Date of manuscript: March 26, 2019

The Implicit Association Test at age 20: What is known and what is not known about implicit bias

Anthony G. Greenwald, University of Washington
Miguel Brendl, University of Basel
Huajian Cai, China Academy of Science
Tessa Charlesworth, Harvard University
Dario Cvencek, University of Washington
John F. Dovidio, Yale University
Malte Friese, University of Saarland
Adam Hahn, University of Koeln
Eric Hehman, McGill University
Wilhelm Hofmann, Ruhr University Bochum
Sean Hughes, University of Ghent
Ian Hussey, University of Ghent
Christian Jordan, Wilfrid Laurier University
John Jost, New York University
Teri Kirby, University of Exeter
Calvin K. Lai, Washington University of Saint Louis
Jonas W. B. Lang, Ghent University
Kristen P. Lindgren, University of Washington
Dominika Maison, University of Warsaw
Brian D. Ostafin, University of Groningen
James R. Rae, University of Massachusetts
Kate A. Ratliff, University of Florida
Colin T. Smith, University of Florida
Adrian Spruyt, University of Ghent
Reinout W. Wiers, University of Amsterdam

The authors are grateful for helpful comments from: Mahzarin R. Banaji, Yoav Bar-Anan, Jan De Houwer, John F. Kihlstrom, Benedek Kurdi, Franziska Meissner, Gregory Mitchell, Brian A. Nosek, Marco Perugini, Klaus Rothermund, Jeffrey Sherman

2nd through 25th authors are alphabetical by last name

Citation:
Greenwald, A. G., Brendl, M., Cai, H., Charlesworth, T., Cvencek, D., Dovidio. J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Jost, J., Kirby, T., Lai, C. K., Lang, J., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., Ratliff, K., Smith, C. T., Spruyt, A., & Wiers, R. W. (2019). The Implicit Association Test at age 20: What is known and what is not known about implicit bias. University of Washington. Retrieved from https://psyarxiv.com/bf97c

The Implicit Association Test at age 20:

What is known and what is not known about implicit bias

Abstract. Scientific interest in unintended discrimination that can result from implicit attitudes and stereotypes (implicit biases) has produced a large corpus of empirical findings. In addition to much evidence for validity and usefulness of Implicit Association Test (IAT) measures, there have been psychological critiques of empirical findings and theoretical disagreements about interpretation of IAT findings. Because of public attention drawn by the concept of implicit bias, commercial and other applications based on the concept of implicit bias have been developed by non-psychologists—some of these applications are not appropriately guided by the existing body of research findings. This article is in 5 parts: (1) review of best practices for research use of IAT measures, (2) summary of what has been confidently learned from empirical research using IAT measures, (3) accepted and controversial theoretical interpretations of IAT findings, (4) significant questions about the IAT and implicit bias that still await answer, and (5) questions arising in attempts to apply research findings to remedy unintended discrimination due to implicit biases.

Keywords: Implicit Association Test, implicit bias, psychometrics, construct validity

Greenwald and Banaji (1995) reviewed methods and findings in an area of research for which they offered the label *implicit social cognition*. They focused on work published in journals featuring social psychology and personality research—and more specifically on research using *indirect* measures of attitudes, stereotypes, and self-esteem. Their concluding sentence was: "Perhaps the most significant remaining challenge is to adapt these methods [i.e., the indirect measurement strategies that they had reviewed] for efficient assessment of individual differences in implicit social cognition."

Greenwald, McGhee, and Schwartz (1998) addressed that challenge in their article, "Measuring individual differences in implicit cognition: The Implicit Association Test". The subsequent body of reports of research using the Implicit Association Test (IAT) now exceeds 2,500 peer-reviewed articles.[1] Setting aside the too-difficult task of reviewing this entire body of work, this article aims to summarize findings and conclusions that should be most useful to researchers who wish to use the IAT in research or application. .

[1] In early March of 2019, the American Psychological Association's PsycNET database contained retrieved 3,608 publications that included "Implicit Association Test" in at least one of the fields of Title, Abstract, Keywords, or Tests and Measures. Inclusion in one of these fields should indicate that the IAT was a focal topic of the retrieved item. The retrieved items included 2,679 peer-reviewed articles and 172 dissertation abstracts. This does not include numerous scholarly publications in disciplines outside of Psychology, including Medicine, Law, Political science, Business, Education, and Economics.

What are 'implicit' measures?

Implicit often appears in the text of psychological publications as an adjective preceding *memory*, *attitude*, *stereotype*, *self-esteem*, *identity*, and *association*. These adjective–noun pairs are often contrasted with pairs in which *explicit* is the adjective. This implicit–explicit contrast has been understood in two ways. Understanding 1 treats *implicit* and *explicit* as properties of psychological *measures*, describing measures that reveal a construct indirectly (implicitly) versus directly (explicitly). Understanding 2 treats *implicit* and *explicit* as properties of *mental processes or mental representations*, which may be conceived as operating in automatic or unconscious fashion (implicitly) or in controlled or conscious fashion (explicitly).

Understanding 2 derives from memory studies of the 1980s, many of which used indirect measures to reveal operations of memory that occurred without conscious recollection (cf. Richardson–Klavehn & Bjork, 1988). By the early 1990s, however, two influential methodological articles (Jacoby, 1991; Reingold & Merikle, 1988) had offered convincing (and subsequently unrefuted) arguments that it was neither justifiable (a) to treat indirect measures as pure indicators of unconscious process, or (b) to treat direct measures as pure indicators of conscious process. Those conclusions justify Understanding 1. Reviewing this history that preceded their 1995 article, Greenwald and Banaji (2017, pp. 861–863) similarly concluded that 'implicit' and 'explicit' are most justifiably used to describe (respectively) measures that reveal psychological constructs indirectly vs. directly—not as synonyms for 'unconscious' vs. 'conscious'.[2] In introducing the Implicit Association Test, Greenwald, McGhee, and Schwartz (1998) used 'implicit' to describe a property of their measure rather than of the construct it was measuring. In their overview of "Implicit measures in social cognition research", Fazio and Olson (2003) even more strongly emphasized indirect measurement as the primary distinguishing property of implicit measures.

The most forceful argument for a mental process understanding of 'implicit' is that of De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009a): "the term *implicit* can best be understood as being synonymous with the term *automatic*," (p. 350). (See commentary on that view by Nosek & Greenwald, 2009 and by De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009b.) Arguably, a virtue of the presently recommended approach (implicit = indirect) is that researchers can readily agree on identifying some measures as indirect, while they find much more difficulty in judging that a measure depends on automatic mental operations.

Those who lean toward the mental process understanding may assume, encountering this article's multiple references to 'implicit attitude' or 'implicit stereotype', that the authors have lapsed into mental process language. In those uses, the authors understand 'implicit X' to mean 'X measured indirectly', not as meaning 'unconscious X'. Furthermore, this article has been

[2] Greenwald and Banaji (1995) defined implicit social–cognitive constructs as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate [response to] social objects" (p. 8). That definition remains useful, but is silent on the measure vs. representation/process distinction that has been focal in subsequent treatments of the definition of 'implicit', such as those of Fazio and Olson (2003) and De Houwer et al. (2009a, 2009b)

written to avoid statements in which a mental-process understanding of 'implicit' should at all complicate or interfere with comprehension.

1. BEST RESEARCH PRACTICES

References to "the IAT" in the remainder of this article will mostly refer to the *standard form* of the IAT, which has seven sets (*blocks*) of *trials*, each of which presents a stimulus (*exemplar*) belonging to one of the IAT's two *target* categories or to one of its two *attribute* categories. (This now-standard procedure, which has substantially fewer trials than were used in the 1998 initial publication of the IAT, is described in greater detail in Appendix A.) Four of the seven blocks (3, 4, 6, and 7) present *combined tasks* in which exemplars of one pair of categories appear on all odd-numbered trials, and exemplars of the other pair appear on all even-numbered trials. The subject's only instructed task is to press a left key or a right key to classify each exemplar into its proper category. What enables the IAT to provide an indirect measure of association strengths is that (a) the same two response keys are used to classify target and attribute concepts and (b) the correct response sides for the two target categories are switched (from those used initially in blocks 1, 3, and 4) between right and left for the second combined task (in blocks 6 and 7). The latency difference between these two combined tasks provides the basis for an indirect measure of relative association strengths (see description of the *D* measure in 2–1 and Appendix B).

This section's recommendations for best research practices have two sources: (a) published experimental studies of procedural variations and (b) informal knowledge accumulated in years of research experience by the present authors and others. Most of the experience-based recommendations have been acquired in pilot testing of previously unused IATs of the type recommended in 1–A8. For some of these practices, the descriptions in this article are their only published description. An asterisk preceding the section header identifies these. All of these practices have a rationale that remains to be confirmed by experimental research. Use of these procedures should be considered advisable but not mandatory. Hopefully, those interested in varying from these recommended procedures will include them in experimental evaluations alongside possibly superior procedures.

1–A. Best Practices for Selection of Categories and Exemplar Stimuli for Use in IAT measures

1–A1. All four categories used in the IAT should be familiar to subjects

Unfamiliar categories cannot be expected to have associations measurable using the IAT. Because of slow classification in responding to exemplars of unfamiliar categories, those categories will appear to be weakly associated with other categories. If exemplars for one of two categories in an attitude IAT are totally unfamiliar, that category will inappropriately appear to be negatively evaluated (as shown by Brendl, Markman, & Messner, 2001). For such categories, an interpretation in terms of the IAT as a measure of association strengths involving the unfamiliar categories is not appropriate. To state this recommendation as simply as possible, the

IAT should not be used to measure associations involving unfamiliar categories. This does not preclude conducting experiments using novel categories for which both category labels and exemplars are previously unfamiliar (e.g., the Pokemon characters used by Olson & Fazio, 2001 and the fictitious Niffians and Laapians used by Ranganath & Nosek, 2008). However, studies involving such unfamiliar categories require prior training involving the novel categories and their exemplars to assure that these are familiar to subjects before the IAT is administered.

