment tool. It was also to our surprise that this untested tool received funding from the National Science Foundation, helping IAT researchers establish a demonstration webpage where the public could obtain feedback on their hidden implicit associations. These events occurred well before the IAT had advanced a reasonable psychometric model (which we argue it still lacks) and before it had been subjected to strong scientific and psychometric evaluation.

It was against these headwinds and in contradiction to a group of high-stature defenders that we set out to scientifically evaluate this new instrument.

Our Research Program: A Summary of Concerns

Here we review eleven summary points that have evolved from our research program, over roughly two decades. We encourage readers to consult the original citations, which provide extensive details about each point. For each one, we clarify if it is an issue unique to the IAT or one that is shared by implicit attitude measures more generally. We begin with matters of measurement and then turn to broader methodological and theoretical challenges that we have identified.

1. The IAT Has Poor Test-Retest Reliability and Is Contaminated by Large Amounts of Random Error

The reliability of the IAT is variable but it generally is far lower than minimal psychometric standards required for meaningful assessment or robust behavioral prediction. This problem likely extends to any implicit attitude measure that scales individual differences using response latencies (see Jaccard, 2023, Chapter 15 this volume, for discussion of this point). Blanton, Burrows, and Jaccard (2016) found the two-week test–retest reliability of an IAT measure of implicit attitudes toward binge drinking in college youth to be 0.24, which is striking given the expected stability of such attitudes in this population. Steffens and Buchner (2003) evaluated test–retest reliability when the IAT retest was an immediate replication of a just-measured IAT implicit attitude. The test–retest correlations ranged from 0.50 to 0.62. Greenwald et al. (2006) report an average test–retest reliability of 0.56

for the IAT, in studies covering a wide range of topics. Cunningham et al. (2001) administered the “race IAT” (measuring the preference for Whites relative to Blacks) across four two-week time periods and obtained an average test–retest reliability of 0.27. Gawronski, Morrison, Phills, and Galdi (2017) reported a one-month, test–retest reliability near 0.50 for both the IAT and other measures of implicit attitudes. Gawronski et al. (2017) reviewed twenty-three studies and found the average IAT test–retest reliability was near 0.40. Keep in mind that a reliability of 0.40 means that $1 - 0.40 = 0.60$ or 60 percent of the variation in the measure is random error. This is like consulting a random number table, choosing a fairly sizeable number, and then arbitrarily adding or subtracting that number to a person’s true score. To us, and to most psychometricians, this is substandard. We think such low reliability would typically prevent researchers from advancing a new measure, such as a new explicit attitude measure, in reputable science journals.

A common defense of the low test–retest reliability for the IAT calls attention to other forms of reliability estimation, where the estimates are more favorable (but still often substandard). For example, IAT researchers often argue that although test–retest reliability is low by conventional standards, reliability estimates that use Cronbach’s alpha or split-half reliabilities suggest higher reliability (Schnabel et al., 2008). Cronbach alphas generally yield estimates in the 0.60 to 0.80 range, with split-half reliabilities somewhat lower than this. What accounts for the disparity between these different methods?

It is well known that different methods of reliability estimation are sensitive to different types of random error. This is especially true of test–retest methods versus methods based on the analysis of items within a testing session. Noise that varies more between sessions

(e.g., varying mood states) than within sessions (e.g., attention span) can produce a pattern of higher internal reliability but lower test–retest reliability. When evaluating the reliability of a measure, one should do so from multiple perspectives. In our opinion, Cronbach’s alpha is problematic when applied to the IAT for multiple reasons. First, the coefficient assumes unidimensionality of items. Trials on the IAT are analogous to items on a multi-item attitude survey so to apply coefficient alpha, one should first affirm unidimensionality among the IAT trials within blocks. Research suggests that unidimensionality of IAT trials likely does not hold. For example, Yamaguchi and Beattie (2020) found task-irrelevant attributes of trial stimuli (e.g., gender of the person depicted in pictures for the race IAT; race of the person depicted in pictures for the gender IAT) impact IAT responses, which can produce a multidimensional structure. Our own research indicates that the two IAT trial blocks (commonly referred to as “compatible” and “incompatible”) are influenced by a common factor independent of the two association-strengths the blocks supposedly measure. This factor is a person’s ability to react quickly to cognitive tasks in general (Blanton et al., 2007), a phenomenon we refer to as “general processing speed” (GPS).

The common influence of GPS on trial responses suggests the presence of what is known as a bifactor model (Gibbons & Hedeker, 1992) rather than a unidimensional model. Multidimensionality also is posited by the “quad model” of the IAT, where Conrey and colleagues (2005) argue that IAT scores generate from “the joint contributions of multiple, qualitatively different processes” (p. 469). Finally, multidimensionality is suggested by studies documenting a range of non-attitudinal influences that contaminate IAT responses (e.g., Rothermund & Wentura,

2004). Given this, reliability estimates of the IAT using coefficient alpha are suspect.¹

Coefficient alpha also is influenced by the number of items/trials on a test. For example, for forty trials per block (the common number used for the IAT), an average correlation of only 0.09 between the items/trials would yield a coefficient alpha of 0.80. A correlation of 0.09 between trial responses implies that “items” that supposedly measure the same underlying construct (implicit attitudes) are dominated by unique/random factors rather than the common variance produced by the presumed influence of an underlying implicit attitude. For the IAT, trial interrelatedness tends to be trivial in magnitude after controlling for GPS. If GPS or other sources of systematic error that bias IAT trials are not controlled when calculating coefficient alpha for the IAT, the reliability estimate becomes misleading because the composite reliability is driven primarily by GPS and other systematic error.

An alternative form of reliability estimation sometimes used for the IAT is split-half methodology, which is based on correlations between total scores for two halves of the items or trials. With split-half estimation, corrections typically are made to accommodate the fact that the reliability estimate is for a half-test rather than a full-test. A weakness of the split-half approach is that subsets of items/trials can be formed in different ways; one can obtain different results depending on the subsets. If the two halves have an equal number of trials/items, it turns out that coefficient alpha will equal the mean (Flanagan-Rulon) reliability of all possible ways of creating the split-halves (Cronbach, 1951). For this reason, coefficient alpha is generally preferred to split-half indices, but for reasons already discussed, the use of alpha with the IAT is problematic. Another weakness of the split-half method is that, like coefficient alpha, the

method is only sensitive to sources of within-session, item-to-item random error. As well, corrections to adjust for half-test length make assumptions (e.g., parallelism of the halves) that may be untenable and if violated can undermine estimates. Finally, like coefficient alpha, the method assumes item unidimensionality which, if violated, can compromise estimates. Raykov and Marcoulides (2011) conclude that split-half methods “have serious limitations that in general preclude them from leading to dependable reliability estimates in empirical and behavioral research” (p. 152). For the IAT, split-half reliabilities tend to be substandard in magnitude, but even when they achieve satisfactory levels, the meaning of the estimates is unclear.

Researchers who argue for the value of the IAT often give little consideration to the reasons why internal reliability estimates can be problematic for the IAT. Rather, they tend to deemphasize the importance of the low test–retest reliability estimates and emphasize the higher internal reliability estimates. For instance, some theorists have cited supposedly reasonable internal reliability estimates and then sought to reconceptualize implicit attitudes as transitory rather than stable. An example of this argument can be found in Jost (2019), who argues that implicit attitudes are highly context-dependent and can shift from one context to the next, hence one should expect low test–retest reliability but high internal reliability. This argument fits a broader pattern of IAT defenses that we document later: Rather than treat potentially damaging findings as potential warning signs

¹ Some researchers believe high coefficient alphas confirm unidimensionality, despite convincing arguments to the contrary (see Cortina, 1993; Davenport et al., 2015). The reality is that Cronbach’s alpha *assumes* but does not *test* unidimensionality.

of a problematic measure, interpretations of the construct are changed and the measure's value preserved. Recall, however, that early on, researchers presented the IAT to the public as a tool to measure hidden biases. What enduring lesson can respondents learn about themselves, if the measure is highly unstable, highly context dependent, and likely to change immediately after the testing session?

We also think Jost's new interpretation fails to seriously engage the magnitude of the challenges the data patterns introduce. For instance, Steffens and Buchner (2003) found test–retest reliability with *immediate* IAT replication to be in the 0.50–0.60 range. This result suggests that the IAT might be hypersensitive to context at levels Jost (2019) does not consider. Research also points to low IAT test–retest correlations for constructs that most would assume are relatively stable. For instance, we found low test–retest reliability over two weeks for IAT measuring attitudes toward drinking alcohol in college students, and this also was true of other implicit attitude measures that used a similar psychometric approach (Blanton et al., 2016). It is hard for us to imagine what could cause college students to shift their attitudes toward alcohol use so dramatically over such a short time period.

Greenwald et al. (2022) and Jost (2019) suggest that by averaging repeated tests of the IAT for an individual, a reasonable estimate of that person's more stable implicit attitude can be had, much like what doctors might do with blood pressure readings. We view this advice as problematic and think it again reflects a tendency not to critically engage data patterns that signal warnings about the measures. Assuming an average test–retest reliability of 0.40, as reported by Gawronski et al. (2017), it can be shown mathematically that one would have to administer the IAT six times in order to achieve an aggregate score

for individuals that still contains, on average, 20 percent random noise. We doubt researchers will want to routinely adopt such multi-testing practices in their studies and we question any measurement approach that requires this large number of time-consuming assessments, just to approach the reliability one would typically expect of brief self-report measures.

A related response to the low IAT reliability has been to simply assert that reliability is high. For instance, Schnabel, Asendorpf, and Greenwald (2008) argued that “The discrepancy between satisfactory internal consistencies and smaller retest reliabilities indicates that IATs capture both stable and occasion-specific variance” (p. 212). Nosek, Greenwald, and Banaji (2007) termed reliability “substantial” (p. 112). Greenwald and Lai (2020) conducted meta-analytic summaries of test–retest reliabilities and labeled correlations of $r = 0.44$ to be “satisfactory” (supplement, p. 3). We respectfully disagree with these characterizations. We do not think that a measure that is contaminated by 56 percent random noise is satisfactory, either for purposes of assessment or behavioral prediction. Suggestions to the contrary reveal a departure from standard scientific norms. Had the early psychometric advocates of attitude measurement (e.g., Thurstone, Guttman, Likert, Osgood) come forth with measures showing these levels of test–retest reliability and then argued that they establish that explicit attitudes can be measured, they would have been rejected outright.

2. The IAT Is Weakly Correlated with Other Measures of Implicit Attitudes, Indicating It Has Low Convergent Validity

A standard often used to evaluate the validity of a measure is its convergent validity with

other measures of the identical construct, a psychometric property known as convergent validity.² There is little evidence of convergent validity for the IAT or for other implicit bias measures. Simply put, measures of implicit bias and attitudes tend to be weakly correlated with one another. For instance, Blanton, Burrows, and Jaccard (2016) evaluated convergence between the IAT and three other popular implicit attitude measures, the personalized implicit association test (Olson & Fazio, 2004), an evaluative priming test (Fazio et al., 1995), and the affective misattribution test (Payne et al., 2008). The average correlation of the IAT with these other measures was 0.10 (and, indeed, none of the measures were highly related to one another). Sherman and colleagues (2003) found the average correlation between the IAT and measures of evaluative priming to be 0.06. Fazio (1999) reports a correlation of 0.05 between the IAT and evaluative priming measures.

As with the low levels of reliability, IAT advocates have tended to show a lack of critical engagement with data patterns suggesting low convergent validity. Implicit bias researchers have suggested that different measures of implicit bias do not converge because they reflect conceptually distinct facets of implicit association (e.g., Gawronski & De Houwer, 2014). This may indeed be the case, but the field then needs a suitable theory that can account for the facet distinctions, and those distinctions must be substantively meaningful. It may be that future work can generate post-hoc theoretical frameworks about the sources of divergence between current implicit attitude measures. However, this does not change the fact that there *currently* is little evidence for convergent validity among them.

