Commentary

No Measure Is Perfect, but Some Measures Can be Quite Useful

Response to Two Comments on the Brief Implicit Association Test

Anthony G. Greenwald[1] and N. Sriram[2]

[1]University of Washington, Seattle, WA, USA, [2]University of Virginia, Charlottesville, VA, USA

Abstract. The comment articles in this issue by Friese and Fiedler (F&F) and by Rothermund and Wentura (R&W) offer perspectives on the validity of the Brief Implicit Association Test (BIAT) (Sriram & Greenwald, 2009; S&G). F&F concluded that construct validity of the BIAT can be established only by conducting studies that experimentally manipulate association strengths. We suggest that this conclusion overvalues experimental strategies and undervalues correlational validation strategies. R&W's critique was predicated on their use of a semantic-network theoretical understanding of the concept of association. In contrast, S&G offered the BIAT as a technique for measuring association strengths in the context of a broader concept of association that has roots in antiquity – and remains widely used in psychology. With this broader understanding of association, some of the phenomena that R&W treated as threats to the BIAT's validity are viewed, instead, as contributors to its validity.

Keywords: Implicit Association Test, implicit measure, social cognition

The two comment articles to which this article replies contain useful expositions that go beyond topics covered by Sriram and Greenwald (2009; S&G hereafter). We start this reply by observing that the Brief Implicit Association Test (BIAT), to which the two comment articles were directed, is at the beginning of its scientific life. The accumulated published evidence for the BIAT's construct validity is limited to two reports (Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009; S&G). Friese and Fiedler's comment article (2010; F&F hereafter) suggested a research strategy that might be useful in future validation studies. Rothermund and Wentura's comment article (2010; R&W hereafter) focused on questions of theoretical interpretation. We consider these in turn.

Research Strategies for Construct Validity

The main point of F&F's comment was this:

> . . . both [correlational and experimental] approaches provide valuable and mutually complementing evidence, but *only* experimental research can conclusively show that the to-be-measured constructs causally influence BIAT scores (p. 228, emphasis added).

We see the use of "only" by F&F as overstating the conclusiveness of experimental evidence and, by implication, underestimating the value of correlational evidence. To elaborate:

Limits of Experimental Validation Designs

Imagine the best possible experimental study in which an experimental manipulation alters the strength of an association that is to be measured by the BIAT. The value of this validation strategy rests on two causal assumptions: (a) the manipulation affected a latent construct of association strength and nothing else and (b) the presumably manipulated association affected the BIAT measure. If these two assumptions are valid, the experimental design has the virtue of isolating the effect of association strength, uncontaminated by other causes. Unfortunately, present technology affords no method of directly observing manipulated association strengths. There can therefore be no way to assure that step (a) occurred: that is, that the manipulation did indeed affect association strength. Equally, there is no way to assure the required absence of any other causes that might, inadvertently, have been set in motion by the experimental manipulation.

To buttress their argument, F&F (p. 229) drew on Pfungst's (1911) account of the pseudo-arithmetically gifted horse, Clever Hans. F&F used Clever Hans's performance as a cautionary tale, implying that nonexperimental validation

DOI: 10.1027/1618-3169/a000075

designs generally risk conclusions as spurious as the conclusion that Clever Hans could do arithmetic. However, any experimental validation study is itself at risk for this type of spurious conclusion. For example, assume that in the presence of his trainer, Hans is randomly given arithmetic problems in Spanish or German, to confirm that Hans understands only German. If Hans's trainer does not understand Spanish, Hans might, as expected, solve only the problems presented in German. Only when the experiment is designed with proper recognition of the alternative (and, as is generally believed, the only effective) causal path – that is, the involvement of subtle communications from trainer to horse – can one construct a presumably proper experiment.