1–A2. The primary criterion for selection of exemplar stimuli for each target and attribute category is that they must be *easy* for subjects to sort correctly

Exemplar stimuli that are difficult to categorize will be responded to slowly in the IAT. As in the case of unfamiliar stimuli (see 1–A1), this slowness can inappropriately cause the category containing those exemplars to appear to be weakly associated with another (target or attribute) category in the IAT. If exemplars for only one of two categories in an attitude IAT are difficult to categorize, that category may inappropriately appear to be negatively evaluated. If exemplars for both categories are difficult to classify, there may appear to be no attitude when an attitude might have been detected with use of easily classified exemplars.

An important contributor to easy classification of exemplars is for those exemplars to be *representative* of their categories. Several empirical findings have shown that non-representative exemplars of categories will produce results different from those obtained with representative exemplars (e.g., Bluemke & Friese, 2006; Govan & Williams, 2004; Steffens & Plewe, 2001). For example, Govan and Williams used *nettles*, *skunkweed*, and *poison ivy* as exemplars for the category *flowers*, and used *butterfly*, *grasshopper*, and *firefly* as exemplars for the category *insects*. Section 1–A8 provides a recommendation method for selecting exemplar stimuli that can be easily classified into their respective categories. Section 1–B6 considers the numbers of exemplars that should be selected.

1–A3. Exemplars for any target category should differ from those for its contrasted target category in just *one* primary feature or *one set* of highly correlated features; the same should be true for exemplars of the two attribute categories

When this practice is followed, subjects can have only one basis for distinguishing exemplars of the two contrasted categories. For an example of violation of this practice: If a (racial) contrast between Asian and African involves only male exemplars of Asian persons and only female exemplars of African persons, subjects can use either gender or race as the basis for distinguishing the sets of exemplars (see Mitchell, Nosek, & Banaji, 2003, for studies with such multiple categorization possibilities). Or, if words for a positive valence attribute are all shown in green font while those for negative valence are all in red font, subjects are free to sort based on font color rather than valence. In an attitude IAT using the race and valence exemplar contrasts just described, the IAT measure might indicate (depending on subjects' choice among the available sorting strategies), difference in valence associations of the race groups, difference in valence association of the gender groups, or differences in associations of the font colors with the race or gender groups. Obtaining an interpretable measure requires selecting exemplars that do not allow subjects such flexibility in sorting strategy.

When target concepts are represented by face images, this practice obliges consideration of the expressions on those faces. Having smiling faces for one political candidate and frowning faces for another in a political attitude IAT is obviously undesirable. One solution is to have all face images lack facial expression, although this may be difficult to achieve when drawing on collections of face photos. Also satisfactory is to use faces with smiling or frowning expressions if those expressions are matched in frequency for the two face categories; this will preclude the expressions serving effectively as alternative guides to classification.

1–A4. For IATs designed to measure stereotypes, avoid confounding the stereotype's contrasted attributes with valence

Some published IAT studies have assessed stereotypes using trait attribute contrasts that unavoidably confounded the contrasted traits with valence (examples: strong vs. weak, smart vs. dumb, sober vs. drunk). Such studies are sometimes intercepted on the path to publication by reviewers or editors who will note that the trait contrast was confounded with a valence contrast and the IAT might therefore provide an attitude measure rather than a stereotype measure (cf. Wittenbrink, Park, & Judd, 1997). This confounding deviates from recommendation 1–A3, by allowing subjects to treat the attribute contrast alternately as one of valence, effectively making the stereotype measure a possible attitude measure. This problem can often be avoided by selecting contrasted trait categories that do not differ in valence. Rudman, Greenwald, and McGhee (2001) used two solutions for this problem in trying to measure a male=strong stereotype. One solution selected exemplars for strong and weak that were matched in valence. The other, which proved easier to implement, was to contrast a presumed characteristic of one group (e.g., strength, expected to be more associated with male) with a similarly valenced characteristic of the contrasted group (e.g., warmth, expected to be more associated with female). The second strategy simultaneously measured two stereotypes (male=strong and female=warm), which might be desirable or undesirable, depending on the aims of the research.

1–A5. Avoid exemplars for one attribute category that are negations of possible exemplars for the contrasted attribute category

Negations have the attractive feature of being easy to produce. However, as demonstrated by Phelan and Rudman (2008; see also Verschuere & Kleinberg, 2017), they cause difficulty in IATs, likely because of an extra processing demand of requiring comprehension of the non-negated meaning before apprehending the negated meaning (see Gilbert, 1991, esp. p. 7). For example, processing 'unhappy' requires activating and then negating the meaning of 'happy'. Some other examples: *trust* and *distrust*, *healthy* and *unhealthy*, *true* and *not true*. The negations in these pairs can be avoided by using instead a synonym (of the negation) that is not in negation form—e.g.: *suspicion* in place of *distrust*, *sick* in place of *unhealthy*, and *false* in place of *not true*.

1–A6. Negations can be satisfactory in category labels

Although negations should not be used in exemplars for target or attribute categories (see 1–A5), they are sometimes satisfactory in category labels. Even so, it can be preferable to avoid using negations in category labels if satisfactory labels not in negation form are available. An

example of a case in which it was not possible to find a good label not in negation was a study of smoking-related attitudes by Swanson, Rudman, and Greenwald (2001). The exemplars for the category *smoking* were pictures containing cigarettes. For the contrasted category's exemplars (the same scenes lacking cigarettes) it was not possible to find a better label than *non-smoking*. Many studies have successfully used *Me* vs. *Not me* as category labels in self–concept or self–esteem IATs, in place of using *self* vs. *other* (see Greenwald & Farnham, 2000) . The *Me* vs. *Not-me* contrast is preferable in using self-related IATs with young children, for whom the contrast of *self* vs. *other* as category labels may pose a comprehension challenge (see Cvencek, Greenwald, & Meltzoff, 2011).

1–A7. In selecting attribute exemplars, avoid ones that have other bases for associations with either of the two target concepts

Some otherwise acceptable attribute exemplars may be compromised by strong association with one of the target concepts in the same IAT. Such problems occur infrequently, and they also tend to be obvious. One such example is selecting *cancer* as a negative valence exemplar in an IAT that measures attitude toward smoking—the problem is due to *cancer* being associated with the target concept of *smoking* (and not with *non-smoking*) through its association with *health* rather than (or in addition to) its valence. However, 'cancer' could be used as an exemplar of *physical illness* in an IAT assessing associations of smoking with physical vs. mental illness. In that case the association of cancer with negative valence would not interfere with the contrast between physical and mental illness.

1–A8. Exemplar stimuli for target and attribute categories are best selected by pilot testing using the category classification tasks planned for the IAT

This recommendation follows on the earlier point (1–A2) about ease of classification being a requirement in selecting category exemplars. Subjects for pilot testing should come from the intended research subject population. The designer of any IAT is often the first pilot subject, which is entirely satisfactory and appropriate if the IAT designer is representative of the planned subject population. A judgment as to whether the exemplars are easy enough to classify can be based on examination of data provided by pilot subjects. The useful data will come from Blocks 1 and 2 of the standard procedure (see Appendix A). Pilot subjects should be able to categorize all stimuli in these two blocks rapidly (600–800 ms for most young adult subjects) and with low error rates (less than 10%).

Exemplars that even a small proportion of pilot subjects find difficult to classify correctly are safely discarded without further consideration. There is no need for selection criteria such as word length, word frequency, or meaningfulness, even though these criteria are appropriate for many other investigations of categorization. An obvious exception to the just-stated observation is that word characteristics should not be confounded with a category contrast, such as by using short words as exemplars for one target or attribute category and long words as exemplars for the contrasted category; this would be a deviation from recommended practice 1–A3.

1–A9. When all four concepts in an IAT are expressed as words, one or more font variations can be used to help subjects distinguish target exemplars from attribute exemplars

In the very first published IAT (an attitude measure that contrasted flowers with insects), all four categories were presented as lowercase words. Some subjects in that experiment pointed out that they were sometimes uncertain whether a target concept's exemplars (e.g., lily or rose) were to be sorted as *flower* (target concept) or as *pleasant* (attribute concept). Likewise, maggot and roach might be classified as *insect* (target concept) or as *unpleasant* (attribute concept). To avoid this difficulty for subjects, a case variation was introduced in the second and third experiments of that first IAT report (Greenwald et al., 1998). Valenced attribute exemplars were displayed in all lowercase and target concept exemplars were displayed with initial capital letters. More substantial font differences between attribute and target concept exemplars are not problematic. The target–attribute distinction can be further enhanced by simultaneously varying font color (e.g., green vs. blue), case (upper vs. lower), and typeface (e.g., Courier vs. Arial) between target and attribute exemplars.

1–B. Best Practices for IAT Administration Procedures

1–B1. Counterbalancing the temporal order of the two combined tasks is generally desirable

With two target categories (call them T1 and T2) and two attribute categories (A1 and A2), the first combined task can assign the same key to T1 and either A1 or A2 (and T2 to the same key as the other attribute category). The earliest IAT studies observed an order effect such that the association of T1 with A1 (and T2 with A2) would appear stronger when T1 and A1 were assigned to the same key in the first combined task rather than in the second. To avoid having this effect of combined-task *order* on the estimated sample mean for an IAT measure, it is generally desirable to counterbalance, across subjects and within treatments, the order of administration of the two combined tasks. One desirable reason for avoiding this order effect is that it will displace the zero-point of an IAT measure. (the IAT's zero point is further treated in 2–2 and 3–12.)