In contrast to the response from the IAT's architects, we think the scientifically prudent response to low convergent validity is to interpret scores from with caution. In this

particular case, evidence for convergence is so low that some degree of skepticism also seems warranted. Had Thurstone, Guttman, Osgood, or Likert brought forth new attitude measures that showed little to no convergence with other measures of the same construct, their measures likely would have been rejected out of hand.

3. The IAT Is Contaminated by Method Variance. The Scoring Algorithm for the IAT Was Designed to Remove It, but It Does Not

Above, we noted that early research on the IAT revealed that raw IAT scores (computed as the simple difference between response latencies, measured in milliseconds) are confounded with a respondent's average response latency, namely general processing speed (GPS) or the tendency to respond to cognitive tasks quickly (Blanton et al., 2006). The slower a person's GPS, the more extreme or biased the person will appear on any given IAT, independent of that person's true implicit attitudes and biases. This source of method variance conflates implicit bias estimates with factors related to GPS, such as respondents' intelligence, age, alcohol use, and any number of the many cognitive deficits that impact reaction times in general. The reliance on reaction times also contributes to the low test-retest reliability discussed earlier, as state changes in a person's alertness or other mental states alter the attitude estimates for that person. GPS and state changes in alertness likely affect all implicit attitude measures that rely on reaction time data.

² In the IAT literature, some researchers use the term convergent validity more loosely to include predictive validity. We use the more traditional definition here. We consider the IAT predictive power later.

The problem of GPS contamination in speeded laboratory tasks is familiar in experimental cognitive psychology. A common way of handling this problem is to subtract one reaction time (RT) from another, such as subtracting the RT measured after the presentation of control stimuli from the RT measured after the presentation of experimental stimuli. This yields an RT facilitation score representing the effect of the experimental manipulation on RT. This approach to addressing GPS contamination for an individual difference measure is lacking (see Jaccard, 2023, Chapter 15, this volume). Early research indicated, for instance, that conceptually unrelated IATs were positively correlated with one another, suggesting that simple differencing does not expunge GPS contamination from the measure (McFarland & Crouch, 2002). In response to such evidence, Greenwald, Nosek, and Banaji (2003) developed a new scoring algorithm in the hopes it would produce an IAT score that was more resistant to “an artifact associated with speed of responding,” as well as “sensitivity to known influence on IAT measures” (p. 197). This scoring algorithm goes beyond simple differencing and involves a host of additional transformations that, in our opinion, are dubious (and see point 8, below).

In our work with the IAT, we have argued for the use of more traditional analytic methods to deal with confounds rather than relying on scoring algorithms. The algorithm approach makes strong assumptions about the universality of the corrections, namely that it offers stable protection against confounds across different populations and different contexts. This assumption seems unnecessarily strong to us, given availability of common methods that do not require it. A more traditional analytic approach is to obtain an independent measure of known confounds and then to statistically control for them in regression-based analytic strategies. This

approach removes confounding effects using empirically derived regression weights as applied to the particular population and context being studied. For example, explicit attitude researchers routinely collect measures of socially desirable responding and then use these measures as covariates to control for social desirability contamination in their measures (Steenkamp et al., 2010). The same logic can be applied to implicit attitude measures and other known confounds that afflict them. For instance, to address GPS, it is straightforward to obtain a person’s average response latency independent of their IAT performance. One can then statistically control for GPS when conducting criterion-prediction studies (see Blanton et al., 2006 for an example).

One drawback to applying this statistical approach to control for a known confound with the IAT is that, although it works reasonably well for research purposes where the goal is to predict criteria from test scores after a confounding influence has been controlled, it does not lend itself to individual assessment. There are no straightforward ways of correcting for a known confound in a measure when the goal is to use a test score to provide respondents with diagnostic feedback, as is done on the IAT website. We cannot be certain, but one reason IAT researchers pursue a nonstandard approach might be that they desire a “pure” score for each person, computed on an individual-by-individual basis. This they can use for purposes of providing individualized feedback on the demonstration website. In our opinion, such a strategy is unrealistic, given the strong influence of known confounds on the IAT and the potential for each of those confounds to vary across measurement contexts.

Another challenge for the IAT is that it has so many confounds that it might be impossible to measure and control for all of them. For example, research has suggested the presence

of confounds associated with stereotype familiarity, empathy for an outgroup, acceptance of external explanations for a group's disadvantage, stereotype threat, task-switching ability, stimulus cross-category associations, figure-ground processing, concern with being labeled a racist, and even social desirability, to name but a few. Each of these constructs presents non-trivial measurement challenges and we question the value of a measure that requires such an exhaustive set of controls. Had Thurstone, Guttman, Osgood, or Likert brought forth attitude measures that showed the levels of method variance found in the IAT, we doubt they would have tried to "algorithm" the problem away. Rather, they likely would have gone back to the drawing board and developed a new measurement strategy.

4. The IAT Was Constructed so That It Measures Relative Attitudes toward Two Objects, Rather than an Attitude toward a Single Object: In Doing so, the IAT Unnecessarily Confounds Attitudinal Dimensions in Ways That Restrict Modeling and Inference

Blanton, Jaccard, Gonzalez, and Christie (2006) note that many implicit attitude measures, including the IAT, assess *relative* implicit attitudes rather than attitudes toward a specific attitude object. For the race IAT, for example, researchers seek to measure the difference between implicit evaluations of Whites relative to Blacks. Greenwald, McGhee, and Schwartz (1998) in their seminal study, provided no strong justification for doing so other than the fact they wanted to test whether their new measure could detect "evaluative differences" between objects (e.g., flowers vs. insects). However, this decision constrains the types of causal theories that can be tested. It ties the hands of researchers attempting to build theories of attitude-behavior relations.

Consider the case in which a researcher wishes to determine if implicit "age bias" predicts hiring discrimination against older job applicants. To this end, the researcher decides to design an IAT. The researcher seeks to measure implicit attitudes toward people who are old. However, owing to the structure of the IAT, the measure must include a contrast category against which implicit attitudes toward older people can be evaluated. Greenwald and Farnham (2000) acknowledged this limitation but argued it was not problematic, stating that "the IAT can nevertheless be effectively used because many socially significant categories form complementary pairs, such as positive-negative (valence), self-other, male-female, Jewish-Christian, young-old, weak-strong, warm-cold, liberal-conservative, aggressive-peaceful, etc." (p. 1023). Per Greenwald and Farnham, the obvious choice of a "complementary pair" for *old* would be *young*. However, by virtue of this constraint, a researcher is (knowingly or unknowingly) limited to testing restrictive causal models, such that hiring behavior is as much determined by disliking of older people as it is by liking for younger people.

Suppose for example that the true implicit relative attitude (TIRA) is defined as

$$\text{TIRA} = \text{TIA}_O - \text{TIA}_Y \quad (12.1)$$

where TIA_O is the true implicit attitude toward older people and TIA_Y is the true implicit attitude toward younger people. To predict discriminatory behavior, a model expressing the behavior to be a linear function of the relative implicit attitude has the form:

$$\text{DB} = a + b \text{ TIRA} \quad (12.2)$$

where DB is a behavioral indicator of discriminatory behavior. If we substitute the right-hand side of Equation 12.1 for TIRA in Equation 12.2 and expand the products, we see that the researcher will be modeling:

$$\begin{aligned} DB &= a + b(TIA_O - TIA_Y) \\ &= a + bTIA_O - bTIA_Y \end{aligned} \quad (12.3)$$

Equation 12.3 reflects a conceptual model that is restrictive in form. It asserts that the impact of the two component implicit attitudes on the criterion is additive and that the effect of each implicit attitude (as reflected by b) is equal in magnitude but opposite in sign to the other.

To the extent that this causal structure does not operate, Equation 12.3 is misspecified and can result in suboptimal prediction and/or faulty causal inference. For example, it may be that discrimination is driven more by implicit disliking of older people (TIA_O) than implicit liking of younger people (TIA_Y); that is, the true regression coefficient for TIA_O predicting DB is larger than the true regression coefficient for TIA_Y predicting DB , making Equation 12.3 misspecified. Even if the IAT is found to predict rejection of older job applicants, unknown is whether this effect reveals the causal impact of anti-old bias, pro-young bias, or some complex combination of the two.

In Blanton, Jaccard, Gonzales, and Christie (2006), we argued that a stronger approach is to measure the two implicit attitudes separately and then to regress any criterion of interest onto both of these measures, allowing the magnitude of the regression coefficients to be determined empirically (see also Jaccard, 2023, Chapter 23, this volume). The component attitudes might indeed combine additively with equal weights per Equation 12.3, or they might combine additively with unequal weights. Or they might combine multiplicatively. The actual integration rule can be tested empirically, if one measures the two attitudes separately, but not if one measures the types of relative attitudes built into the IAT.

We pursued a strategy of measuring distinct implicit attitudes separately, but took it a step further by separating out positive and negative

associations for each attitude object. We did this using a measure we termed a Simple Association Test (SAT; Blanton et al., 2006). We noted that the conceptual model underlying an IAT specifies four types of associations relevant to implicit attitudes. Using the old–young example, they are (1) negative implicit associations associated with older people, (2) positive implicit associations with older people, (3) negative implicit associations associated with younger people, and (4) positive implicit associations with younger people. Separation of positive and negative associations allows for flexible modeling of both the approach and avoidance dynamics that might influence hiring decisions (see Cacioppo et al., 1997; Cacioppo et al., 2012), as well as the differential influence of older versus younger associations on hiring decisions.

Our concern in Blanton, Jaccard, Gonzales, and Christie (2006) was with racial bias rather than ageism. We developed a version of the SAT that separately measured each of the four associations that are confounded in a single IAT score. These were $B+$ (the strength of implicit positive associations for Blacks), $B-$ (the strength of implicit negative associations for Blacks), $W+$ (the strength of implicit positive associations for Whites), and $W-$ (the strength of implicit negative associations for Whites). We found a correlation of 0.22 between $B+$ and $W+$ and a correlation of 0.32 between $B-$ and $W-$, after controlling for GPS. These correlations suggest that people who associate positive characteristics with Blacks also tend to do so for Whites, and people who associate negative characteristics with Blacks also tend to do so for Whites. This correlation pattern suggests the traditional race IAT cannot accurately map individual differences in implicit race associations (see Jaccard, 2023, Chapter 15, this volume, for details). Interestingly, Blanton et al. (2006) found that the only statistically significant

predictor of the outcome was B—, a causal effect that cannot be detected with an IAT but that made theoretical sense.³

In sum, the psychometric model underlying the traditional IAT imposes causal constraints that can be overly restrictive in many if not most criterion-prediction contexts. It also produces a single score that cannot logically sort people based on their true attitudes toward a single object. A more viable psychometric model is needed (see Jaccard, 2023, Chapter 15, this volume).

5. The Metric of the IAT Is Arbitrary, with an Empirically Unverified Zero Point; It Cannot Support Statements of Bias Prevalence

The frequency distributions generated by the IAT webpage are routinely invoked to make assertions about the prevalence of different implicit biases. From the distribution of the IAT designed to measure race attitudes, for instance, researchers commonly assert that roughly 75 percent of respondents possess implicit biases or “automatic preferences” against people who are Black. The key to such claims is that IAT researchers assume that a computed IAT of zero is the dividing line between a person with one set of biases (e.g., an implicit preference for Whites relative to Blacks) versus an opposing set of biases (e.g., an implicit preference for Blacks relative to Whites). In formal psychometric terms, researchers who use an IAT to diagnose hidden biases assume the IAT they are using has a *rational zero*, i.e., that a score of 0 on the IAT maps onto a true zero point of neutrality that separates opposing implicit biases.

In Blanton and Jaccard (2006a), we argued that this practice is problematic because the IAT has an *arbitrary metric*. The issue of metric arbitrariness applies to most psychological inventories, but researchers who employ

psychological inventories in their research rarely assert a rational zero-point and metric non-arbitrariness as the IAT architects do. The assumption of a rational zero for the IAT seems to have been accepted by many others without question – by journalists, the public, and researchers, alike.