Value of Correlational Validation Designs

Experimental validation designs, as urged by F&F, have the obvious virtue of ruling out any (backward) causal path from dependent variable to independent variable. However, many correlational studies can likewise afford great confidence that a nonmanipulated "independent" variable is free of causal influence of the "dependent" variable. An example: Suppose that one wishes to develop a measure of weight using photographic images of people. A correlational study might involve photographing numerous nude models, computing numerous body dimensions from the photographs, and then trying to statistically capture useful weight indicators via correlational analyses. Researchers can be confident that the image-derived measures did not cause variations in the (previously measured) weights of the models. Also, the assumption that actual weights of the models are properly understood as causes of the image-derived measures will be highly plausible. The conclusion: Correlational designs can be very effective in limiting possible paths of causal explanation.

Many scientifically adequate measures cannot be subjected to the type of experimental validation advocated by F&F. For example, the use of spectral red shift as a measure of astronomical distances cannot depend on experimental validation in which distances between observers and hugely distant light sources are experimentally varied. In psychology, sophisticated correlational validation methods based on multitrait multimethod designs (Campbell & Fiske, 1959) have been widely (and successfully) used in construct validation studies. For example, the most widely accepted measures of intelligence have never had the benefit of randomized experiments to manipulate either the presumed genetic or environmental causes of intelligence. Validation of these intelligence measures depends on correlations in samples that allow natural variations of the presumed causes.

Theoretical Interpretation

In commenting on the BIAT, R&W recapitulated points that they had offered in two previous critiques of the Implicit Association Test (IAT; Rothermund & Wentura, 2001, 2004). As explained by Greenwald, Nosek, Banaji, and Klauer (2005) in their reply to the 2004 critique, Rothermund and Wentura's nonacceptance of the IAT as a measure of association strengths was based on their preference for a semantic-network definition of association, which was different from and considerably narrower than the definition preferred by Greenwald et al. Here is how Rothermund, Wentura, and De Houwer (2005) described the difference.

> Drawing on the philosophical writings of Aristotle and Hume, Greenwald et al. argued for such a broad usage because the term association can indicate very different things, like contiguity (in time or place), frequency, similarity, contrast, or causation…. This usage can be contrasted with theoretically more constrained accounts that use the term association to indicate links in a semantic network structure… (p. 427).

Using their definition of association, R&W could not conceive IAT measures as measures of association strengths. They therefore proposed an alternative explanation of the IAT in terms of "strategic recoding." As Greenwald et al. (2005) pointed out:

> [W]hat appeared to be a central disagreement in interpretation of the IAT between Greenwald et al. (1998) and Rothermund and Wentura (2004) proved to be no more than different preferences for defining the concept of association. This definitional disagreement has implications for the choice of language to describe results that are expected to occur in similar empirical form by both Greenwald et al. and Rothermund and Wentura (p. 421).

No Measure Is Perfect

The two comment articles identified "unwanted factors" (F&F, p. 230) or "nonassociative influences" (R&W, p. 234) that influence the magnitude of IAT effects, including order of combined-task blocks, composition of stimulus sets, salience asymmetries, social context, familiarity, mere acceptance effects, processing fluency, polarity correspondence, salience asymmetries, perceptual similarity, task-switching ability, and strategic recoding. This is not an appropriate place to review details of evidence in support of these influences. Nevertheless, we can agree that several of these do have supporting empirical evidence and also that some of these bear on the validity and usefulness of IAT measures. We see these influences as falling into two categories that deserve to be distinguished.

Undesired and Often Avoidable Influences

Those who are unaware of undesired influences on IAT measures are at risk to use the IAT's method ineffectively and, thereby, to construct inadequate IAT measures. To the extent that these undesired influences are understood, they can either be procedurally avoided or statistically controlled (see, for example, the recommendations in Lane, Banaji, Nosek, & Greenwald, 2007; Nosek, Greenwald, & Banaji, 2007). For example, under the heading of stimulus selection, IAT measures are known to be impaired when categories are represented either by very unfamiliar stimuli or by atypical stimuli (e.g., butterfly, rather than mosquito, for insect; stinkweed, rather than tulip, for flower – see Govan & Williams, 2004). These influences can be avoided by not including such stimuli as category exemplars in IAT measures. The undesired effect of order of combined tasks on IAT measures (first documented by Greenwald, McGhee, & Schwartz, 1998) can be attenuated by adding practice on the reversed discrimination before administering the second combined task (Nosek, Greenwald, & Banaji, 2005); it can also be statistically managed by using order of combined tasks as a blocking factor in data analyses.