In published articles, several researchers have reported that they avoided counterbalancing the order of combined tasks out of concern that variance associated with this counterbalanced procedural variable would reduce estimates of correlations between IAT measure(s) and other measures with which correlations were expected. There are three reasons for this not being a concern. First, when counterbalanced, order of combined tasks can be used as a covariate to correct the estimated correlation of the IAT with other variables for the possible order effect. Second, the effect of order of combined tasks is typically small enough so that its effect on correlations with other variables will be quite small (even without a covariance adjustment).[3]

[3] Correlations of the order of administering combined tasks with IAT measures ranged from –.02 to .25 in the data analyzed by Greenwald et al. (2003) in the article introducing the *D* measure. For an observed correlation of $r = .4$ between the IAT and another variable of interest in a study with counterbalancing of order, the largest of the observed order-effect correlations (.25) used as a covariate would result in an increase of that correlation from .400 to .413.

Third, when counterbalancing is not used, the order effect influences observed sample means of the IAT measure. This perturbation can sometimes be damaging to theoretically based hypothesis tests (see 3–13) and to estimates of scores for which closeness to the zero value matters (see 3–14).

The foregoing observations notwithstanding, the zero-point's location is not always critical. Accordingly, there are numerous hypotheses that can be tested satisfactorily in experiments that do not counterbalance the order of the IAT's two combined tasks.

1–B2. Counterbalancing of sides initially assigned to each category is desirable

Effects on IAT measures of which attribute category is associated with left or right key and the side to which each target concept is initially assigned have not been demonstrated in any published studies. There is nevertheless a suspicion that positioning the positive valence category to the right side may produce a small effect of faster responding than if negative valence is assigned to the right key.[4] The main reason for this counterbalancing is the general principle that conceivable extraneous sources of influence on data should be avoided. This counterbalancing is relatively easy to achieve and is especially desirable in studies with large respondent samples, in which small effects may prove statistically significant.

1–B3. Target and attribute category trials are *always* strictly alternated in the standard IAT's procedure for combined-task blocks

The desirability of this procedure was discovered informally (and repeatedly) in variations of IAT procedures tested in 1994–1995 by the authors of the first IAT publication. The main supporting evidence was that measured IAT effects had larger effect sizes when this procedure was used. The strict alternation procedure was described in five places in the initial publication of the IAT (Greenwald et al., 1998, pp. 1464, 1465, 1467, and 1469). Section 3–4 provides an explanation for how maximizing task switches between target concept and attribute concept classification should both increase facilitatation of IAT performance in one IAT combined task and interfere in the other combined task. Most published reports of IAT measures presumably use this standard alternation, although without reporting its use. Occasional reports mention deviating from the strict alternation for a specific research purpose (e.g., Mierke & Klauer, 2003; Rothermund, Teige–Mocigemba, Gast, & Wentura, 2009). No published report has yet indicated that deviation from strict alternation improves either the IAT's psychometrics or its correlation with conceptually related measures. Although reports of variations from strict alternation have not been designed to test their effects on psychometrics or correlations, it is a near certainty that they both impair psychometrics and reduce correlation magnitudes.

The present recommendation is to use the standard alternation between target and attribute discriminations in combined tasks blocks, which has been used in most publications. Most

[4] This possible effect depends on an assumption that a *right side=positive* cultural association can inflate positivity of a positively valenced concept when the right key is associated with positive valence in Blocks 2, 3, 4, 6, and 7 of the standard IAT. This expected inflation requires the (also untested) assumption the right=positive association does not equally inflate apparent valence when a negatively valenced target concept is assigned simultaneously to the right key.

published reports since the original publication have not explicitly stated that they were used the standard alternation or any variation from it. Researchers should assume, if they do not report otherwise in describing an IAT's procedure, that readers will assume that they were using the standard alternation strategy in combined task trial blocks. It would be desirable if researchers using a different procedure would try to establish that their chosen procedure does not produce results statistically different from those produced using the standard procedure described in Appendix A

1–B4. Intertrial intervals should be brief

Greenwald, McGhee, and Schwartz (1998) varied the interval between occurrence of the response on Trial *n* and presentation of the stimulus for Trial *n*+1 among values of 100 ms, 400 ms, and 700 ms. They found no effect of this variation on magnitude of effects using the IAT. After that early finding, researchers have tended to use quite brief intertrial intervals (250 ms is a commonly used value). This conserves time in a procedure that often has a few hundred trials—adding 1 second to the intertrial interval increases the duration of the standard 190-trial IAT procedure (described in Appendix A) by about 3 minutes. A suspected additional virtue of the brief intertrial interval—albeit one not studied systematically—is to limit intertrial time that can be used to allow mental rehearsal of the correct response key assignments. Greater intertrial time would plausibly reduce difficulty in combined tasks assign the same key to two non-associated concepts; such opportunity to rehearse instructions between trials may permit faster responding. which might in turn reduce the IAT's sensitivity to differences in association strengths. The reasoning here is a close relative to the explanation for larger IAT effects when target and attribute concepts are strictly alternated in combined tasks (see 1–B3).

1–B5. Initial practice in classifying the two target concepts (first block) should precede initial practice in classifying the two attribute concepts (second block)

This conclusion was drawn from never-published exploratory studies conducted prior to the first IAT publication. The explanation: If attribute concept practice comes first, the attribute initially assigned to the left key can acquire some association with that key, such that the ensuing first practice classification of the target categories should boost the association of the target concept assigned to the left key to the attribute previously practiced on that key (and similarly for the right key). The psychological principle underlying this recommendation is *mediated generalization* (Cofer & Foley, 1942), a process by which two categories (e.g., *pleasant* and *insect*), both associated with the same response (e.g., *left key*), can thereby become associated with each other. In this example, when target concepts are practiced first, *left key* acquires an association with *insect* in the first bock. In the second block, *insect* gains some association to *pleasant* by mediated generalization (due to their sharing the left key during the two practice blocks). In the non-recommended procedure, *pleasant* acquires association with *left key* in the first block; *insect* gains an association with *pleasant* in the second block due to mediated generalization. Despite the operation of mediated generalization regardless of order of the first two blocks, there is a theoretically expected asymmetry. In the second block the direction of association formation should be from the category practiced in the second block to the one practiced on the same key in the first block. The expected stronger effect of practicing *insect* in

the second block is that it is easier to form the insect-to-pleasant association than the pleasant-to-insect association. This asymmetry is explained by Paivio's (1969) 'conceptual peg' hypothesis, based on his experiments showing stronger acquisition of associations in noun–adjective (i.e., target–attribute) direction than in adjective–noun (i.e., attribute–target) direction.[5]

1–B6. It is desirable to use at least 3 exemplars for each category in the IAT

In the only experimental study that varied number of exemplars for IAT categories, Nosek, Greenwald, and Banaji (2005) found that as few as two exemplars could be used to represent categories of pleasant, unpleasant, young, old, male, female, science, and liberal arts. Use of a single item per category (the category) label did not fail totally, but was clearly inferior. These results should be generalized cautiously because of the limited number of categories and IAT measures investigated. This caution is applied in recommending a minimum of three items per category. In published studies using the IAT, the numbers of exemplars per category are mostly in the range of four to six. Using four or more exemplars should minimize risk that the category's effective definition in the IAT is distorted by the specific exemplars chosen.

From another perspective, some authors have recommended using two or more interchangeable sets of exemplars for categories when it is easy to generate sufficient numbers of easily classifiable exemplars (as it is for categories such as positive/negative valence, male/female gender, young/old age, and Black/White race (and many others). Wolsiefer, Westfall, and Judd (2017) analyzed the effects of exemplar choice in IAT measurement. They found that variation due to use of different sets of exemplars was smaller in IAT measures than in other indirect measures of social cognition. In response to a personal communication inquiring about implications of their findings for the desirability of using multiple sets of exemplars for IAT categories, along with multilevel modeling of the variance contributed by exemplars, Wolsiefer wrote that such use of exemplar sets and multilevel analysis "doesn't appreciably change individual level bias scores . . . . [W]e also examined whether accounting for stimulus variance in the IAT would appreciably change the predictive validity of the IAT. We found no evidence that this was the case." Even though it is often a desirable feature of research design, it does not appear necessary to develop multiple alternative sets of exemplars for target and attribute concepts in IAT measures. This proves fortunate because in many cases easy-to-classify exemplars are in short supply.

1–B7. It is desirable (not essential) for the number of trials in any block to allow each target exemplar stimulus to be presented the same number of times within the block, and likewise for the exemplars in each attribute category

The desirability of this practice is the usual desirability of minimizing sources of extraneous variance in data due to differences in procedures experienced by research subjects. Adoption of this practice can run into complications in managing equal appearances due to the numbers of exemplars selected for each target category and each attribute category. To achieve equal

[5] Paivio's analysis almost certainly also explains why evaluative priming experiments generally use the concept categories as primes (racial groups, ethnic groups, gender groups) rather than using the attribute categories as primes (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986).

appearances of all attribute-concept or of all target-concept stimuli, trials in combined-task blocks must be twice the smallest value that is simultaneously an integer multiple of the number of unique target exemplars and unique attribute exemplars. For example, with 4 exemplars per target category (total = 8 exemplars) and 5 exemplars per attribute category (total = 10 exemplars), the smallest number that is an integer multiple of both 8 and 10 is 40, requiring a combined task block to have twice that number, or 80 trials, which may be an excessive block length for some subject populations. An acceptable alternative is to distribute the total of 80 trials across the two blocks of each combined task (an example is described in Appendix A). When equal numbers are not possible, it is generally easy to manage stimuli so that no exemplar of a target category is presented more than once more per block than any other exemplar of a target category (and similarly for attribute categories).