Contrary to common practice and interpretations, the rational zero assumption was fully lacking in empirical evidence when the IAT demonstration webpage launched. Indeed, the IAT interpretations and assessments were made available to the public in 1998, months before the first publication on the method (Greenwald et al., 1998). This one publication gave no attention to the zero-point assumption needed to validate the feedback the public was already receiving. In the two decades since the launch of their webpage, there still is no strong empirical evidence to support the assumption. However, there is empirical evidence that calls it into question.

Blanton, Jaccard, Strauts, Mitchell, and Tetlock (2015) re-analyzed studies that linked racial and ethnicity IATs to preferential behavior for Whites over Blacks and minorities. For the behavioral outcomes, researchers had calculated a difference score whereby a score of 0 indicated the behavior was equally favorable toward Whites and minorities, positive scores meant behavior was increasingly more favorable toward Whites than minorities, and negative scores meant behavior was increasingly more favorable toward minorities than Whites. Blanton et al. applied a regression method developed by Greenwald, Nosek, and Sriram (2006) to gain perspectives

³ Following our work, versions of a single object IAT have been becoming more popular (Bluemke & Friese, 2008), but implicit measures that segregate positive and negative associations have not taken hold.

on the zero point, a method that was thereafter rejected by Greenwald et al. (2019), after observing our results. Our results indicated that individuals who had behavioral difference scores of zero across multiple published studies tended to have IAT scores that were indicative of implicit bias against Blacks relative to Whites. Blanton et al. (2015) argued that this data pattern suggested a “right bias” in the IAT, because a frequency distribution of IAT scores will show a higher percentage of respondents with positive IAT scores (i.e., scores to the right of zero) than would be predicted based on people’s actual behavior. When Blanton et al. (2015) recalibrated the IAT to adjust for this bias, characterizations of the prevalence of implicit bias in the studies were substantially reduced, from an average of 83 percent to an average of 28 percent.⁴

A better strategy for evaluating a rational zero point is to draw on strong, accepted theory that permits one to make predictions about how data should pattern themselves as one moves across the dimension of interest and through the true zero point. To isolate where the true zero-point occurs on the metric, the theory should predict a distinct data pattern for the zero-point. If one observes the predicted data pattern for the true zero point at a given number on the scale, then one has a basis for interpreting that number as reflecting the true zero point. Greenwald, Banaji, Rudman, Farnham, Nosek, and Mellott (2002) performed such analyses for a self-esteem IAT and found inconsistent support for an IAT rational zero (the predictions failed about 50 percent of the time). Their tests were problematic, however, because the theory they used to conduct the rational zero tests was tenuous (see Blanton & Jaccard, 2006b).

Again, assumptions of a rational zero point are unusual in psychological research, but many IAT advocates have treated the

assumption with a casualness suggesting it can be commonly assumed. For instance, Greenwald et al. (2002) argued that a rational zero will be found with *explicit* attitude measures if one simply subtracts one rating from the other, stating that “The thermometer and Likert measures had rational zero points because of their construction as difference scores” (p. 12). Differencing two explicit attitude indices does not guarantee that the computed zero maps onto the true (latent) zero point of the psychological construct of interest. Consider the case of a multi-item attitude inventory where the positive items are more positive in nature than the negative items are negative. In this case, someone who equally endorses the positive items and the negative items likely has a positive rather than a neutral attitude, despite the fact when the two subscales are differenced, a score of zero results. A similar dynamic could operate for the average latencies for the compatible and incompatible trials of the IAT (which are differenced).

In response to our critique Greenwald, Rudman, Nosek, and Zayas (2006) noted that with physical quantities (e.g., length, weight, calories), a zero value on a measure “indicates the absence of the quantity being measured”

⁴ Greenwald et al. (2019) not only criticized the method advocated by Greenwald, Nosek, and Sriram (2006) but also offered another version of it (see also Cvencek et al., 2021). The alternative method is flawed because it relies on standardized scores when calculating regression intercepts and fails to account for the complications that such standardization introduces in the estimation of regression intercepts and tests of them. Their revised test as well as the original test by Greenwald et al. (2006) cannot be viewed as strong tests of a true relative zero. However, the results of Blanton et al. (2015) are unequivocal: People who behave *equally favorably* toward Blacks and Whites often have scores on the IAT that signify they have more negative implicit attitudes toward Blacks than Whites despite their egalitarian actions.

(p. 171). This statement refers to an absolute zero, not a rational zero for a bipolar dimension. As well, the IAT does not measure a physical dimension (e.g., time) but rather a psychological dimension (e.g., bias). Although it is true that a difference measure of time, measured in milliseconds, may indeed have a rational zero as an indicator of differences in time per se, such properties should not be assumed when the measure is used to estimate a psychological quality, e.g., an implicit attitude. Even something as simple as a count, like the number of times a child hits another child on a playground, can have an arbitrary origin when it is mapped onto an underlying psychological dimension, like “aggression.” Physical measures that possess rational zero points to quantify physical dimensions lose this quality when they are used as proxies for psychological constructs.

The assertion of a rational zero for the IAT is also dubious given (1) the many method variance confounds that affect the IAT, as described earlier, and (2) the (unspecified) complex combination of four distinct groupings of positive and negative associations for two attitude objects that the IAT assumes is operating. There also are so many substantive and methodological variants of the IAT in the literature, we find it difficult to accept that all of them, regardless of what they measure, produce a metric with a rational zero. Our general point is that assertions that the IAT zero point maps onto true zero points should not be taken at face value, especially for measures used to deliver assessment feedback on a range of socially sensitive topics. The zero point should be tested relative to *some* external standard. The absence of a clear mapping of the IAT zero onto a true zero for relative implicit attitudes undermines the frequent statements in the literature about the prevalence of implicit bias.

6. The Criteria for Classifying People Into Ordinal Implicit Bias Categories on the IAT Website Are Arbitrary

On the IAT website, people are told not just the direction of their bias, but they are also given labels, indicating whether they have “slight,” “moderate,” or “strong” automatic preferences. These labels can have consequence, as people told that they have “strong” implicit biases against minorities might feel greater distress than people who are told that they have a “slight” implicit bias. Unfortunately, these labels are not empirically based. In one of our early papers (Blanton & Jaccard, 2006a), we argued that the criteria for classifying people into the ordinal categories of bias used by the IAT website are arbitrary, because they were chosen without the benefit of any empirical data linking IAT scores to real-world benchmarks. We examined this issue further in Blanton and Jaccard (2008), where we noted a curious shift in scoring practices on the IAT webpage. We observed that in 2002, 48 percent of the untold numbers of individuals taking the online IAT were classified as having a strong “automatic preference” for Whites over Blacks. However, by 2007, that percent had dropped dramatically, down to 27 percent. What happened? Did some historic event occur that caused people to discard their old biases? Was the IAT campaign to promote greater awareness of implicit biases working?

It turns out the change occurred because the IAT architects changed the standards for assigning people to categories, but they did so without any empirical referents. Part of the reason for this shift appears to be the development of the new scoring algorithm, which altered the distribution of scores from what had been used previously to deliver feedback to the public. Prior to 2002, IAT researchers apparently classified people into the categories by transforming raw IAT scores (measured in

milliseconds) into an individualized Cohen's d . The traditional d values of 0.20, 0.50, and 0.80 were then used as the basis for classification. Greenwald, Nosek, and Sriram (2006) pushed back against our claim that their diagnostic practices were arbitrary, suggesting that the Cohen standards lent their diagnostic practice "consequential validity." Interestingly, Cohen himself argued against such uses of his standards. Specifically, Cohen stated that he viewed his standards as arbitrary, context sensitive, and in need of empirical validation. To quote him directly, he stated that offering his benchmarks was "an operation fraught with many dangers" because "my definitions are arbitrary" (1988, p. 12). Cohen also stated that his "proposed conventions were set forth ... with much diffidence, qualifications, and invitations not to employ them if possible," noting that "the values have ... no more reliable basis than my own intuition" (1988, p. 532). Cohen laments that the criteria "were needed in a research climate characterized by a neglect of attention to issues of magnitude" (1988, p. 532) and stressed the importance of conducting research to derive meaningful, evidence-based benchmarks.

Interestingly, in 2007, after the introduction of their new scoring method, the IAT website authors changed their criteria, no longer using Cohen's original criteria but now using, as best as we could determine from personal correspondence with the website authors, scores that mapped onto Cohen's d values of 0.30 (slight preference), 0.70 (moderate preference), and 1.30 (strong preference). Blanton and Jaccard (2008) noted that had the original d criteria been used, instead of a decrease in extreme bias, there would have been an *increase* in extreme bias among IAT respondents; almost 60 percent of the people taking the IAT would have received this label. By altering the benchmarks applied to the new scoring algorithm, the distribution more closely approximated

the distribution of the older algorithm that Greenwald et al. (2006) were rejecting.

We fully recognize the possible need to recalibrate classification criteria when a new scoring method is implemented. However, it became evident to us that the criteria being used were arbitrary and lacked a satisfactory logic model. We personally find it unacceptable to provide the public with feedback about ostensibly hidden implicit preferences (especially given the random error and method variance that pervades the IAT), absent empirical justification for these diagnoses. The distributions created out of this arbitrary process are also used by many IAT researchers to characterize the prevalence of implicit bias in their research reports, media statements, and in legal depositions where the data are used to support accusations of implicit bias in organizations and the minds of individuals.

7. Researchers Who Employ IAT Measures to Predict Discrimination Often Report Their Data in Ways That Suggest Discriminatory Bias, When Discriminatory Bias Was Not Observed

We described earlier how our research reanalyzing published studies found preliminary evidence that common IAT measures might be "right biased." Such bias has implications for the interpretation and reporting of research findings. This is illustrated by one of the first studies where we first noticed the problem, as reported in Blanton, Jaccard, Klick, Mellers, Mitchell, and Tetlock (2009). There we re-analyzed data from an early, widely cited study of behavioral prediction using the IAT that was reported by McConnell and Liebold (2001). These researchers had research participants interact with a White experimenter, then complete a race IAT, after which they interacted with a Black experimenter. Both interactions were unobtrusively

recorded on videotape. Two judges rated the videotapes on the favorability of the interactions, and McConnell and Liebold derived a behavioral preference measure for each participant by subtracting a standardized index of favorability of behavior toward the Black experimenter from a standardized index for the White experimenter. A difference score of zero ostensibly reflected equally favorable behavior toward the White and Black experimenter, whereas a positive difference score reflected more favorable behavior toward the White experimenter. McConnell and Liebold reported positive associations between the behavioral index and the race IAT, suggesting that the IAT predicted racial bias.

Our reanalysis found something quite different. Although 90 percent of the participants in the study had positive IAT scores (which suggests that 90 percent are implicitly biased against Blacks, if the assumption of a rational zero point is accurate), over 70 percent of the participants were rated by judges as acting more *positively* toward the Black experimenter than the White experimenter. Our re-analysis also found that individuals with the highest IAT scores were the *least* behaviorally biased and acted in more egalitarian ways toward the Black and White experimenters. Essentially, as scores on the IAT increased, the tendency to behave in a biased manner against Whites relative to Blacks decreased. This pattern was hidden from readers, because McConnell and Liebold had standardized each of the two behavioral indices before differencing them, thereby forcing both to have a mean of zero and removing the evidence of out-group favoritism in the study.

We, of course, did not suggest in our re-analysis of McConnell and Liebold (2001) that Whites in general treat Blacks preferentially or that this study supports any strong inference about the relative prevalence of in-group versus out-group favoritism of any sort.

We instead pointed to a host of methodological explanations for why this specific difference might have been observed in this specific study. Perhaps the Black experimenter chosen for this study was more outgoing and approachable than the White experimenter chosen for the study. Perhaps participants projected positivity to the Black experimenter to signal that they were not racially biased. Perhaps the participants realized they were in a study assessing implicit bias and so they acted in ways to manage impressions.⁵ Our goal was not to interpret the findings more broadly but rather to consider how incomplete reporting in this seminal study could reinforce arbitrary interpretations of IAT scores that are not empirically supported.