Desired Influences

Two of the suggested "unwanted" factors – task switching and strategic recoding – are ones that we understand, actually, as contributors to the validity of IAT and BIAT measures. Research by Greenwald et al. (1998) prior to their first publication of the IAT had found that IAT effects were larger when combined-task blocks were constructed with a task switch (between the two component 2-choice discriminations) on every trial of the block. These task switches prove to be easy when the two categories that share a key are associated, but they prove to be difficult when the two key-sharing categories are unassociated. This is why we regard task switching, which is maximized by the standard IAT procedure of strictly alternating the two 2-choice discriminations, as contributing to the IAT's usefulness as a measure of association strengths.

We likewise understand strategic recoding, which allows a subject to treat two key-sharing categories as one superordinate category (e.g., treating flower names and pleasant words as members of a superordinate category of "good things") as something that depends on the strength of association between those two categories and, therefore, contributes to the IAT's ability to measure that association strength. This observation applies at least partly to the recoding-related processes identified as salience asymmetry (Rothermund & Wentura, 2001, 2004), mere acceptance (Mitchell, 2004), polarity correspondence (Proctor & Cho, 2006), and perceptual similarity (De Houwer, Geldof, & De Bruycker, 2005). When these permit simplified representation of a combined task's two nominal categories, they may contribute to an IAT or BIAT measure's validity.

We hasten to add that the preceding two paragraphs' brief treatments of task switching and strategic recoding are likely oversimplifications. Consider two subjects who hypothetically have the same strengths of valence associations for flowers and insects, but differ in their possession of the cognitive skills needed for task switching or strategic recoding. Because of the involvement of task switching and strategic recoding in IAT performance, these two subjects may not produce identical measures on an IAT or BIAT flower-insect attitude measure. If the resulting invalidity is nontrivial, one might cope with it by measuring the relevant cognitive abilities for each respondent and using those measures to statistically adjust IAT- or BIAT-measured association strengths. We hope to see some demonstrations of usefulness for such validity-increasing strategies soon (cf. Klauer, Schmitz, Teige-Mocigemba, & Voss, in press). From our perspective, the alternative of designing IAT or BIAT measures to eliminate task switching or strategic recoding (e.g., Houben, Rothermund, & Wiers, 2009; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009; Teige-Mocigemba, Klauer, & Rothermund, 2008) risks impairing those measures by eliminating processes that contribute to their validity.

Undesired influences are inevitable, which explains our title assertion that "No measure is perfect." In terms of its metric qualities, the IAT's properties approximately resemble those of sphygmomanometer blood pressure (BP) measures that are used to assess hypertension. Both IAT and BP have good, but not outstanding, test-retest reliability. Both IAT and BP have multiple unwanted influences. For BP, the unwanted influences include effects of time of day, recent activity, anxiety levels (white coat syndrome), concurrent medications, and technician competence. These influences on BP measures obviously have not proved convincing as arguments against using them to predict or diagnose cardiovascular disease. As for BP, the value of IAT or BIAT measures will ultimately depend on their having the usefulness in research and in applications that BP measures are well known to have. For the IAT, there is now substantial evidence of usefulness in the form of predictive validity (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). For the BIAT – in its infancy – present evidence of usefulness rests mainly on demonstrations of functional similarity to IAT measures (S&G), but the first test of predictive validity of a BIAT measure of race attitudes indicated that measure's usefulness in predicting vote in the 2008 United States Presidential election (Greenwald, Smith, et al., 2009).

In offering their list of unwanted influences on IAT measures, R&W observed, "The existence of these nonassociative influences undermines the fundamental claim – expressed in the very name of the task – that an IAT effect is a pure measure of associations between the nominal categories of a specific IAT" (p. 234). As firm believers in the imperfection of all psychological measures, we wonder whether anyone has ever made this "fundamental claim" of purity. (We certainly have not.) As a parallel, we do not think that the use of "Intelligence" in the name of IQ measures implies that IQ measures are to be understood as "pure" measures of intelligence.