1–B8. Runs of more than four consecutive same-key-correct trials in combined-task blocks are undesirable

Runs of consecutive trials that require the same (left or right) key for a correct response allow subjects to increase their performance speed in the IAT due to a well-known repetition priming process (e.g., Horner & Henson, 2008) that is unrelated to strengths of associations between categories that share the same key. If these runs occur in one combined task and not in the other, they can inappropriately affect a subject's IAT measure. And if they occur more for some subjects than others, they can similarly add statistical noise to estimates of means or correlations involving the IAT measure. Lengthy same-key-correct runs are avoidable in combined tasks by randomizing trials independently within each consecutive subset of four trials. Trials 1–4 would then randomly present a stimulus from one target concept on Trial 1 and from the other target concept on Trial 3, and a stimulus from one attribute concept on Trial 2 and the from the other attribute concept on Trial 4; and so on for Trials 5–8, 9–12, etc., with independent randomization for even-numbered and odd-numbered trials in each group of four trials. This strategy limits maximum same-key-correct runs to four trials. For comparison, randomization within groups of 8 trials will allow (very) occasional same-key-correct runs of up to 8 trials.

1–B9. In correlational studies, statistical power can be increased by using 2 or more administrations of the IAT for each subject

This strategy produces an IAT measure with greater test–retest reliability than is expected for a single IAT completion. (The statistical basis for this recommendation is described in 2–6.) Increased test–retest reliability will reduce unsystematic variance in estimated sample means, providing both greater power in tests of experimental treatment effects and increased magnitude of correlations between IAT measures and conceptually related variables An alternative to gaining power for both of these purposes is to increase subject sample sizes.

1–B10. Weaker correlations involving an IAT measure will be observed if the subject population shows little variation in that IAT measure

This expectation is a statistical consequence of restriction of range (see, e.g., Cohen, Cohen, West, & Aiken, 2003, p. 57). As example, if one assesses a correlation between gender identity (which varies widely between male and female) and gender attitude (which is correlated with

gender identity), one observes a stronger correlation when the sample includes both male and female subjects than when one samples just males or just females from the subject population. Similarly, a race attitude IAT (which varies in mean substantially between African Americans and European Americans) will be more strongly correlated with a parallel self-report measure and with other related measures in a sample that includes both racial groups than in a sample limited to one of the two groups. This increased sensitivity to correlations is a justification for not subdividing a sample on demographics when one or more variables being correlated differ non-trivially between the demographic groups that would thereby be analyzed separately.

1–B11. In laboratory research, when IAT-including experiments are administered by multiple experimenters, treatment conditions should be distributed equally across experimenters

This generally advisable research practice is recommended here because of its known significance in research using IAT measures. The effect of experimenter race on subject performance on race attitude IAT measures was first demonstrated by Lowery, Hardin, and Sinclair (2001). Effects of other experimenter characteristics have not been established so clearly as for race of experimenter in the race attitude IAT, but are easily conceivable.

1–B12. Desirable procedures for pretest–posttest IAT administrations

The first IAT ever completed by a subject is known, on average, to show a more polarized result (i.e., greater difference from zero) than will a second or subsequent IAT completion (first reported by Greenwald, Nosek, & Banaji, 2003; see also Lai et al., 2016). This not-yet-fully-understood effect may be due to the first administration having slower responding on combined tasks than do subsequent administrations, if this slowing may occurs more on the combined task that is more difficult for the subject. There are two ways to deal with the resulting expectation of a mean spurious difference between the first and second IAT in a pre-post design: (1) Use a no-treatment control group that also receives both pretest and posttest (used first with IAT measures by Dasgupta & Greenwald, 2001), or (2) give all subjects pre-experimental IAT completion experience, which need not use the same IAT intended for the pretest–posttest design. Without one of these approaches, there is a risk of mistakenly interpreting an observed attenuation of IAT in the posttest as a treatment-caused reduction of the IAT.

2. WHAT IS CONFIDENTLY KNOWN FROM EMPIRICAL RESEARCH USING IAT MEASURES

Because of the quantity of published research using IAT measures, it is not feasible to attempt description of the entire body of confidently established findings. This section aspires nevertheless to present a core of established knowledge about the IAT. The first part of this section's list focuses on findings that describe metric properties of IAT measures. These are followed by established characteristics of correlational findings involving IAT measures.

As explained also at the beginning of Section 1–A, in each of the IAT's two combined tasks subjects use two response keys to classify exemplars of four categories. These most often include two contrasted *target* concept categories and two contrasted *attribute* concept categories. For IAT measures of a subject population's widely shared attitudes and stereotypes, one attribute category is found to be strongly associated with one concept category, and the other attribute category more strongly to the other concept category. In the age attitude IAT (for example) the two concept categories are *young* and *old*, and the two attribute categories are *pleasant* and *unpleasant*. For most who complete this IAT, *young* is more strongly associated to *pleasant* than to *unpleasant* and *old* is associated more strongly to *unpleasant*. The young–old concept pair represents an age dimension and the pleasant–unpleasant attribute represents an attitude (or valence) dimension. The IAT measure can then be understood as a measure of association between the age dimension and the valence dimension; the strength of this (indirectly measured) association is generally understood as the subject's (implicitly measured) attitude for the age dimension.

2–1. The *D* measure is presently the most useful summary statistic for the IAT

Greenwald, Nosek, and Banaji (2003) investigated many possible algorithms for computing a summary measure from latencies recorded in the IAT's two combined tasks. Using a set of multiple performance criteria, they found that a modified effect size measure (*D*) was psychometrically strongest. Effect size measures often calibrate a measure provided by a subject's performance relative to a measure of variability in that performance across subjects. The *D* measure is computed with a numerator measured as the difference between mean latencies for the two combined tasks and a denominator that is a standard deviation (SD) computed from *all* latencies in the IAT's two combined tasks. This differs from a Cohen's *d* measure of effect size, for which the denominator is the pooling of the two SDs computed separately from the IAT's two combined tasks. By thus including the variability of the mean difference between the two combined tasks, the *D* measure denominator's *inclusive SD* contains a corrective related to individual differences in speed of performance, which has largely to do with differences in executive function rather than association strengths. Also unlike Cohen's *d*, the variability that calibrates the subject's performance is the subject's *own* variability in latency, rather than that of an entire subject sample. Appendix B gives details for the steps in computing the *D* measure, as well as for a few possible variations. A recent investigation (Glashouwer, Smulders, de Jong, Roefs, & Wiers (2013) demonstrated that the *D* measure outperforms other available scoring algorithms for the IAT in laboratory experimental studies. (The tests of the *D* measure in its

initial publication were based not on laboratory research, but on data obtained at an educational internet site.)

No improvement on the *D* measure has yet been found. However, it has been found that a non-parametric variant of the *D* measure (computed after converting all latencies to ranks) performs approximately as well as the *D* measure (Sriram, Nosek, & Greenwald, 2007). In addition, Richetin, Costantini, Perugini, and Schönbrodt (2015) tested some variants of the *D* measure (created by adjustments of distribution tails) that perform as well or slightly better than another variant of the original *D* measure (not the one described in present Appendix B and in Table 4 of Greenwald et al., 2003). Alternative approaches to scoring IAT data are also provided by several multi-process models described in Section 3–9. Publications reporting these alternative strategies have not yet evaluated them on criteria that permit comparison of their psychometric properties and construct validity with the *D* measure. An additional approach that has been suggested for use with IAT measures is hierarchical linear modeling, also known as multilevel analysis. However, multilevel analysis has not yet produced an individual-subject summary measure that even approximates performance of the *D* measure.

2–2. The zero point of an IAT measure indicates equal strengths of two complementary dimensional associations

The numerator of the IAT's *D* measure equals zero when the mean response latencies in the procedure's two combined tasks (one in Blocks 3–4 and the other in Blocks 6–7, see Appendix A) are equal. This zero value is interpreted as indicating equal strengths of the complementary associations between the dimensions represented, respectively, by the two target concepts and the two attribute concepts. The IAT's two combined tasks differ in that one estimate is reversed in direction from the other. Considering the age attitude IAT as an example, one combined task measures the association between age and a valence dimension such that faster responding indicates stronger association of the young end of the age dimension with the pleasant end of the valence dimension. For the other combined task (with reversed key assignments for the young and old categories) faster responding indicates stronger association of the old end of the age dimension with pleasant. When these two tasks are performed at the same mean speed, these two opposed-direction associations are as assumed to be equal in strength.

The just-stated rational-zero assumption of the IAT will be problematic if the IAT's procedure plausibly allows determinants of combined-task performance speed other than the two complementary dimensional association strengths. Three procedural choices about use of the two response keys are possible sources of concern for this rational zero-point interpretation: (a) the order in which the two combined tasks are done, (b) the side (left or right) to which the two attribute categories are assigned (these stay constant throughout the IAT), and (c) the sides to which the two concept categories are initially assigned (these are reversed in Block 5). Only the first of these has been empirically demonstrated to have an effect, although there are conceivably small effects of the other two variations that have not yet been empirically established. These three procedural choices should not be troublesome if it can be assumed that any left-right asymmetry in their effects on latency will be equal and opposite if each procedure is left–right reversed (counterbalanced) across subjects in the research procedure. The desirable strategy is

therefore to counterbalance all three procedures orthogonally. However, because any two of these three effectively determine the third, counterbalancing is needed only for any two of the three.

The only one of the three counterbalanced procedures of the preceding paragraph that is known to affect the IAT measure is the order in which the two combined tasks are encountered. This order effect was described by Greenwald et al. (1998), who found that performance on either combined task was reliably more rapid when it was the first one encountered. Displacement of a subject's IAT measure in opposite directions by this order effect therefore *does* threaten interpretation of the IAT's zero value. Counterbalancing order of the combined tasks fixes this problem at the level of group means, but not at the level of individual subjects. Therefore, it should be assumed that individual subjects may have IAT displacements that depend on the order in which they encountered the two combined tasks. Nosek et al. (2005) reported that this order effect can be moderated or eliminated by increasing the number of trials in the fifth block of the standard procedure (the block that gives practice with reversal of key assignments for the concept categories prior to the second combined task). However, neither the counterbalancing nor the adjustment of number of 5th-block trials eliminates perturbations of the zero-point for individual subjects. Counterbalancing should assure, however, that this minor fluctuation of the zero-point contributes only minor error variance to findings. See 3–13 for further consideration of questions related to the IAT's zero point; see also 2–16 about desirability of counterbalancing order of IAT and self-report measures.