This was not the only instance of deceptive reporting when the data undermined researchers' interpretations. For instance, in another widely cited study, Green et al. (2007) seemingly found that the race IAT was predictive of physician clinical treatment of Whites versus Blacks with thrombolysis. Specifically, Green et al. found differential treatment decisions were made for Whites versus Blacks and that the size of this difference correlated with the race IAT in a way that they suggested indicated discrimination against Blacks. However, Dawson and Arkes (2009) re-analyzed the Green data and found the same pattern of data we observed in the McConnell and Liebold study. Dehon, Weiss, Jones, Faulconer, Hinton, and Sterling (2017) later conducted an extensive meta-analysis of the effects of implicit bias on clinical decision-making in

⁵ This possibility was explored by Heider and Skowronski (2007, Experiment 2), who reported replication of McConnell and Liebold (2001) with social desirability demands reduced. Blanton and Mitchell (2011) reanalyzed data from this study and found that key data had been fabricated and that the original study did not in fact replicate McConnell and Liebold (2001).

medical settings. They found that only two of the nine studies that met inclusion criteria showed evidence of a relationship between the race IAT and clinical decision-making, one of which was the Green et al. study. The other study found a similar result to Green et al., namely that IAT scores suggesting implicit preferences for Whites were associated with selecting less ideal treatment for *both* Black and White pediatric pain patients.

We found this pattern again in an experimental study utilizing a race IAT to predict competitive behavior. Heider and Skowronski (2007, Experiment 1) reported that the race IAT predicted more competitive interactions with a Black opponent in a prisoner's dilemma game. This was technically accurate: when playing against a Black opponent, White participants with higher IAT scores were more competitive than those with lower IAT scores. However, in a reanalysis of their data, Blanton and Mitchell (2011) found that the researchers failed to report that the race IAT also predicted more competitive interactions with a *White* opponent, in the same game. There actually was no racial bias to predict in their study.

Again, we do not think one should draw any inferences about the prevalence or patterning of racial bias in everyday life from these laboratory results. The important point is that researchers who use the IAT to predict discrimination at times report their data in ways that suggest discrimination was observed when it was not. By extension, their reporting suggests that IAT scores commonly interpreted as revealing of prejudicial implicit biases are predictive of discriminatory behavioral biases, even when the data do not support such conclusions. It is our view that, in the future, studies that seek to predict discriminatory behavior from IAT scores should include careful comparative analyses of behavior, to determine if such assertions are warranted.

Attributions of discrimination are complex and study interpretations should reflect this.

8. The Revised IAT Scoring Algorithm Artificially Equates Unreliable Responding with Reduced Implicit Bias

Although this particular issue is unique to the IAT, it has broader implications for the use of scoring algorithms for implicit measures more generally. Recall that Greenwald et al. (2006) developed their D scoring algorithm as a means of expunging IAT scores of known sources of method variance. They settled on their D score method after nominating it as one of "five candidate methods for computing [an IAT score]" (p. 200). Each of the candidates not only had different algorithms, but they also had different ways for handling extreme trial responses, for coding error trials, and for excluding outlier participants. From this heterogeneous bag of possible approaches, Greenwald and colleagues chose the D score as their winner, based on their own *a priori* assumptions of what properties an implicit attitude measure should exhibit. Some of these criteria seem questionable. For instance, candidate algorithms were preferred if they maximized the correlation between implicit and explicit measures. This seemed an unusual choice for a measure that was supposed to reflect associations that are introspectively unidentifiable and outside of conscious awareness. To justify many of their choices, Greenwald and colleagues cited findings from the old algorithm that they thought the new one should replace. This approach seems to us to put the new algorithm at risk of having old problems.

Ultimately, the scoring method they chose introduced a major new flaw in the IAT, which we documented in Blanton, Jaccard, and Burrows (2015). It turns out that for the D scoring algorithm, as reliability of the IAT

improves as reflected by lower levels of trial error, the measure yields a more *extreme* bias score independent of the true bias of the respondent taking the test. The key to understanding this effect is to consider the meaning of trial error in the IAT. For the IAT, trial error refers to random fluctuations in responses from one trial to the next, among trials in the same block (e.g., all compatible trials or all incompatible trials). Trial error is analogous to the random fluctuations one expects for a multi-item attitude survey from one item to the next, despite the fact that all of the items are thought to measure the same construct. With a response latency task, trial error can be caused by such factors as momentary lapses in concentration, fleeting distractions occurring in any given testing environment, brief fluctuations in attention, mood, hunger, and the like.

The more trial error that is present, the more variability there will be in response times from one trial to the next within a given block of trials. With the D score, this trial error is formally incorporated into the scoring of the IAT in the denominator of the index. Blanton, Jaccard, and Burrows (2015) showed mathematically that the D score must range from -2.0 to $+2.0$ and that if two individuals have the exact same average difference in response times between compatible and incompatible trials (which in previous scoring algorithms would have been interpreted as evidence of equivalent implicit biases), they will be diagnosed as having different levels of implicit bias, due to differences in the magnitude of trial error during test taking. As trial error decreases, IAT scores are pushed toward the values of -2.0 or $+2.0$. Interestingly, if there is no trial error for a given respondent (i.e., if their reactions to different trials are consistent), then the D score must equal the extreme, either $+2.0$ or -2.0 , no matter what the average difference in reaction time is between the compatible and incompatible blocks. As such,

with the D algorithm, the more care researchers take to reduce momentary distractions in their experimental settings and the more successful they are at minimizing other sources of random trial error while respondents are taking the IAT, the more they will characterize test takers as highly biased. In contrast, a researcher testing the same participants under less controlled settings will find evidence for lower bias levels.

Although it is true that trial error will rarely approach zero, the property of having the estimated score of an individual difference impacted by inter-item/trial error is psychometrically dubious. It is analogous to using the standard deviation across items in a multi-item self-report inventory of self-esteem to determine that person's self-esteem. It is not enough to sum or average responses to an inventory, the logic goes; we also need to bring to bear the variability of item responses. Blanton, Jaccard, and Burrows (2015) showed in a computer simulation that when the IAT achieves reliability levels of 0.70 or higher in terms of trial error, the distortions in bias characterizations can be substantial, misattributing true bias by up to 45 percent in a population. Moreover, any individual difference variable correlated with susceptibility to trial error will be correlated with IAT scores, independent of true implicit bias. This could cause people with some cognitive styles (e.g., high attention spans) to be labeled as more biased than people with contrasting styles (e.g., low attention spans), independent of their implicit attitudes.

In our opinion, the IAT needs a new scoring algorithm that is better grounded in a formal psychometric model (see Jaccard, 2023, Chapter 15, this volume). This also is true of many of the other implicit attitude measures, although we are not aware of any that share this unusual feature in the newer IAT scoring algorithm.

9. The IAT Predicts Behavior, on Average, No Better Than Attitude Measures from the 1960s and Early 1970s that Caused a Crisis in Social Psychology and Forced Attitude Theorists to Re-Examine the Utility of the Attitude Construct

There have been several meta-analyses of the relationship between the IAT and behavior. Across these meta-analyses, no matter which data-selection rules were followed, no matter how the data were aggregated, and no matter which statistical approach was employed to analyze the data, the mean effect sizes for predicting behavior from the IAT have been small or very small and often not in line with theoretical or common-sense expectations. The most visible meta-analyses are those of Greenwald, Poehlman, Uhlmann, and Banaji (2009) and our own effort (Oswald et al., 2013; see also the exchange between us and the Greenwald team after the publication of our meta-analysis – Greenwald et al., 2015 and Oswald et al., 2015). The Greenwald et al. meta-analysis focused on nine different domains, (1) race (White vs. Black), (2) other intergroup behavior, (3) gender/sexual orientation, (4) consumer preferences, (5) political preferences, (6) personality traits, (7), alcohol and drug use, (8) clinical (e.g., phobias, anxiety), and (9) close relationships. They also used diverse outcomes, behavioral and otherwise. More specifically, the exclusion criteria for the outcomes used by Greenwald et al. were that the outcome could not be another implicit construct, nor could it be an explicit measure of the construct being measured by the IAT. As a result, many of the outcomes included in their analysis were non-behavioral and included judgments, preferences, response latencies measured by computer-based tasks, and neurological activity measured via neuro-imaging techniques. The average correlation

that Greenwald et al. reported between the IAT and outcomes was 0.27. In contrast, our meta-analysis focused exclusively on the race IAT and behavior. We also approached our analysis in a statistically more rigorous way and corrected for many coding errors in the original meta-analysis that had biased estimates upward. For the race IAT, Greenwald et al. reported an average zero-order correlation between the IAT and outcomes of 0.24, whereas we found an average zero-order correlation with behavior of 0.15. A more recent meta-analysis by Kurdi et al. (2019) focused on studies of implicit attitudes, stereotypes, and behavior in the area of intergroup relations, using outcomes they labeled as behavior but that were so broadly defined as to include outcomes that typically would not be viewed as such (e.g., person perception, ideology, neurological measures). Kurdi et al. did not report an average correlation but they reported a 90 percent prediction interval that ranged from $-.14$ to $.32$ with a likely midpoint near 0.10.

None of these correlations are impressive in magnitude. In fact, when viewed in historical context, the correlations have an ironic quality to them. In 1969, Allan Wicker published a highly influential review of the relationship between attitudes and behavior that found the correlation to be so weak that it caused social psychologists to question the relevance of the attitude construct. Wicker reported an average attitude-behavior correlation of 0.26, which is virtually the same as that reported by Greenwald et al. and treated by them as an affirmation of the utility of implicit attitudes in general and the IAT in particular.

The Wicker review generated a much different response from attitude researchers. The attitude construct mostly languished into disrepute until another influential review by Ajzen and Fishbein (1977). These theorists

advanced what has come to be known as the principle of correspondence. Ajzen and Fishbein argued that attitudes are indeed often relevant to behavior but that what is critical for behavioral prediction is that the attitudes and behaviors be measured at correspondent levels. They noted that behavior has four key elements that must be considered when both the behavior and attitude are measured. Specifically, behavior takes a specific form (e.g., criticizing) performed in relation to a specific target or “attitude object” (e.g., a person who is Black or a person who is White) and occurs in a given context (e.g., at the workplace or at a dinner party), at a given point in time or time period (e.g., later that afternoon or over the course of the next month). If one seeks to predict a specific behavior, one must measure the attitude in a way that takes into account all of these components of the behavior, i.e., one measures the attitude toward performing the particular action performed toward the specific target, in the specific context, at that particular point in time. If one does so, then attitudes often will be predictive of the behavior. However, if the measured attitude does not incorporate these elements, such that the attitude measure lacks correspondence with the criterion of interest, then the attitude may or may not be predictive of the behavior. For example, a supervisor’s general attitude toward Black people should not be expected to predict that supervisor’s decision to promote a specific Black manager at a given point in time, in a given work context. This is because the general attitude fails to capture the many facets of the specific target behavior that are operating. Ajzen and Fishbein’s seminal review observed patterns of results consistent with their predictions. Their principle of correspondence has since been reaffirmed multiple times in the research literature. It is a core theorem in attitude research.

The principle of correspondence is not the only reason modern attitude researchers invoke as a basis for attitude–behavior disparities. Ajzen and Fishbein (1977, 1981; Fishbein & Ajzen, 2010) and many other attitude theorists also recognized that (explicit, not implicit) factors other than attitudes can impact and shape behavior and that these variables need to be integrated into attitude theory accordingly. It is now widely recognized that even when attitudes are measured in ways that are correspondent with the behavioral criterion, they can still show weak or no relationships with behavior if these other causal factors dominate. Examples include social norms, self-efficacy, social prototypes, emotions, behavioral intentions, and knowledge necessary to perform the behavior.

In sum, the evolution of modern attitude theory suggested multiple reasons why the IAT mission would likely fail at behavioral prediction (setting aside measurement validity and model misspecification), and recent meta-analyses suggest it has failed if judged on historic standards. Contrary to modern attitude theory, most implicit measures are designed in ways that they measure broad attitudes toward general objects (e.g., Blacks vs. Whites) with none of the attitudinal specificity that has proven consequential in behavioral predictions for explicit attitudes. Nor do IAT researchers acknowledge the operation of multiple behavioral determinants that must be controlled for in order to make correct causal inferences between attitudes and behavior. In the early 2000s, when we saw these dynamics emerging, we predicted to each other the IAT would rarely achieve real-world behavioral prediction above 0.25, and this indeed is what has come to be (and it is true of other implicit attitude measures as well). Rather than push the lessons learned in past attitude work aside, we believe that current work on implicit attitudes should stand on the shoulders of prior science.