Corrected Characterizations of the IAT and BIAT

Scoring Method of the Standard IAT

In their Footnote 1, R&W described the scoring of the standard IAT as being based on analysis of just two of the seven trial blocks of its procedure. However, since the publication of an improved scoring method (Greenwald, Nosek, & Banaji, 2003) most researchers have used four of the seven blocks in scoring IAT measures.

Relevance of Nonfocal Categories

In describing the BIAT, R&W asserted that the BIAT's "focusing manipulation eliminates the relevance of the nonfocal categories for the task" (p. 233). No empirical tests of that assertion have yet been conducted. We expect that this assertion will likely prove incorrect when such tests appear. Nevertheless, it is true that one of the four categories is not explicitly mentioned in the instructions for a BIAT. Even so, in administrations of the BIAT, it is considered advisable to present the items for all four categories prior to any data collection, so that subjects will not be surprised by the first appearance of items for a category that was not mentioned in instructions.

Types of Categories in the BIAT

In their Footnote 2, R&W observed that "in the BIAT, one attribute category is combined with two different target categories." This might more accurately say that, in constructing pairs of focal categories for the BIAT, one category of any type is combined, successively, with each of two other categories of any type. The distinction between target and attribute categories characterized the very first IAT attitude measures (Greenwald et al., 1998), but that distinction became inconsequential after the development of identity and stereotype IATs, which sometimes had no attribute categories. Examples with no attribute categories can be found in S&G's gender-stereotype BIAT (the four categories of which were *male*, *female*, *arts*, and *science*) and their gender identity IAT (which used *male*, *female*, *self*, and *other*).

Psychometric Effect of Task Shortening

The BIAT typically has a third or so of the number of trials of a standard IAT. In commenting on this reduced number of trials, R&W commented "Apparently, shortening the task in this way does not substantially impair the psychometric quality and predictive validity of the BIAT" (p. 233). Although the impairment may not be "substantial", there was indeed impairment. The findings reported by S&G in their Table 2 found that internal consistency, test-retest reliability, and implicit-explicit correlations were all lower for the BIAT than for the standard IAT. Nevertheless, these reductions were small enough to suggest that BIAT measures might psychometrically outperform IAT measures when both were based on the same number of trials.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

De Houwer, J., Geldof, T., & De Bruycker, E. (2005). The Implicit Association Test as a general measure of similarity. *Canadian Journal of Experimental Psychology, 59*, 228–239.

Friese, M., & Fiedler, K. (2010). Being on the lookout for validity. Comment on Sriram and Greenwald (2009). *Experimental Psychology, 57*, 228–232.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by redefining the category labels. *Journal of Experimental Social Psychology, 40*, 357–365.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

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

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the IAT: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General, 134*, 420–425.

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

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Race attitude measures predicted vote in the 2008 U.S. Presidential Election. *Analyses of Social Issues and Public Policy, 9*, 241–253.

Houben, K., Rothermund, K., & Wiers, R. W. (2009). Predicting alcohol use with a recoding-free variant of the Implicit Association Test. *Addictive Behaviors, 34*, 487–489.

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology, 63*, 595–619.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV. What we know (so far). In B. Wittenbrink & N. S. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 59–102). New York: Guilford Press.

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test (IAT). *Journal of Experimental Social Psychology, 40*, 366–373.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166–180.

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.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.

Pfungst, O. (1911). *Clever Hans*. New York: Holt.

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 3*, 416–442.

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology, 62*, 84–98.

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift für Experimentelle Psychologie, 48*, 94–106.

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

Rothermund, K., & Wentura, D. (2010). It's brief but is it better? An evaluation of the brief implicit association test. *Experimental Psychology, 57*, 233–237.

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

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology, 56*, 283–294.

Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment, 24*, 237–245.

Received January 6, 2010
Revision received March 24, 2010
Accepted March 27, 2010

Anthony G. Greenwald

Department of Psychology
University of Washington
Box 351525
Seattle
WA 98195-1525
Tel. +1 206 543 7227
E-mail agg@u.washington.edu