2–3. Corollary of the zero-point assumption: IAT measures indicate relative strengths of associations

Interpretation of the zero value of an IAT measure as indicating equal strength of two complementary dimensional associations (2–2) allows interpretation of non-zero IAT values as indicating that one of these associations is greater than the other. This warrants interpretation of the IAT as a measure of *relative* strength of the two complementary dimensional associations. As an example: Assume that a Black–White race attitude IAT measure is scored so that higher scores indicate greater preference for racial White relative to racial Black. As a thought experiment, assume that a score of 1.0 (not 0.0) indicates absence of relative preference. That would mean that a score of 0.0 indicates preference for Black relative to White. These numbers may correlate with a separate measure of relative preference, but the numerical values cannot be directly interpreted as indicating a relative racial preference.

Although the relative association-strength interpretation is widely used in research reports, it is not universally accepted. Alternative interpretations started to appear a few years after the IAT's initial 1998 publication. These included *criterion shift* (Brendl, Markman, & Messner, 2001), *figure ground asymmetry* (Rothermund & Wentura, 2001), *task-switching* (Mierke & Klauer, 2001), *salience asymmetry* (Rothermund & Wentura, 2004), *category recoding* (Rothermund et al., 2009), and *executive function* (Ito et al., 2015; Klauer, Schmitz, Teige–Mocigemba, & Voss, 2010). Comparison of the association-strength interpretation with these alternatives is considered in 3–8 and 3–9. Although these views provide alternative identifications of processes that work in opposed ways in the IAT's two combined tasks, they do

not provide a basis for questioning either the interpretation of the IAT's zero value as indicating equal strengths of opposed processes or the understanding that the IAT provides a measure of relative strength of opposed processes, however they are conceived.

Two aspects of the IAT's being a relative measure have been suggested to be liabilities that should be avoided. Both have to do with the IAT being constructed as a difference score. The first presumed liability is that difference scores often have lower test–retest reliability than do either of the two measures that compose them; from this perspective, if the IAT were not computed using a difference (i.e., the difference between means of the IAT's two combined tasks) in its numerator, it should have superior reliability and, consequently, higher correlations with other measures. However, as Williams and Zimmerman (1996) showed, the conclusion of reduced reliabilities of difference scores depends on assumptions of (a) equal variance of the two means being compared and (b) high correlation between the two means. Neither of these assumptions is appropriate for the pairs of means used in the numerators of IAT measures. Second, various researchers have suggested that a measure of strength of association of a single target concept to an attribute dimension should be preferable to the IAT's measure of a difference in strengths of associations of two target concepts with the attribute dimension. However, as Bar-Anan and Nosek (2014) found (see 4–1), attempts to measure single associations using IAT-like procedures are typically have weaker psychometrics than the IAT. The explanation may be that the IAT's procedure, due to its control of factors extraneous to association strengths that might influence the measures, has the type of increased power that a within–subject design often provides relative to a between–subjects design.[6]

2–4. IAT measures retain their measurement properties with repeated use on the same person

Many people have voluntarily subjected themselves to repeated administrations of one or more IAT measures—in some cases, many repetitions. They often report variation in their IAT scores across repetitions, but describe this variation as occurring within a relatively narrow range. The only deviation from this observation of stability across multiple repetitions of the same IAT is the more polarized result obtained for a first IAT (relative to later ones) that was described in 1–B12 (see also 2–6). The previously anecdotal observation of stability across repetitions of an IAT is now bolstered by data from Lindgren et al.'s (2018) study in which subjects took the same three IATs in 8 sessions separated by 3-month intervals (Lindgren et al., 2018). The not-yet-published analyses of these ancillary findings establish that, excluding the first IAT taken, means and correlations of Lindgren's three alcohol-use-related IATs were quite stable over the two-year period (a portion of these results is described in 3–12).

[6] Those who are familiar with thermometer measures of attitudes (ratings of warmth toward an object) may be aware that thermometer difference measures (e.g., difference of warmth toward two competing political candidates) predict vote substantially more strongly than does the **single** thermometer measure for either candidate.

2–5. Test–retest reliabilities of IAT measures of most social attitudes and stereotypes are no better than moderate

In a few domains—most notably, political attitudes and consumer brand preferences—IAT measures have relatively large test–retest reliabilities of approximately $r = .70$ (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; see 2–9). However, measures of socially sensitive attitudes and stereotypes typically reveal test–retest reliabilities nearer to .50 and often lower. This limited reliability is generally characteristic of latency-based measures of implicit social–cognitive constructs (see Greenwald & Lai, 2020 [in press]).

For purely statistical reasons, the IAT's moderate test–retest reliability limits magnitudes of correlations of IAT measures with conceptually related variables. Test–retest reliabilities of IAT measures therefore also limit magnitudes of correlations reported in published meta-analyses of the IAT's predictive validity (e.g., Kurdi et al., 2019). The moderate reliabilities of IAT measures can be further impaired by requirements of some research situations. For examples, research conducted with limited available time, often true of internet data collections, may oblige reducing the number of trials in the IAT(e.g., Lai et al., 2014, Study 4; Lai et al., 2016, all studies). Limited attention span of young children can similarly oblige reduced numbers of data collection trials (e.g., Cvencek et al., 2011).

2–6. Test–retest reliabilities of IAT measures can be improved by aggregation across the repetitions of the measure

The average of two separated administrations of any measure to the same person is expected to have greater test–retest reliability than does a single administration. The Spearman–Brown prophecy formula (Brown, 1910; Spearman, 1910) captures this statistical expectation:

$$r_k = k(r_1) / [1 + (k - 1)r_1],$$

where r_k is the test–retest reliability of the aggregate of k repetitions of the measure and r_1 is the test–retest reliability of a single administration (i.e., the correlation between two separated single administrations). When $k = 2$, the formula is $r_2 = (2 \cdot r_1) / (1 + r_1)$. For a single measure with test–retest reliability = .5, $r_2 = (2 \cdot 0.5) / (1 + 0.5) = .67$. Applying the same formula for three and four repetitions, $r_3 = .75$ and $r_4 = .80$.

The initial study of the *D* scoring algorithm revealed that, for three of the four measures for which tests were conducted, the first administrations of the IAT showed a significantly larger mean displacement from zero than did subsequent administrations (Greenwald et al., 2003, pp. 210–211; see also 1–B12 and 2–4). A useful parallel to this observation is found in cardiology research, where initial blood pressure readings regularly show higher values than ones taken relatively soon thereafter in the same setting. Because of the low test–retest reliability of single sphygmomanometer administrations (lower than those for single IAT administrations), medical studies of hypertension routinely average multiple blood pressure readings on each testing occasion. Use of 3 or more blood pressure measures separated by at least a few minutes is standard in the hypertension research literature (see Stergiou, 2002). Recent data showing

increase of IAT measures' test–retest reliability afforded by repeated administrations of IAT measures are described in 3–11.

2–7. Socially sensitive attitudes and stereotypes are often more polarized and more pervasive when measured by IAT than by parallel self-report measures

In the first large sample study of internet-administered attitude and stereotype IAT measures (Nosek et al., 2007), a striking difference between IAT and parallel self-report measures was documented: IAT measures were generally more polarized (further from neutral) than were parallel self-report measures. Measured in standard deviation units (Cohen's *d*), Nosek et al. found that the mean self-report race attitude measure was $d = 0.31$, which revealed a level of explicit preference for racial White that is conventionally between small and medium.[7] The race attitude IAT for the same respondents (N=732,881) had a mean *d* of 0.86, revealing a substantially more polarized average level of implicit (than explicit) White preference.

Applying statistics of the normal distribution, a Cohen's *d* of 0.31 corresponds to 54.4% of a population having at least a conventionally small ($d = 0.2$) level of self-report-measured White preference; $d = 0.86$ corresponds to 74.5% (20% more) having more than conventionally small IAT-measured White preference. Across 14 social attitude and stereotype measures for which Nosek et al. (2007) reported findings (sample sizes ranging from 28,816 to 732,881) the weighted mean absolute value of *d* for was 0.51 self-report measures, compared to 0.87 for IAT measures. Mean IAT was more polarized than mean self-report for 12 of the 14 attitude or stereotype topics. In contrast, for three political attitude topics IAT means were less polarized than were self-report means.

The greater polarization of IAT than of self-report measures of social attitudes and stereotypes indicates that, for these topics, respondents generally had stronger IAT-measured than self-report-measured attitudes or stereotypes. A necessary consequence is that, for IAT measures in comparison to self-report measures of these attitudes and stereotypes, more respondents meet criteria for having scores that deviate from neutrality—meaning that the data reveal that implicit biases are more pervasive than explicit biases. There are occasional important exceptions to this generalization, including the observation that some groups have stronger ingroup-favorable explicit than implicit attitudes (see 4–13).