10. When Statistically Controlling for Explicit Attitudes in Tests of IAT Predictive Utility, Researchers Routinely Employ Suboptimal and Outdated Measures of Explicit Attitudes, a Practice that Can Inflate Estimates of the Impact of Implicit Attitudes on Behavior

A notable feature of attitude–behavior research with the IAT as well as other implicit attitude measures is its reliance on simplistic explicit attitude concepts and measures, particularly in studies designed to competitively test the predictive utility of implicit versus explicit attitude measures. This puts explicit measures at a competitive disadvantage and can produce the false appearance that implicit measures are needed to account for variance in criteria that explicit measures are unable to account for. The primary shortcoming of the explicit measures used by IAT researchers is that they are outdated. Implicit attitude researchers almost exclusively conceptualize and measure explicit attitudes in ways typical of studies reviewed by Wicker (1969), ignoring the major advances of modern attitude theory from the last five decades. To us, this represents a “straw man” approach, in that researchers pit the IAT against explicit measures that were long ago rejected by attitude theorists. For example, we noted previously the importance of taking into account the four elements of a behavior (action, target, time, and context) when specifying an attitude from which to predict behavior. Such elemental analysis almost never occurs when explicit attitudes are measured in implicit attitude research.

In a prospective study, Blanton, Burrows, and Jaccard (2016) tested for incremental explained variance of the IAT as applied to binge drinking, when explicit attitudes were measured appropriately, along with other known explicit predictors of behavior that

included constructs from the Theory of Reasoned Action and social prototype models of behavioral willingness. The explicit constructs accounted for about 40 percent of the variance in binge drinking behavior in the ensuing two weeks, an effect-size far larger than what is typically reported for explicit attitudes in implicit attitude research. The IAT did not add any statistically significant incremental explained variance to these constructs. Another implicit attitude measure, the affective misattribution test (AMP), revealed about 8 percent incremental explained variance in behavior over and above the classic feeling thermometer measure that is widely used in implicit attitude research. When the modern attitude constructs were statistically controlled instead, incremental explained variance for the AMP dropped to 2 percent.⁶ This example illustrates how weak and outdated methodological practices can upwardly bias estimates of the influence implicit measures have on behavior. There exist a smattering of studies that have sought to incorporate constructs related to the Theory of Reasoned Action (TRA) when evaluating incremental explained variance by implicit attitudes (e.g., Payne et al., 2016), but most of these studies (including Payne et al.) have not been faithful to proper implementation of TRA.

A counterargument to the issue of specificity/correspondence that has been suggested by critics of our work is that, because the IAT is designed in such a manner that it will lack correspondence with specific criteria, it is only “fair” to pit it against non-correspondent explicit measures. By showing incremental improvements in prediction using both a non-correspondent IAT and a non-correspondent

⁶ The AMP, like the IAT, is confounded with other constructs, so the source of this incremental variance is not unambiguous; see our point 11.

explicit measure, the true power of implicit attitudes is demonstrated. Our response is that two wrongs do not make a right. It is not the case that implicit measures like the IAT predict variance that eludes our ability to predict via self-report. Rather it is simply that implicit measures like the IAT predict variance that might be predicted easily via self-report – if self-report questions are competently constructed.

11. The IAT Is Confounded by Many Influences Other than Implicit Attitudes: As Such, It Can Support False and Counterproductive Narrative About Implicit Bias

As noted, the IAT is artifactually confounded with many variables, not only method variance artifacts but also substantively meaningful artifacts like explicit attitudes, intelligence, similarity, neuroticism, and age, among others. As a result, behaviors that some might attribute to implicit bias could be due to these other influences. Instead of simply assuming that an act of discrimination must be due to implicit bias, we think researchers should ask whether the effect could plausibly be attributed instead to explicit bias. If White people are less likely to help a Black person in need of roadside assistance than a White person in need of assistance, is this because of hidden associations about race that are “introspectively inaccessible” or is it because of a racist explicit cognition that rapidly and temporarily becomes salient in working memory, that “stopping for a Black person might be dangerous”? If a manager does not hire a well-qualified obese applicant for a position, is it because of hidden associations about obese people that are largely unknown to the manager but that “leaked” into this person’s decision, or is it because of a prejudicial explicit cognition that rapidly and temporarily

became salient in working memory at the time of applicant evaluation, that “I don’t really want to be around fat people” or “fat people won’t be accepted by their co-workers”?

Nothing in the evidence presented can answer this question. However, for many, an assumption of implicit bias has become the default account for documented biases. IAT research has been notable for its focus on simple correlations between IAT scores and behavior, while ignoring plausible confounds. Indeed, this is the case in the studies considered in the Greenwald et al. (2009), Oswald et al. (2013), and Kurdi et al. (2019) meta-analyses examining the relationship between implicit associations and behavior. Each of the studies estimated zero-order correlations between implicit attitudes and behavior, while generally ignoring confounds and alternative explanations and giving minimal attention to the assessment of explicit constructs other than a crude measure of explicit attitudes. In the absence of rigorous research that competitively tests implicit bias effects against explicit constructs and other known confounds, one simply cannot be certain that implicit bias is the causal agent.

Routinely invoking implicit attitudes and cognitions as explanations is particularly troublesome in the case of discrimination, because doing so runs the risk of obscuring the many explicit biases that society needs to address to lessen it. We, as a field, may actually be setting back efforts to address discrimination, including the misallocation of resources and the wasting of time developing and evaluating faulty interventions to address “implicit bias” when harmful discrimination arises from other causes. For example, Forscher and colleagues (2019) reported a meta-analysis to determine the effectiveness of interventions to reduce discrimination through implicit bias reduction. They analyzed data from 87,418 participants taken from 492 studies and found

little evidence that the interventions had meaningful effects on behavior, much less on implicit bias, *per se* (see also Lai et al., 2016). In our view, the presence of 492 interventions using an approach that, on average, fails to bring about meaningful behavioral change represents a non-trivial misallocation of resources in the social scientific pursuit of interventions to reduce discrimination.

In our experience, however, discriminatory behavioral patterns are routinely attributed to implicit bias as the default. For instance, Jost and colleagues (2009) published an article under the lead, “The existence of implicit bias is beyond reasonable doubt.” They listed ten studies that they suggest “provide a compelling empirical case for the predictive validity of implicit bias measures, including the IAT” (p. 46). The studies either failed to control for explicit attitudes or did so by measuring explicit attitudes in ways that reflected pre-1970 attitude theory. One of their ten studies was Green et al. (2007), which we noted earlier actually found out-group favoritism.

Summary of Our Research Program

Like the many attempts to measure unconscious, subconscious or “hidden” attitudes that preceded it, the IAT, in our opinion, has not lived up to expectations. It generally lacks reliability and convergent validity. It has considerable method variance. Its metric is arbitrary, but researchers routinely rely on it to advance conclusions that require an inventory with a non-arbitrary metric. Its zero point is also arbitrary and possibly biased, although researchers routinely advance conclusions that necessitate the assumption of a rational zero point. It incorporates an algorithm that scores more reliable people as more biased. Research on this measure cannot support statements about the prevalence of implicit bias. Its correlations with behavior generally are low and there is insufficient research to determine if most

correlations are driven primarily by implicit attitudes or by a host of confounds, including constructs measurable via self-report. The IAT, as well as most other implicit attitude measures, lacks a viable psychometric model. It confounds different types of association strength and attends to relative attitudes in ways that restrict the exploration of plausible causal dynamics, i.e., it imposes a model that is likely misspecified, in most common applications.

One place where we agree with IAT researchers (Greenwald & Banaji, 2017) is that the study of implicit attitudes has been characterized by a fresh, creative approach to the attitude construct. In that sense, we are supportive of this line of research. However, we also see the initiative as a step backward, because it tends to ignore lessons learned from traditional attitude research conducted over the past forty years. After over twenty years of research on the approach and now with limited support for early claims and predictions, we think the field (and society) needs to take stock of the utility of the IAT. We have sought to bring healthy scientific skepticism to the implicit attitude research table. In all of our research, we have offered constructive suggestions for ways to improve the problems we have identified. We urge readers to carefully examine our body of work (as well as that of others who adopt other perspectives) and form their own judgments.

Alternative Conceptualizations of Implicit Measures

Based on past experience, we anticipate that IAT advocates will counter-argue the weaknesses identified in our research program, shifting from argument to argument as needed to justify current practice and perhaps criticizing us personally as well. We do not think it possible to anticipate and react to every argument that might be deployed, but some responses deserve closer scrutiny because they

have been advanced in one form or another. One argument prioritizes methods over theory, arguing that theoretical frameworks of the sort we embrace threaten creativity and discovery. Another argument invokes conceptualizations of implicit attitude measures as instruments that assess general attitudes or broad dispositions, rendering concerns about attitude specificity moot. A third argument posits that measures of implicit attitude reveal biases at the level of the aggregate, not the level of the individual. The fourth argument seeks to defend the IAT on nonscientific grounds. We consider each in turn.

1. Methods as Theories

Greenwald and Banaji (Banaji, 2001; Greenwald & Banaji, 2017) have argued that the traditional orientations we and others bring to the implicit attitude field threaten to stifle creativity and limit discovery. This view posits that the genius of the IAT and related measures is found in their tendency to defy old ways of thinking. Implicit attitude research challenges not just common views on attitudes but also on measurement. It is in the creativity of researchers building new measures that new psychological phenomena will be revealed. Measures of implicit bias have the potential not just to teach us about the origins of bias and discrimination but also about the nature of human decision-making more generally. By contrast, the tired, conventional ways that the two of us approach methods and theory will lead the field down sleepy, familiar roads. Needless to say, we do not agree with these suppositions.

The argument that IAT researchers make is that there is value in placing methods before theories; that creative development of implicit bias measures leads to creative breakthroughs in implicit bias theory. One risk researchers take in adopting this philosophy, however, is that implicit bias research might become dominated by many small theories, each designed

to explain responses to a given implicit measure and how that measure differs from other measures, error variance and all. Lost will be theories that address the more fundamental psychological constructs that underlie human judgment and decision making. As an example, the quad theory of the IAT (Conrey et al., 2005), for example, was offered to explain how people make responses to the IAT, providing an accounting of sources of systematic error of the measure. But every implicit attitude measure contains systematic error. If this logic is expanded out, we likely will be left with theories that are as much about systematic error as they are about implicit attitudes.

For psychometric reasons, we think it important to understand and build theories that help explain systematic differences in method-specific error variance, so that the sources of systematic error can be addressed when constructing measures. However, a focus on the methods by which implicit attitudes are assessed can be counterproductive if it draws attention away from substantively important phenomena. There is precedent for misdirection of this sort. Cognitive dissonance theory was initially tested using the classic forced compliance paradigm (Festinger & Carlsmith, 1959). Researchers became as enamored with the method as the theory and so for decades, the study of cognitive dissonance essentially became a study of the forced-compliance paradigm. In the process, the theory narrowed around the method.

We see similar forces at play in the implicit attitude field, where many implicit bias researchers are stalking method variance. Although they state that they seek to gain a better understanding of implicit bias and how it might impact consequential judgments in daily life, their attention is turned toward method variance specific to different measures of implicit bias. The situation has been exacerbated by the lack of strong conceptual definitions of implicit attitude and implicit bias. For

instance, Greenwald and Banaji (1995) originally defined implicit attitudes as “introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects” (p. 8). This definition has ambiguities and there is nothing in it that would prioritize the IAT over almost any other implicit measure. However, just six years after this description of implicit attitudes, Banaji (2001) had largely moved from expressing interest in implicit attitudes generally to an interest in determining the meaning of the IAT. She argued researchers should set aside what she termed the “validity problem” (p. 132) to allow IAT researchers time to learn what the IAT means. She had faith this measure meant something important, a faith that preceded validity assessments. This same faith was also evidenced by the IAT demonstration webpage, which was brought to the public before any validity assessments were conducted. IAT researchers apparently felt their measure revealed something important, and they advanced this view before their claims could be evaluated.