2–8. IAT measures are almost invariably positively correlated with parallel self-report measures, but these correlations vary widely in magnitude

A very large proportion of published correlations between IAT and parallel self-report measures are numerically positive, meaning that the two types of measures tend to agree in direction. In a meta-analysis of 126 studies, Hofmann, Gawronski, Gschwendner, Le, & Schmitt (2005) reported an average correlation between IAT and self-report measures of $r = .24$. In Nosek et al.'s (2007) data set (see 2–7), correlations between IAT and self-report measures were positive for all 17 included IAT measures, with weighted average $r = .27$, but varying from

[7] The established conventions for Cohen's *d* interpret values of 0.2, 0.5, and 0.8, respectively, as small, medium, and large effect sizes.

r = .13 for age attitude to *r* = .75 for attitudes toward the main candidates (George W. Bush and Albert Gore) in the U.S. presidential election of 2000. Greenwald et al.'s (2009) meta-analysis found positive average correlations between IAT and self-report of .21, based on 155 independent samples, varying from .09 (for 10 studies involving close relationships) to .54 (for 9 studies of political preferences). In 57 experimental studies, Nosek (2005) reported an average correlation of .36, ranging from –.05 to .70 (p. 572). For 95 experimental studies, Nosek and Hansen (2008) reported an average correlation of .36, ranging from .07 to .70 (p. 579). Kurdi et al. found an average correlation of .120 (personal communication), for 160 studies limited to the domain of intergroup behavior.

2–9. Magnitudes of correlations predicting attitude-relevant behavior from IAT attitude measures are consistently positive, but vary widely in magnitude

The only meta-analysis that assessed predictive validity data for IAT measures in a diversity of domains found substantial variation across domains in magnitude of these correlations (Greenwald et al., 2009). The nine domains reviewed by Greenwald et al. are listed here in order of increasing weighted average predictive validity correlations (with 95% confidence intervals and number of independent samples [*k*] indicated): close relationships (*r* = .171 ± .094, *k* = 12), gender/sexual orientation (*r* = .181 ± .081, *k* = 15), intergroup behavior not including race (*r* = .201 ± .093, *k* = 15), alcohol and drug use (*r* = .221 ± .069, *k* = 16), Black/White race (*r* = .236 ± .062, *k* = 32), personality traits (*r* = .277 ± .064, *k* = 24), clinical measures (e.g., phobia and anxiety) (*r* = .296 ± .068, *k* = 19), consumer preferences (*r* = .323 ± .049, *k* = 40), and political preferences (*r* = .483 ± .071, *k* = 11). Overall, the weighted average predictive validity correlation was *r* = .275 (± .029, *k* = 184).

2–10. Predictions of discriminatory judgments and behavior by IAT attitude measures are, on average, small

Three of the nine domains in the Greenwald et al. (2009) meta-analysis involved intergroup behavior (gender/sexual orientation, Black/White race, and other intergroup behavior). These studies, about one third of those in the meta-analysis, had relatively small aggregate correlational effect sizes (.181, .201, and .236, respectively). Two subsequent meta-analyses focused specifically on predictive validity of IAT measures in the domain of intergroup behavior. Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013) included only studies involving race or ethnicity, finding an aggregate effect size of *r* = .140 for 86 samples. Kurdi et al. (2019) added studies involving gender, sexual orientation, overweight, and disabilities (physical and mental). While not reporting an overall aggregate effect size, Kurdi et al. noted that a large proportion of the studies they reviewed were deficient in attending to reliability of measures and power of studies, urging future researchers to contribute methodologically stronger studies.[8]

In combination, the three meta-analyses establish that predictive validity correlations of IAT measures in the domain of intergroup discrimination are relatively small. Small correlations between IAT measures and measures of intergroup discrimination are consistent with an assumption that discriminatory judgments and behavior have important determinants in addition

[8] Benedek Kurdi reported in a personal communication that this aggregate correlation was *r* = .097.

to the possible roles of implicitly measured attitude and stereotype associations. The roles of attitudes and stereotypes, in relation to other plausible determinants, are considered further in 3–8 to 3–11. The practical and societal significance of these correlation magnitudes is considered in Section 3–3.

2–11. Correlations of IAT attitude and IAT stereotype measures with intergroup discrimination are stronger when the intergroup discrimination measure *compares* judgment or behavior toward the two categories contrasted in the IAT

The two IAT meta-analyses that were limited to intergroup behavior (Kurdi et al., 2018; Oswald et al., 2013) both found that correlations between IAT measures and criterion measures of intergroup discrimination were stronger for criterion measures obtained in relative rather than absolute fashion. That is, correlations were greater when the criterion measure compared behavior or judgment toward the IAT's two contrasted groups (i.e., target concepts), rather than having a measure of behavior directed toward just one of the two groups. For measures referencing only one of the two groups, correlations were positive but weaker, and were stronger for measures of behavior toward the stigmatized group than toward the non-stigmatized group. This methodological moderator was similarly found to be influential in studies of correlations between IAT and parallel self-report in the meta-analyses of Hofmann et al. (2005) and Greenwald et al. (2009).

This observation of stronger correlations for measures of relative behavior toward the two target concepts may also apply to IAT–criterion correlations for political and consumer preferences. However, because so few studies in the political and consumer domains have used non-relative criterion measures, the comparison between relative and non-relative criterion measures is not effectively testable for those domains.

2–12. Predictive validity correlations of IAT measures are higher to the extent that IAT measures and parallel self-report measures are positively correlated

Three meta-analyses reported correlations of IAT measures with behavioral measures (implicit–criterion correlations: ICCs) and with parallel self-report measures (implicit–explicit correlations: IECs). These ICCs and IECs were significantly positively correlated in the two meta-analyses that reported this relationship (Greenwald et al., 2009, Table 4; Kurdi et al., 2019, p. 11). Larger IECs were also accompanied by significantly larger correlations of self-report attitude measures with behavior (explicit–criterion correlations: ECCs) in the Greenwald et al. meta-analysis. The absence of this finding by Kurdi et al. could have been due to the limitation of Kurdi et al.'s study to intergroup behavior, which has narrower ranges of observed correlations for all three types of correlation (ICC, IEC, and ECC). Stronger IECs and stronger correlations for both IAT and self-report with criterion measures may occur when the constructs measured by IAT and self-report have more shared causes, evident in larger implicit–explicit correlations.

2–13. Correlations of IAT measures with relevant judgment and behavior criteria (ICCs) are less variable in magnitude than are correlations of self-report measures with those same criterion measures (ECCs)

This difference between ICCs and ECCs was observed both by Greenwald et al. (2009) and Kurdi et al. (2019). The greater variability in predictive validity correlations for self-report measures could be due to what Greenwald et al. (2002, p. 17) described as greater susceptibility of self-report measures to 'response factors', which include demand characteristics (Orne, 1962), evaluation apprehension (Rosenberg, 1969), and subject role-playing (Weber & Cook, 1972). IAT measures appear relatively free of those influences. That relative immunity was evidenced by a substantially smaller negative influence of topic social sensitivity on predictive validity correlations for ICCs than for ECCs in the Greenwald et al. meta-analysis.

2–14. When used in combination with self-report measures to predict discriminatory behavior, IAT measures provide incremental validity

All three predictive validity meta-analyses reported that, although IAT measures and parallel self-report measures were correlated as predictors of intergroup discriminatory behavior, they were not entirely redundant predictors. That is, both IAT and self-report significantly predicted criterion variance that was not predicted by the other. Kurdi et al.'s (2018) meta-analysis confirmed this mutual incremental predictive validity, improving on the methods of Greenwald et al.'s (2009) and Oswald et al. (2013) by using a structural equation method described by Westfall and Yarkoni (2016).

2–15. Predictions of balanced identity theory are more strongly confirmed with IAT measures than with self-report measures

Building on Heider's (1958) balance theory, balanced identity theory (BIT: Greenwald et al., 2002) predicts correlational patterns involving attitudes, stereotypes, self-esteem, and social identities. Greenwald et al. found that these correlational predictions were confirmed when IAT measures were used to assess the four types of constructs. Greenwald et al. did not find these predicted correlations in parallel tests using self-report measures. This same pattern of confirmation of BIT predictions with IAT measures and not with self-report measures was also obtained in a subsequent small meta-analysis by Cvencek, Greenwald, and Meltzoff (2012). However, a larger meta-analysis by Cvencek et al. (submitted) did find statistical support for BIT's predictions with self-report measures, although the support was significantly weaker than that obtained when IAT measures were used. Stronger confirmation of BIT predictions for IAT than self-report measures was also found in a study with elementary school children in Grades 1–5 (Cvencek, Greenwald, & Meltzoff, 2011), indicating that the associative knowledge underlying implicit measures may already be established by early childhood.

2–16. Order of measuring IAT and self-report measures in research studies does not systematically influence magnitude of observed IAT effects or magnitude of correlations between IAT and self-report (or other) measures

The effect of order of administration of IAT and self-report measures was examined as a procedural variable in both the Greenwald et al. (2009) and Kurdi et al. (2018) meta-analyses. Neither found consistent effects of the order in which the two types of measures were administered. However, the meta-analytic findings justify only concluding that such order effects do not occur on average — they do not justify a conclusion that such order effects *never*

occur. Considering the ease of varying the order of these two types of measures in most research studies, there seems little reason not to continue the routine practice of most investigators, which has been to counterbalance the order of administering IAT and self-report measures.

2–17. IAT performances can be faked when subjects are instructed on how to fake—this faking is at least partly detectable

The possibility of faking on IAT measures was first tested by Banse, Seise, and Zerbes (2001) and by Kim (2003). Although few subjects spontaneously discover how to fake effectively when asked to fake an IAT result, most subjects can easily follow the (effective) instruction to give slow responses for one of the IAT's two combined tasks. For example, a faked preference for insects relative to flowers can be obtained by asking subjects to respond slowly in the combined task that requires response on the same key for flower names and pleasant words. Cvencek, Greenwald, Brown, Snowden, and Gray (2010) conducted tests of instructed faking on gender-identity IATs. They found that most subjects were able to fake an opposite implicit gender identity (association of opposite of own gender with self) when instructed to respond slowly in a combined task that required the same key press for own-gender names and words referring to self. Cvencek et al. went further to develop a statistical indicator of such deliberate slowing, showing that this index had a 75% success rate in (blindly) detecting both instructed faking of group identities (e.g., gender identity, national identity, etc.) and uninstructed motivation to fake (e.g., convicted pedophiles likely wishing not to be identified as pedophiles by an IAT measure). Agosta, Ghirardi, Zogmaister, Castiello, & Sartori (2011) similarly reported success in using a statistical method to identify fakers in the aIAT, a lie-detector application of the IAT developed by Sartori, Agosta, Zogmaister, Ferrara, & Castiello(2008).

2–18. Multiple interventions proposed to reduce implicit biases have produced desired effects when those effects are measured immediately, but not when the test of intervention impact is delayed by a day or more

The term 'malleability' first appeared in studies of effects of interventions created to alter implicit measures of attitudes or stereotypes in two articles published in 2001 (Dasgupta & Greenwald, 2001; Rudman, Ashmore, & Gary, 2001). Dasgupta and Greenwald had observed an effect of an intervention that they found to be evident 24 hours later. Rudman et al. found that, at the end of a semester (compared to the beginning of the semester), "students enrolled in a prejudice and conflict seminar showed significantly reduced implicit and explicit anti-Black biases, compared with control students" (p. 856). In their review that included fourteen studies that using IAT intergroup attitude measures, Blair (2002)concluded that there was "a strong case for the malleability of automatic stereotypes and prejudice" (p. 242). In Lai et al.'s (2014) reported of a multi-laboratory study, 'malleability' was used multiple times to describe findings that "[e]ight of 17 interventions were effective at reducing implicit preferences for Whites compared with Blacks" (p. 1766). These interventions had been tested "an average of 3.7 times each in four studies with combined N of 17,021" (p. 1766).

The Oxford English Dictionary defines 'malleable' as "capable of being hammered or pressed out of shape without a tendency to return to the original shape". The conclusion that

changes in IAT measures following brief interventions indicate *malleability* of implicitly measured attitudes or stereotypes supposes that the observed changes are durable. That supposition was later shown to be premature. Thirteen of the 14 studies reviewed by Blair (2002) along with all eight of those later reported as successful interventions by Lai et al. (2014) tested effects of the interventions on IAT measures only during the same relatively brief session in which the intervention was administered. In a subsequent series of studies Lai et al. (2016) found that "all [of the 8 previously effective] interventions immediately reduced implicit preferences. However, none were effective after a delay of several hours to several days" (p. 1001). This important follow-up observation effectively removed the basis for the previous conclusions that about malleability. The one study in Blair's (2002) review that had used a delayed test was by Dasgupta and Greenwald (2001), who observed a (barely) significant effect ($p = .049$) after a 24-hour delay. That isolated finding by Dasgupta and Greenwald is now best treated as a possible (even likely) Type I error. [9] The question of malleability bears important questions about effectiveness of methods now actively being offered for use to reduce implicit biases in many workplaces (see 5–1 to 5–3).

[9] It is difficult to know how to understand 'delay' before posttest in the semester-long courses of the Rudman et al. (2001) study. These posttests were administered while the course was still in progress. There is a similar problem in understanding posttest delay in other studies of multiple spaced interventions when the posttest was conducted in the same setting in which interventions were administered.

3. THEORETICAL UNDERSTANDING OF THE IAT

Interpretation of the IAT in its 1998 initial publication was limited to the bare statement—contained in the article's title—that the IAT provided a measure of association strengths. A reviewer of the initial submission of that article recommended rejection because the article offered no theoretical interpretation of its findings (an entirely correct observation). Fortunately, the editor invited a revision, leading to eventual publication. The theoretical understanding that now exists is presented here as a list of questions. Some of these have confident answers; others have controversial answers.

3–1. Are results obtained on IAT measures consciously controllable?

There is no doubt that performances on the IAT's two combined tasks (Blocks 3, 4, 6, and 7 of the standard procedure) require processes that are generally understood as conscious processes, including: (a) attention (needed to perceive the stimuli), (b) decision making (needed to select the correct response for each stimulus), and (c) working memory (needed to retain the instructions for the two response keys). A more theoretically developed conception of conscious process involvement in the IAT is offered in Section 3–9's consideration of the quadruple process model (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005). The role of automatic processes in the IAT is assumed to be in influencing speed of response selection for decision making when exemplar and assigned category are strongly associated, or in interfering with speed of response when they are not. The possible role of association strengths in these presumed-automatic processes is considered more fully in 3–4, 3–5 and 4–2.

Even in the presence of automatic processes that affect IAT performance, subjects can (consciously) control their IAT scores by deliberately slowing their responding in one or the other combined task. Additionally, numerous non-durable intervention effects reported in published research (see 2–18) may involve some conscious (even if unidentified) influences on IAT performance. There is no evidence, however, that IAT responders can control their scores by trying to *increase* speed (relative to performance following standard instructions) in the combined task that they find more difficult. Studies of deliberate faking (see 2–17) have found that few subjects spontaneously discover the effective strategy (selective slowing) when they are instructed to try to fake an opposite-from expected result. However, almost all can fake when provided with the needed instruction.

3–2. Are constructs measured by the IAT consciously accessible?

In general, it is difficult to empirically answer questions of the form: "Does X occur unconsciously or consciously?" Past research has addressed variations on this question in which "X" was replaced by memory of various forms, perception of visually masked stimuli, problem solving, hypnotic suggestion, conditioning, learning of artificial grammars, and processing of unattended stimuli. In most of these research areas, long-running debates about whether established phenomena occur with or without conscious cognition continue without resolution.

In implicit social cognition, there has been much speculation about the role of conscious versus unconscious process, but no attempt to address this question empirically until a set of

studies reported by Hahn, Judd, Hirsh, and Blair (2014). Hahn et al. sought to establish that their subjects had introspective (i.e., conscious) access to knowledge that provided the basis for their performance on IAT measures. Hahn et al. did this by showing (a) that their subjects could predict relative degrees of preference that would be shown on five IATs that they had not yet completed and (b) that sources of knowledge on which their subjects could self-report (e.g., explicit thermometer ratings of attitude) predicted less well than did the forecasted IAT relative preferences. Hahn et al. concluded that their findings were "contrary [to the positions of] most academic and popular representations, [in which] implicit attitudes are portrayed as 'unconscious' and inaccessible to introspection" (p. 1389).

The history of attempts to establish presence or absence of a conscious cognitive basis of any putatively unconsciously controlled behavior strongly suggests that the question to which Hahn et al. (2014) addressed their research will not soon have a consensual answer. As just one example of possible further research, it may be useful to investigate differences between subjects who do and do not show defensive reactions of the type reported by Howell, Redford, Pogge, and Ratliff (2017). Defensive reactions presumably indicate surprise on discovery of one's IAT results, in turn suggesting lack of access to the knowledge that, if introspectively accessible, would allow them to predict their unwelcome IAT preferences (also noted by Hahn et al., p. 1389).

3–3. Is prediction of intergroup discrimination by IAT measures statistically too weak to be of practical value?

In considering their meta-analytic findings, Oswald et al. (2013) concluded that "the IAT provides little insight into who will discriminate against whom" (p. 188), and Oswald et al. (2015) similarly concluded that "IAT scores are not good predictors of ethnic or racial discrimination, and explain, at most, small fractions of the variance in discriminatory behavior in controlled laboratory settings" (p. 562). As described in 2–10, two meta-analyses in addition to theirs have established that predictive validity correlations in studies of intergroup discrimination are small. In response to their assertion that demonstrated predictive validity correlations were too small to be of practical significance, Greenwald, Banaji, and Nosek (2015, pp. 557–560) presented statistical simulations establishing that the effect sizes even smaller than the magnitudes found in all three meta-analyses (Greenwald et al., 2009; Kurdi et al., 2018; Oswald et al., 2013) "were large enough to explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons" (p. 553). Oswald et al. (2015) did not contest the validity of Greenwald et al.'s simulations but maintained their belief that demonstrated predictive validity correlations were not 'large enough' to have 'substantial societal significance' (p. 565).

3–4. How does the IAT work to measure association strengths?

Sections 2–2 and 2–3 described IAT measures as providing a *relative* measure of strengths of complementary dimensional associations. 'Strength' is easily understood as a characteristic of (physical) forces such as those produced by muscles, engines, and explosives. The same does not apply when 'strength' is used to describe mental forces produced by habits, attitudes, and associations. Even those who have professional education on these mental constructs may lack

intuitive understanding of their strength aspect. Necessarily, therefore, researchers rely on operational definitions.

Operational definitions of association strength often rest on the assumption that speed of responding an unpredictable stimulus measures strength of associations that link the stimulus to its appropriate response. For one example, in the evaluative priming task (Fazio et al., 1986) subjects are asked to categorize words as pleasant or unpleasant using two different keys. They respond to multiple (target) words, each preceded by a (prime) stimulus that they are (often) instructed to ignore. Variations in speed of responses to target word stimuli as a function of their evaluative congruence with valences of prime stimuli are taken to measure variations in strengths of associations between the prime and target stimuli. The IAT's combined tasks provide a more complex example, involving four categories of words and instructions to respond to words of two of those categories with a left key and the two other categories with a right key. IAT results are taken to indicate stronger association between the pairs of categories that share keys in whichever of the two combined tasks (see Appendix A) is performed with greater speed and accuracy.

Figure 1 schematizes associations involved in responding to a gender–science stereotype IAT for a person who is assumed to have associations of *male* with *science* and *female* with *family*. Figure 1 captures two levels of associations. The associations between categories (*female* with *family*; *male* with *science*) are assumed to have been formed by many experiences of encountering (in life or in media) male people more frequently than female people in scientific roles and female people more frequently than male people in family roles. At the second level, associations of categories to their exemplar stimuli are assumed to be established by many experiences of contiguity in text and speech between these exemplars and their associated category labels.