To their credit, IAT researchers acknowledge their early fascination with their measure, and they justify their preoccupation based on the view that this assessment method has expanded theory. In 2012, Greenwald argued “there is nothing so theoretical as a good method” (p. 99) and in 2017, Greenwald and Banaji (2017) advocated that the “sequence of introduction of innovative research methods producing previously unobservable findings that in turn demand new theory occurs regularly in scientific work” (p. 869). With their approach, new theories arise as serendipity when researchers pursue new methods.

We acknowledge that there certainly are cases in which serendipity has advanced theory or produced other scientific advances. From penicillin, to X-rays, to gram staining, science

often stumbles on discovery. We think these examples affirm the view that a curious method can, in fact, lead to important scientific advances. However, we do not think that methodological serendipity is a sound philosophy for implicit attitude researchers to embrace as a *modus operandi*. Serendipitous science, by its nature, is hard to harness and prone to frustrate. Implicit measurement researchers working in this spirit also risk narrowing their field of inquiry, from seeking to explain human behavior to seeking to explain responses on a measure, per the example of the forced-compliance paradigm. In contrast, a strong theory produces many predictions, and it can be tested through a wide range of methods, in pursuit of greater understanding of a broad array of social phenomena.

An intriguing measure certainly might produce a bold new theory that will expand the field, but not if it draws attention to the idiosyncratic ways different people react to that measure. At present, so many methods of measuring implicit attitudes occupy the field that we fear the challenge of understanding how people respond to situations they confront in daily life as a function of implicit attitudes and biases is losing focus. We see signs of this in Brownstein, Madva, and Gawronski (2020), who argued that unexplained divergence between different implicit measures offer theory-expansion opportunities:

Ultimately, the relatively low correlations between implicit measures is not so much an anomaly that threatens the field as it is a pedestrian empirical finding, which has begun to be explored (e.g., Bar-Anan and Nosek 2016; Moran et al., 2017). We would certainly welcome further theory-based and experimental investigation into the mechanisms explaining performance on these distinct measures (cf. Bishara and Payne, 2009; Conrey et al., 2005; Payne et al., 2010), which should illuminate when and why they come apart, and, in turn, which

specific measures are most apt for which specific aims, contexts, psychological processes, and behavioral predictions. (p. 15)

As noted, we understand the importance of having a strong theory linked to measurement efforts and the need for researchers to understand the measures they are using. However, we endorse the more traditional approach, where clear conceptual definitions and the positioning of constructs in a broader conceptual system guide the development of measures. This seems particularly important in the implicit attitude literature, where methodological norms can encourage researchers to develop inventories based on whatever new technologies strike them as interesting. The result will be the proliferation of theories to explain human behavior that are, in fact, theories that explain method variance.

2. General Implicit Attitudes and Global Implicit Dispositions

Another response to our critique, one that we have encountered in various forms in reviews of our work, is that the implicit attitudes measured by the IAT and other current implicit measures reflect global attitudes. As such, they should predict broad-based behavioral patterns, even if they are ill-suited to predict isolated or specific behaviors of substantive import. Prescient to this view of how implicit measures might operate is the theorizing on general (explicit) measures advanced by Fishbein and Ajzen (1975; Ajzen & Fishbein, 1977) some forty years ago in their classic discussion of general attitudes and multiple-act criteria. These theorists posited that any single behavior likely will be shaped by situational contingencies, limiting the predictive utility of a global attitude estimate. Across a wide range of behaviors, however, these situational contingencies can cancel one another

out when forming an aggregate behavioral index, such that a global attitude estimate will be predictive of behavior in the aggregate. This orientation also can be used to understand the causal influence of general dispositions (e.g., self-esteem, neuroticism) on behavioral aggregates. General dispositions tend to be weak predictors of specific behaviors but stronger predictors of general behavioral tendencies and outcomes (Swann, Chang-Schneider, & Larsen McClarty, 2007).

We have applied Fishbein and Ajzen's (1975; Ajzen & Fishbein, 1977) framework in our analyses of the IAT and implicit measures, elaborated in greatest detail in a paper titled "*A Theory of Implicit Reasoned Action*" (Jaccard & Blanton, 2007). An argument we made there was that implicit measures like the IAT are limited relative to self-reports, because they do not lend themselves to the prediction of specific acts of prejudice. Bodenhausen and Petsko (2023, Chapter 20, this volume) propose much the same framework as our own, but they use it to argue for new standards against which the IAT and measures like it should be judged in the future, based on their ability to predict broader behavioral patterns:

Most of the existing implicit measures of race bias seek to assess a global kind of bias that applies broadly toward many or most members of the racial group in question, in a variety of different contexts or circumstances. However, most tests of the predictive power of implicit measures involve only measuring one particular kind of behavior, directed at particular targets in one particular situational context. The likelihood of robust prediction under these circumstances is limited. As such, the predictive value of very general indicators of racial bias would better be evaluated in terms of their relation to more general patterns of intergroup behavior, as assessed in multi-act criteria that collapse across different actions, targets, contexts, and time parameters. (Manuscript, p. 15)

This same logic was part of our reason for predicting that meta-analyses would show weak behavioral prediction (as turned out to be the case). From the perspective Bodenhausen and Petsko (2021) seek to advance now, however, researchers might nonetheless view an IAT designed to measure implicit bias as useful, even if it yields only weak or modest prediction of specific behaviors.

Greenwald, Banaji, and Nosek (2015) have embraced such an argument in a reinterpretation of meta-analytic studies showing weak attitude-behavior correlations with the IAT. They take the logic a step further by predicting that the many small effects aggregate in ways that have consequential effects on individuals and society. We think that by acknowledging the weak performance of the IAT in studies designed to predict specific behavioral acts of discrimination, the arguments in Greenwald, Banaji, and Nosek (2015) represent progress. This language contrasts with the claims in their original meta-analysis, where Greenwald, Poehlman, Uhlmann, and Banaji (2009) framed the weak predictive utility of the IAT optimistically (and inaccurately) as “relatively high,” (p. 32) when compared to the predictive utility obtained from (impoverished) self-report measures.

Ultimately, the predictive utility of global measures like the IAT remains an empirical question. Despite two decades of research on this measure, evidence for this assertion is limited. Kurdi et al. (2019) analyzed a handful of studies in their meta-analysis and argued that the IAT performs better when measurement correspondence is high, but their selection of studies and their coding of studies is far from rigorous. Bodenhausen and Petsko (2023, Chapter 20, this volume) go no further than to cite this same study and also note the limited number of studies to address their prediction. It may indeed be the case that this optimistic frame on the IAT will

be supported by the data, but at this point, it is speculation.

Also speculative is the argument that the IAT predicts behaviors for individuals that will aggregate in ways that matter. We fully accept that individual acts of discrimination and disadvantage accumulate and have harmful effects on members of marginalized groups and society, but that is not the question. The question is whether the IAT or measures like it help researchers predict broad-based behavioral patterns that have harmful consequences through aggregation. Greenwald, Banaji, and Nosek (2015) offer no empirical evidence to show that the weak-to-moderate correlations observed in IAT studies have this quality. As is common with earlier defenses of this measure, this new assertion again affords their measure an assumption of relevance that it has not earned empirically. We think the dispassionate and scientifically responsible view is that, until and unless confronted with empirical evidence to the contrary, we should assume the null, that this measure has unknown utility.

There also are sound theoretical reasons to anticipate that the IAT and measures like it will fail with respect to this more limited role as a global measure that predicts behavioral patterns. This is because the IAT and measures like it do not have available to them the same assessment strategies that self-report measures use to tap global attitudes or broad psychological constructs. With explicit measures, there are two dominant strategies used to measure global constructs, and these do not import well to the logic or methodological practice of implicit measurement.

One strategy is to ask respondents explicit questions designed to elicit reflection on the global attitude or broad disposition of interest. Consider self-esteem, which is typically viewed as a global attitude toward the self. The first question in the most commonly used self-esteem instrument, the Rosenberg (1965)

self-esteem measure, asks respondents to rate their agreement with the statement, "On the whole, I am satisfied with myself." To generate a rating for this question, respondents likely reflect on autobiographical memories available to them or on formed opinions they have generated over a lifetime of observations and that also are salient to them at the time of assessment. The hope with such questions is that, from this process of self-reflection, an informed rating will be made.

This strategy cannot be imported to implicit attitude measures. By design, these instruments avoid deep probes of memory and opinions that explicit measures pursue. To the contrary, the goal with each trial in an IAT and other implicit measures like it is to have respondents react quickly to the stimuli they encounter without deep reflection, so as to allow underlying hidden associations to manifest themselves. We simply do not know if the particular IAT tasks can plumb the depths of self-esteem or constructs like racism.⁷ What is clear is that the IAT task does not have the same inherent face validity as the explicit measure designed to elicit deliberate self-reflection. It is not *prima facie* clear that a small sample of pictures and a small set of descriptive adjectives can perform a deep dive into psychological tendencies.

However, there is a second strategy used to measure global attitudes or other broad constructs via self-report. This involves asking many questions of the individual that focus on the many different ways a construct might manifest itself in everyday life. This larger set of questions is then combined to create a single score that represents an estimate of the person's global position on the construct. The Coopersmith (1987) Self-Esteem Inventory (CSEI) employs this approach. Respondents rate their level of agreement or disagreement with such questions as, "I'm easy to like," "I'm pretty sure of myself," and "I'm proud of my

work." Any one of these items alone would likely tap into only a small slice of their global self-regard, but the CSEI uses fifty-eight such items and aggregates them to triangulate the broader construct.

As with the first strategy for measuring global constructs, this second strategy does not translate well to the IAT or other similar implicit measures. The IAT measures only a limited number of construct facets in a single test built around a small set of interchangeable stimuli (e.g., Black and White faces, positive and negative words). If one conceptualizes implicit attitude as a global construct that can influence many behaviors or the same behavior across many behavioral instances, then a stronger approach is to develop new measures that account for the wider range of manifestations of the global attitude, in the spirit of a questionnaire measuring the many different ways an explicit construct might be exhibited.

In summary, the notion that the IAT or other implicit attitude measures meaningfully tap into global attitudes or dispositions that are introspectively hidden but nonetheless consequential, in our opinion, needs both a stronger evidence base and a stronger theoretical rationale. The basis for expecting the IAT to reflect a broad construct that impacts behavioral patterns has not been elaborated with any degree of clarity but rather appears to our eyes to be another attempt to dismiss a reasonable concern about it (i.e., that it cannot be constructed in a manner that has correspondence to specific criteria and as a result has weak criterion–prediction validity).

⁷ Burhmester, Blanton, and Swann (2011) examined this question and concluded the IAT cannot do so in the case of self-esteem.

3. Bias in Crowds

The move toward aggregation is a means of saving the IAT measure by changing assumptions about the nature of the construct being measured. The third argument against our critiques follows in this same vein but changes the nature of the construct in a different way. Specifically, it abandons *any* attempt to predict the behavior of individuals, either single behaviors or behavioral patterns. Instead, the new focus is on predicting the behavior of broad groups of people at the aggregate level. Such is the approach of Payne and colleagues (Payne et al., 2019; Payne et al., 2017). Payne argues that implicit attitudes likely reflect broader contextual influences on individuals who share contexts so that implicit attitude scores aggregated across members of groups will be probative about life in that group context. He argues that the implicit bias “score” of a group (as reflected by the group IAT mean) can reveal the amount of biasing influence exerted on the group.

Although this theoretical framework may have merit, it does not eliminate the need for a valid measure, and this is one of the questions at hand. If an IAT is constructed such that it cannot logically sort people based on their true individual attitudes, it cannot logically sort aggregates on their true group attitudes, unless the systematic error cancels out when aggregated across individuals (which is unlikely). An arbitrary metric does not become non-arbitrary when averaged. A measure that has “right bias” does not become unbiased when averaged. A measure that is contaminated by multiple sources of systematic error does not become uncontaminated by that systematic error when averaged. The interpretation of an average IAT D score of -1 is no less ambiguous than an individual D score of -1 .

Nevertheless, Payne, Vuletic, and Lundberg (2017) argued that validation

efforts for the IAT should focus on group-level modeling rather than individual level modeling. They examined the correlates of mean race IAT scores for different geographic regions in the United States and reported that mean race IAT scores tended to be higher in states and counties that had stronger historic ties to slavery (as measured by the percentage of the population that was enslaved prior to the Civil War). They interpreted this pattern as evidence that regions with stronger ties to slavery today have higher levels of structural and/or systemic racism, making prejudicial associations more accessible for individuals living in those states.

This interpretation of implicit measures has non-trivial implications. It suggests that over the past twenty years, researchers who have been seeking behavioral validation of the IAT and similar measures have been fighting a losing battle. The “bias in crowds” focus also calls into question the common use of this measure as an assessment tool, that is, as an instrument that can reveal to individuals their hidden biases. A group-level interpretation not only shifts attention away from the study of individuals, it also leads to an inability to empirically build and validate theories of the determinants of individual-level discrimination. For example, it is well known that causal relationships that operate at the group level do not necessarily generalize to the individual level, a phenomenon known as the ecological fallacy (see Jaccard, 2023, Chapter 15, this volume). We believe that this attempt to salvage the IAT and similar implicit attitude measures comes at too great a cost, namely giving up on a strong science of individual racism and intolerance.

4. Non-Serious Critiques

There is a final category of response to our critiques that, for lack of a better term, we

simply call non-serious. These are responses that fail to engage our critiques in any substantive manner, other than to find a means of dismissing them or us out of hand. We have sought to engage implicit attitude researchers in a set of concerns that we believe derives from an internally consistent logic and that adheres to empirical standards that we apply in the evaluation of research on any topic, including our own research. In contrast, we have sometimes found that critics of our work play argumentative sleight of hand, often to the point of self-contradiction. In Blanton et al. (2007, p. 407), for instance, we document how Nosek and colleagues criticized us for taking a position that they themselves advocated in an earlier paper. This type of contradiction also arose when we examined the IAT assumption of a rational zero point. Specifically, in Blanton et al. (2015), we used a regression method suggested by Greenwald, Nosek, and Sriram (2006; see p. 58) to evaluate zero-point bias and found results counter to the operation of a rational zero. Later, Greenwald et al. (2019) attributed the test to us and argued against its use on the grounds that it was internally inconsistent. They stated: “Blanton et al. (2015) offered a regression analysis method” (p. 33) and then criticized us for using the very test they originated – only to offer a new flawed test. A similar turn was made by Greenwald, Banaji, and Nosek (2015). They falsely characterized and roundly criticized the data aggregation and analytic methods we employed in the meta-analysis of Oswald et al. (2013). However, later Greenwald and Banaji presented this same approach as their “analytic strategy” in Kurdi et al. (2019, p. 575). We could point to many other contradictory shifts we’ve observed over the years but feel the above make the basic point.

Another class of non-serious criticisms has been to question our motives or character.

This has sometimes been subtle, as when Greenwald, Nosek, and Sriram (2006) stated they were “unable to resist speculating” (p. 60) why their measure attracts attention from researchers like us, after which they suggested it might be linked to the “unsettling reaction” (p. 60) that IAT feedback generates in others. Other times, it has not been so subtle, as in an article in *New York* magazine. One of the IAT architects was so clear in questioning our motives and character that it caused the reporter, Singal (2017), to comment that “it’s a rather big deal for the chair of Harvard’s psychology department to accuse researchers who are engaging in thoroughly commonplace methodological critiques of being animated by racism and possibly mental illness.” We also have grown accustomed to researchers openly speculating that we are driven by conservative ideological motives, or as (Jost, 2019) suggested, “for ideological reasons, including motivated system defensiveness” (p. 14).

We have both been around long enough to have seen our share of bitterness and unfairness in “scholarly” exchanges. However, with the IAT, the level of personal attacks has surprised even us. We are more than open to fair criticisms of our science and urge readers to scientifically evaluate our empirical work first-hand.

In Lieu of Flowers, Science

Science by its nature is meant to learn from its mistakes, in different ways at different times. Sometimes failure sparks innovation. Early attempts at measuring attitudes were based on assumptions that global measures of attitude would predict specific behaviors, but failures of this assumption promoted conceptual and methodological advances that have made strong attitude-behavior relations attainable in many if not most behavioral research contexts. Other times, failure sparks

acceptance of the failure and redirects attention toward more profitable pursuits. For example, we reviewed various early attempts to measure hidden attitudes that have now been mostly or completely dropped by attitude researchers, and we view these redirections as signs of healthy science.

Over twenty years after its introduction, there is mounting evidence that the IAT has failed and science must choose how it will respond. The psychometric properties of this measure are far below par for the behavioral sciences, meta-analyses are failing to yield promising evidence of criterion prediction, measurement confounds are well-documented, interventions based on the underlying theory are failing, and now multiple researchers allied to this measure have felt it necessary to acknowledge (and refute) the growing perception that the IAT is “dead” (Jost, 2019; Kurdi & Banaji, 2017). So, what will be the response?

In the short term, we expect the response will be continued interest in the IAT, despite its scientific death. The economist Paul Krugman (2020) coined the term “zombie ideas” to characterize beliefs that have been “killed” by empirical evidence but that nevertheless remain animated as a result of common and deeply held beliefs that they must be true. We see signs that the scientific value of the IAT has now become a zombie idea. Researchers have killed it time and again, yet it thrives. We do not know for certain, but we suspect there are many factors contributing to this measure’s resilience to its own death. For one, it has filled a niche that many in our field have wanted to have filled. There is a “truthiness” to this measure, in that it feels true based on beliefs that many hold. We hazard a guess that most social psychologists (ourselves included) and many in the general public believe that bias and discrimination pervade society. Many believe (as do we) that most people

harbor prejudicial views that they don’t fully understand or acknowledge. We, like many people, think there is value in locating and confronting one’s hidden wells of intolerance. We also suspect that many if not most people who wish to promote social justice (ourselves included) have at some point turned their focus inward to interrogate their own beliefs and values, to try to find and confront their biases. The IAT seems to validate these ways of thinking, by backing them with the feel of science.

We suspect there also are other dynamics at play that have resulted in the continued popularity of the IAT. One factor is the way the IAT architects marketed this instrument heavily early on, through the press and their webpage. Their premature claims and predictions and their early distribution of this measure for purposes of self-assessment might now make it harder for researchers to openly acknowledge reasonable concerns that have arisen. This is not just an issue for the primary architects of this measure. Many social psychologists promoted this measure and its utility in their classes or worked IAT exercises into their textbooks. Many funding agencies have supported this enterprise with tax-payer dollars. Rejection of the IAT on scientific grounds could now tarnish the reputation of social psychology.

Finally, we think the measure has survived simply because it intrigues a lot of people, research psychologists included. Journals are filled with studies that essentially ask the question, “What does the IAT mean?” Explanations vary from one analysis to the next – often contradicting one another – but they proliferate. Bodenhausen and Petsko (2023, Chapter 20, this volume) suggest IATs tap global attitudes that exert influence across situations and settings, whereas Jost (2019) suggests they “measure a particular attitude at a given time in a specific social context”

(p. 14), whereas Payne and colleagues (2017) argue that they do not measure qualities of individuals but rather features of their shared social context. These are very different views on what the IAT measures, and all represent departures from the IAT starting point, but each helps the IAT transcend death.

Also common across each of these perspectives and others are calls for “more science” to help us better understand what IAT scores mean. The assumption driving these calls is that the IAT must mean something. This is a view that we think researchers should now question. The most striking example that we have seen, of the IAT-must-mean-something orientation, is the chapter by Jussim and colleagues (2023), Chapter 13 in this volume. Jussim and colleagues present an excoriating review of IAT research and logic. They argue that (1) “what the IAT measures remains unclear, even after twenty years of research,” (2) “it is likely that IAT scores explain only a modest portion” of the variance in discrimination, and (3) completely eliminating the biases the IAT measures would likely exert only “minimal” influence on discrimination. Nevertheless, they conclude their chapter by arguing that, despite its limitations, the IAT should not be abandoned, but that, even after twenty years, much more research is needed to fully understand what the IAT measures and explains. By contrast, criticisms of the IAT on almost any ground typically are greeted with swift rebuttals and rejoinders.

For these and other reasons, we think the current phase of current interest in a dubious measure will continue for some time. But we hold out hope that logic and data eventually will prevail and so it seems reasonable to ask what might come next. One possibility is for researchers to confront and constructively engage this failure, following the example of attitude researchers from the 1970s. Our work has often advanced critiques of this

measurement approach, but we have always striven to point to better strategies that researchers can adopt to build stronger implicit measurement theory and methods. These involve but are not limited to the development of viable psychometric models prior to the construction of measures, the embrace of measurement correspondence, the pursuit of assessment techniques less contaminated by random error, and so on (see Jaccard, 2023, Chapter 15, this volume). We suspect such advancement will require that the current wave of measures not just be modified but retired.

Regardless, we hope that efforts to improve or replace current implicit measures will not distract researchers away from the pursuit of stronger *explicit* measures. We live in a time when better explicit measures of bias are sorely needed. We truly believe that implicit bias is a problem in society and that it is to society’s benefit that we better understand and confront this social problem. However, we also think that recent history has revealed an urgent need for psychology to promote a healthy science of explicit bias research. We say this because we believe that much of the bias that currently threatens society is consciously held. Recently, membership in hate groups reached an all-time high (Southern Poverty Law Center, 2019); hate crimes also are on the rise (Oudekerk, 2019). These spikes are set against the backdrop of particularly shocking mass killings, where shooters targeted “Mexicans” in El Paso, African Americans in Charleston, and gays and Latinos in Orlando, and they follow the more recent murder of George Floyd. Throughout all of this, many social psychologists have been focused on trying to figure out “what the IAT means” and advancing what now appear to be ineffective strategies for reducing discrimination, because these approaches target the types of prejudice the IAT and other

measures like it ostensibly measure. A recent comment by Greenwald (2020), the primary architect of the IAT, in an interview with *Knowable Magazine* about programs to reduce implicit bias was particularly striking to us:

I see most implicit bias training as window dressing that looks good both internally to an organization and externally, as if you're concerned and trying to do something. But it can be deployed without actually achieving anything, which makes it in fact counterproductive. After 10 years of doing this stuff and nobody reporting data, I think the logical conclusion is that if it was working, we would have heard about it. (Greenwald in Mason, 2020)

The two science-based meta-analyses on changing implicit attitudes referenced earlier (Forscher et al., 2019 and Lai et al., 2016) reach a similar conclusion as Greenwald about the documented overall ineffectiveness of implicit-based change strategies. At this time in history, the biases that most concern us are explicit and structural in character, and it is in these domains that we believe strong scientific advances are desperately needed.

References

- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84(5), 888–918. <https://doi.org/10.1037/0033-2909.84.5.888>
- Ajzen, I., & Fishbein, M. (1981). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, J. S. Nairne, I. E. Neath, & A. M. Surprenant (Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder*. Washington, DC: American Psychological Association, pp. 117–150.
- Blanton, H., Burrows, C. N., & Jaccard, J. (2016). To accurately estimate implicit influences on health behavior, accurately estimate explicit influences. *Health Psychology*, 35(8), 856–860. <https://doi.org/10.1037/hea0000348.supp> (Supplemental)
- Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist*, 61(1), 27–41. <https://doi.org/10.1037/0003-066X.61.1.27>
- Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist*, 61(1), 62–71. <https://doi.org/10.1037/0003-066X.61.1.62>
- Blanton, H., & Jaccard, J. (2006c). 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(1), 155–166. <https://doi.org/10.1037/0033-295X.113.1.155>
- Blanton, H., & Jaccard, J. (2006d). Postscript: Perspectives on the reply by Greenwald, Rudman, Nosek, and Zayas (2006). *Psychological Review*, 113(1), 166–169. <https://doi.org/10.1037/0033-295X.113.1.166>
- Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, 34, 277–297. <https://doi.org/10.1146/annurev.soc.33.040406.131632>
- Blanton, H., Jaccard, J., & Burrows, C. N. (2015). Implications of the Implicit Association Test D-transformation for psychological assessment. *Assessment*, 22(4), 429–440. <https://doi.org/10.1177/1073191114551382>
- Blanton, H., Jaccard, J., Christie, C., et al. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology*, 43(3), 399–409. <https://doi.org/10.1016/j.jesp.2006.10.019>
- Blanton, H., Jaccard, J., Gonzales, P. M., et al. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, 42(2), 192–212. <https://doi.org/10.1016/j.jesp.2005.07.003>
- Blanton, H., Jaccard, J., Klick, J., et al. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, 94(3), 567–582. <https://doi.org/10.1037/a0014665>

- Blanton, H., Jaccard, J., Strauts, E., et al. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, 100(5), 1468–1481. <https://doi.org/10.1037/a0038379>
- Blanton, H., & Mitchell, G. (2011). Reassessing the predictive validity of the IAT II: Reanalysis of Heider & Skowronski (2007). *North American Journal of Psychology*, 13(1), 99–106.
- Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, 38(6), 977–997. <https://doi.org/10.1002/ejsp.487>
- Bodenhausen, G. V., & Petsko, C. D. (2023). Complications in predicting intergroup behavior from implicit biases: One size does not fit all. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge: Cambridge University Press.
- Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding implicit bias: Putting criticism into perspective. *Pacific Philosophical Quarterly* (online). <https://doi.org/10.1111/papq.12302>
- Buckhout, R. (1966). Changes in heart rate accompanying attitude change. *Journal of Personality and Social Psychology*, 4(6), 695–699. <https://doi.org/10.1037/h0021194>
- Buhrmester, M. D., Blanton, H., & Swann, W. B., Jr. (2011). Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*, 100(2), 365–385. <https://doi-org.srv-proxy2.library.tamu.edu/10.1037/a0021341>
- Cacioppo, J. T., Berntson, G. G., Norris, C. J., et al. (2012). The evaluative space model. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (vol. 1, pp. 50–72). London: SAGE Publications. <https://doi.org/10.4135/9781446249215.n4>
- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, 1(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Lawrence Erlbaum.
- Conrey, F. R., Sherman, J. W., Gawronski, B., et al. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487. <https://doi.org/10.1037/0022-3514.89.4.469>
- Coopersmith, S. (1981, 1987). *Self-esteem Inventories*. Palo Alto, CA: Consulting Psychologists Press.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104. <https://doi.org/10.1037/0021-9010.78.1.98>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, 12(2), 163–170. <https://doi.org/10.1111/1467-9280.00328>
- Cvencek, D., Meltzoff, A. N., Maddox, C. D., et al. (2021). Meta-analytic use of balanced identity theory to validate the implicit association test. *Personality and Social Psychology Bulletin*, 47(2), 185–200. <https://doi-org.srv-proxy1.library.tamu.edu/10.1177/0146167220916631>
- Davenport, E. C., Jr., Davison, M. L., Liou, P., et al. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, 34(4), 4–9. <https://doi.org/10.1111/emip.12095>
- Dawson, N. V., & Arkes, H. R. (2009). Letter to the editor: Implicit bias among physicians. *Journal of General Internal Medicine*, 24, 137–140. Retrieved May, 2020 at: www.ncbi.nlm.nih.gov/pmc/articles/PMC2607500/
- Dehon, E., Weiss, N., Jones, J., et al. (2017). A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine*, 24(8), 895–904.

- Fazio, R. H. (1999). Correlation between the IAT and the Bona Fide Pipeline. Unpublished raw data.
- Fazio, R. H., Jackson, J., Dunton, B. C., et al. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69(6), 1013–1027.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58(2), 203–210. <https://doi.org/10.1037/h0041593>
- Fishbein, M., & Ajzen, I. (1975). *Beliefs, Attitudes, Intentions and Behavior*. New York, NY: Addison-Wesley.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and Changing Behavior: The Reasoned Action Approach*. New York, NY: Psychology Press.
- Forscher, P. S., Lai, C. K., Axt, J. R., et al. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3), 522–559. <https://doi.org/10.1037/pspa0000160>
- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2nd ed.). New York, NY: Cambridge University Press.
- Gawronski, B., Morrison, M., Phills, C. E., et al. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, 43(3), 300–312. <https://doi.org/10.1177/0146167216684131>
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436. <https://doi.org/10.1007/BF02295430>
- Green, A. R., Carney, D. R., Pallin, D. J., et al. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients. *Journal of General Internal Medicine*, 22(9), 1231–1238. <https://doi.org/10.1007/s11606-007-0258-5>
- Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, 7(2), 99–108. <https://doi.org/10.1177/1745691611434210>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Journal of Personality and Social Psychology*, 102, 4–27.
- Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, 72(9), 861–871. <https://doi.org/10.1037/amp0000238>
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, 108(4), 553–561. <https://doi.org/10.1037/pspa0000016>
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., et al. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1), 3–25. <https://doi.org/10.1037/0033-295X.109.1.3>
- Greenwald, A. G., Brendl, M., Cai, H., et al. (2019). *The Implicit Association Test at age 20: What is known and what is not known about implicit bias*. University of Washington. Retrieved May 20, 2020 from: <https://psyarxiv.com/bf97c>
- Greenwald, A. G., Brendl, M., Cai, H., et al. (2022). Best research practices for using the Implicit Association Test. *Behavioral Research Methods*, 54(3), 1161–1180. <https://doi.org/10.3758/s13428-021-01624-3>
- Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, 79(6), 1022–1038. <https://doi.org/10.1037/0022-3514.79.6.1022>
- Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, 71, 419–445. <https://doi.org/10.1146/annurev-psych-010419-050837>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://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(2), 197–216. <https://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(1), 56–61. <https://doi.org/10.1037/0003-066X.61.1.56>
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., et al. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17–41. <https://doi.org/10.1037/a0015575.suppl> (Supplemental)
- Greenwald, A. G., Rudman, L. A., Nosek, B. A., et al. (2006). Why so little faith? A reply to Blanton and Jaccard's (2006) skeptical view of testing pure multiplicative theories: Postscript. *Psychological Review*, 113(1), 180. <https://doi.org/10.1037/0033-295X.113.1.180>
- Heider, J. D., & Skowronski, J. J. (2007). Improving the predictive validity of the Implicit Association Test. *North American Journal of Psychology*, 9(1), 53–76.
- Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, 212(4), 46–54. <https://doi.org/10.1038/scientificamerican0465-46>
- Jaccard, J. (2021). Methodological issues in the study of implicit attitudes. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge: Cambridge University Press.
- Jaccard, J., & Blanton, H. (2007). A theory of implicit reasoned action: The role of implicit and explicit attitudes in the prediction of behavior. In I. Ajzen, D. Albarracín, & J. Hornik (Eds.), *Prediction and Change of Health Behavior: Applying the Reasoned Action Approach*. Mahwah, NJ: Erlbaum Associates, pp. 69–81.
- Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, 76(5), 349–364. <https://doi.org/10.1037/h0031617>
- Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, 28(1), 10–19. <https://doi.org/10.1177/0963721418797309>
- Jost, J. T., Rudman, L. A., Blair, I. V., et al. (2009). The existence of implicit bias is beyond reasonable doubt. *Research in Organizational Behavior*, 29, 39–69. <https://doi.org/10.1016/j.riob.2009.10.001>
- Jussim, L., Careem, A., Goldberg, Z., et al. (2021). IAT scores, racial gaps, and scientific gaps. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge: Cambridge University Press.
- Krugman, P. (2020). *Arguing with Zombies: Economics, Politics, and the Fight for a Better Future*. New York, NY: W.W. Norton & Company.
- Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, 28(4), 281–287.
- Kurdi, B., Seitchik, A. E., Axt, J. R., et al. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, 74(5), 569–586. <https://doi.org/10.1037/amp0000364>
- Lai, C. K., Skinner, A. L., Cooley, E., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, 145(8), 1001–1016. <https://doi.org/10.1037/xge0000179>
- Mason, B. (2020). *Curbing Implicit Bias: What Works and What Doesn't*. Knowable magazine. Downloaded at: www.knowablemagazine.org/article/mind/2020/how-to-curb-implicit-bias.
- 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(5), 435–442. <https://doi.org/10.1006/jesp.2000.1470>

- McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition*, 20(6), 483–510.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*. New York, NY: Psychology Press, pp. 265–292.
- Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, 86(5), 653–667. <https://doi.org/10.1037/0022-3514.86.5.653>
- Oswald, F. L., Mitchell, G., Blanton, H., et al. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171–192. <https://doi.org/10.1037/a0032734.supp> (Supplemental)
- Oswald, F. L., Mitchell, G., Blanton, H., et al. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, 108(4), 562–571. <https://doi.org/10.1037/pspa0000023>
- Oudekerk, B. (2019). Hate Crime Statistics: Briefing prepared for the Virginia Advisory Committee, U.S. Commission on Civil Rights, Panel 1: Hate Crime History in VA, Current Legal Framework, Enforcement and Data. Retrieved May, 2020 at: www.bjs.gov/content/pub/pdf/hcs1317pp.pdf.
- Payne, B. K., Govorun, O., & Arbuckle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition & Emotion*, 22(2), 238–271.
- Payne, B. K., Lee, K. M., Giletta, M., et al. (2016). Implicit attitudes predict drinking onset in adolescents: Shaping by social norms. *Health Psychology*, 35(8), 829–836. <https://doi.org/10.1037/hea0000353.supp> (Supplemental)
- Payne, B. K., Vuletic, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, 116(24), 11693–11698.
- Payne, B. K., Vuletic, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, 28(4), 233–248.
- Porier, G. W., & Lott, A. J. (1967). Galvanic skin responses and prejudice. *Journal of Personality and Social Psychology*, 5(3), 253–259. <https://doi.org/10.1037/h0021205>
- Rankin, R. E., & Campbell, D. T. (1955). Galvanic skin response to Negro and white experimenters. *The Journal of Abnormal and Social Psychology*, 51(1), 30–33. <https://doi.org/10.1037/h0041539>
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. London: Routledge.
- Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, 3(2), 113–123.
- Roose, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, 114(2), 363–375. <https://doi.org/10.1037/0033-2909.114.2.363>
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, 133(2), 139–165. <https://doi.org/10.1037/0096-3445.133.2.139>
- Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Assessment of individual differences in implicit cognition: A review of IAT measures. *European Journal of Psychological Assessment*, 24(4), 210–217. <https://doi.org/10.1027/1015-5759.24.4.210>
- Sherman, S. J., Rose, J. S., Koch, K., et al. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, 22(1), 13–39. <https://doi.org/10.1521/jscp.22.1.13.22766>

- Singal, J. (2017). Psychology's favorite tool for measuring racism isn't up to the job. *New York Magazine*, 11. <https://www.thecut.com/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html>
- Southern Poverty Law Center. (2019). *Gate groups reach record high*. Retrieved May 2020 at: www.splcenter.org/news/2019/02/19/hate-groups-reach-record-high
- Steenkamp, J.-B. E. M., De Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research*, 47(2), 199–214. <https://doi.org/10.1509/jmkr.47.2.199>
- Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology*, 50(1), 33.
- Swann, W. B., Jr., Chang-Schneider, C., & Larsen McClarty, K. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist*, 62(2), 84–94. <https://doi-org.srv-proxy2.library.tamu.edu/10.1037/0003-066X.62.2.84>
- Webb, E. J., Campbell, D. T., Schwartz, R. D., et al. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago, IL: Rand McNally.
- Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 25(4), 41–78.
- Yamaguchi, M., & Beattie, G. (2020). The role of explicit categorization in the Implicit Association Test. *Journal of Experimental Psychology: General*, 149(5), 809–827. <https://doi.org/10.1037/xge0000685>