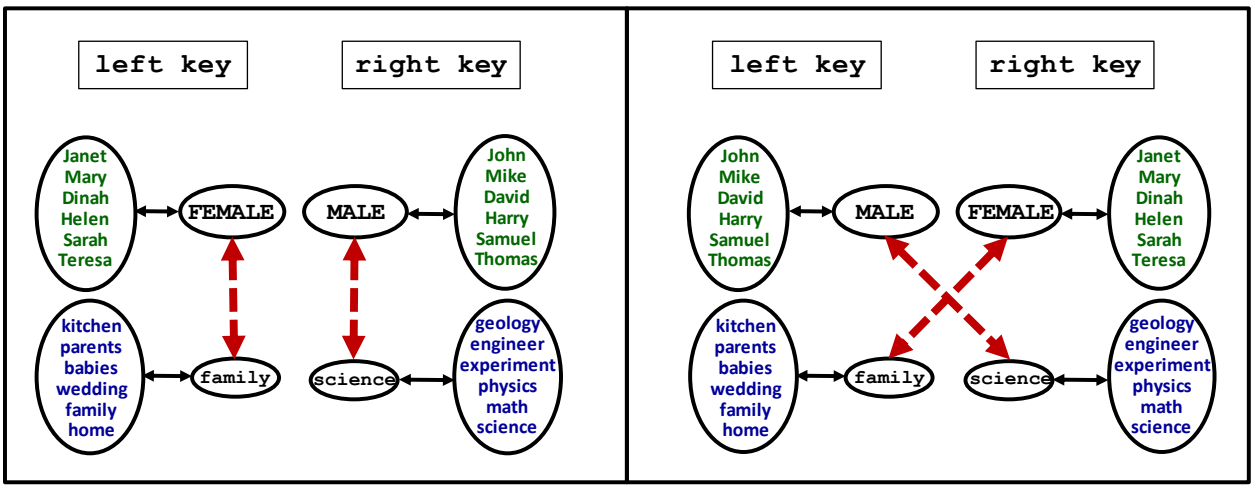

Figure 1. Representation of associations involved in responding to an IAT gender–science stereotype measure. The left panel shows four categories in a stereotype-consistent structure, with associations linking all categories and exemplars for which instructions request response to each key. The red arrows represent the stereotype-consistent associations. In the right panel, these associations cross between the keys, comprising a source of interference in giving the instructed responses.

A minimal theoretical interpretation of the IAT is that, when two categories are associated, the association between those categories makes it easy to give the same keyboard response to exemplars of both. When instructions assign the same key to *female* and *science*, the *female=family* association should interfere with producing the instructed key-press required for the *science* category on trials that present *female* exemplars.

One way to understand the effect of association between categories (e.g., *female–family* or *flower–pleasant*) is that these associations make it simple to retain instructions for a combined task in which two associated categories are assigned to the same response key. In combined-task blocks in which two non-associated categories require the same key response, IAT respondents often pause between trials, perhaps because the response-key instructions were momentarily lost from working memory, requiring active mental retrieval. No such pause for retrieval may be needed when the two categories are associated.[10]

3–5. How do association strengths measured by the IAT influence social behavior?

The challenge to answer this question has been addressed in multiple dual-construct theoretical conceptions. The two constructs in such theories often distinguish modes of mental operation that may play separate or joint roles in determining social behavior. The duality has been formulated in terms of mental representations (e.g., associations vs. propositions), or mental processes (automatic vs. controlled), or systems (impulsive vs. reflective), or research operations (implicit vs. explicit), or abstract categories (e.g., Type 1 and Type 2). The large number of these dual-mode formulations makes it impractical to mention more than a few in this article. Consideration of similarities and differences among them can be found in an overview by Stanovich, West, and Toplak (2014). Often, one of the two modes is conceived as being simpler, faster, and operating without awareness, while the other is conceived as more complex, slower, and accompanied by awareness. Most of these dual-construct conceptions are flexible enough so that the alternative conceptions rarely appear to be empirically at odds with one another. Most of the conceptual development going beyond the *automatic* versus *controlled* distinction adopted by many cognitive psychologists following Shiffrin and Schneider (1977) has been a re-branding of the two constructs. The theories have mostly not generated novel empirical predictions.

In the following, the two modes will be referred to as *associative* and *rule-based* (borrowed from Sloman, 1996 and Smith & DeCoster, 2000). These terms are used to simplify reference, with no intent to suggest preference among the multiple available dual formulations. Among dual-construct conceptions, Strack and Deutsch (2004; 2012) most directly considered both (a) how associative knowledge might, by itself, produce behavior and (b) how it might cooperate with rule-based knowledge in producing behavior. They suggested two possibilities for direct causation of behavior by associative knowledge. One is *ideomotor action*, by which the thought or perception of an action may elicit performance of that action (James, 1890). The second was the hypothesis that "semantic concepts can be directly connected to motor programs", for which

[10] This interpretation has never been validated by empirical test. However, an easy test should be available within many existing data sets. Finding that the more slowly performed combined task contains occasional trials with substantially longer latencies than the mean for that task would be consistent with the hypothesis of pauses to retrieve instructions.

Strack and Deutsch offered Bargh, Chen, and Burrows's (1996) finding of *stereotype activation* as an example.

In addition to a direct path from associative knowledge to behavior, there are two other forms of explanation for the frequent findings of correlations between IAT measures and behavior. One is that association strengths measured by the IAT and the behaviors with which they correlate are shaped by some (perhaps many) of the same influences. The other is that associative and propositional processes—to use Gawronski and Bodenhausen's (2006; 2011) designations—may cooperate in influencing behavior. An example of this last type of explanation is Greenwald and Banaji's (2017) proposal that IAT-measured association strengths may influence attitude-relevant judgments and behaviors by (automatically) shaping the content of conscious thought. The resulting associatively shaped conscious thoughts may more immediately guide the judgments and decisions that produce correlated behavior. Available research findings do not now provide a basis for preference among these three types of explanations—automatic effects on behavior, shared influences, and cooperative causation.

3–6. Can the IAT establish that a concept is associated with positive or negative valence?

The IAT was identified as a *relative* measure in its initial publication ("The IAT effect index is proposed as a measure of subjects' *relative* [emphasis added] implicit attitudes toward the categories under study" [Greenwald et al., 1998, p. 1468]). There have been multiple attempts to produce an IAT measure that would reveal the valence associated with a *single* concept—in effect, an *absolute* measure of a concept's associated valence. These attempts are, at best, approximations—IAT measures that likely are closer to an absolute valence measure than is the standard IAT's relative measure. The next paragraph describes the success (or non-success) of these approximations.

The first attempt (Gemar, 2001) was to subdivide IAT trials into those on keys representing positive valence and negative valence in each combined task. Nosek et al. (2005, Study 1) put this strategy to a comprehensive empirical test, finding that the IAT could not effectively be decomposed in this fashion. Other attempts sought to establish that a target category is more associated with positive valence than with *neutral* valence (alternatively, more with negative than with neutral valence). Those attempts encountered two barriers: First, it may be impossible to select exemplars for 'neutral' that are totally lacking in valence (D. E. McGhee, 2001, described in Pinter & Greenwald, 2005). Second, even if a set of neutral exemplars could be identified, their use would violate the recommended practice (1–A3) of avoiding multiple bases for discriminating between two contrasted categories. That is, a neutral category would differ from a positive or negative valence category not only in position on the valence dimension, but also in presence vs. absence of valence, and likely also in presence vs. absence of one or more properties of the presumed-neutral items (e.g., the possibly 'neutral' *middle* and *center* categories have a spatial position attribute that would likely be absent from exemplars of a contrasted non-neutral valence category). Still another approach is to construct 'single-category' variants of the IAT, including the Single-Category IAT (Karpinski & Steinman, 2006) and the Single-Task IAT (Bluemke & Friese, 2008). Lastly, there is the Brief IAT (Sriram & Greenwald, 2009), which—even while using two contrasted target concepts—can focus attention

more on one of those than on the other. These last three options are plausibly closer to being measures of absolute valence associations than is the standard IAT, but there is not yet empirical support for the conclusion that they effectively achieve this goal. Summary: There is not yet a variant of the IAT that can confidently be treated as providing an absolute measure of valence associated with a concept.

3–7. Does the IAT measure prejudice and racism?

Within a few years after the first publication of the IAT, the measure's creators and its most active developers stopped using the words 'prejudice' or 'racism' in published descriptions of what the IAT measured. Among the reasons for this rhetorical change were, first, the accumulation of early findings that made clear a divergence between what was revealed by IAT measures and what was revealed by parallel self-report measures that frequently accompanied IAT measures in research studies (see 2–7). Second, there was nothing about the IAT's procedure that would prompt subjects, while their classification latencies were being recorded by the IAT's procedures, to have in mind the hostility or antipathy that is central to most definitions of prejudice (cf. Greenwald & Pettigrew, 2014, p. 684). Third, an IAT score indicating preference for racial White relative to Black can be obtained by someone who likes both racial groups but likes Whites more. In contrast with IAT measures, self-report measures of racial attitudes often oblige subjects to actively contemplate hostile or disparaging statements about outgroups.[11]

3–8. What alternatives or additions to an associative strength interpretation of the IAT have been proposed?

Adding to Brendl et al.'s (2001) *criterion-shift* explanation of the IAT (see 2–3), Rothermund and Wentura (2004) described four alternatives to the association-strength interpretation of IAT measures.

(a) *Differential familiarity of stimulus items* in contrasted categories. Regarding this, Rudman, Greenwald, Mellott, & Schwartz (1999), Dasgupta et al. (2000), and Ottaway, Hayden, and Oakes (2001) had previously reported substantial evidence that familiarity variation of stimulus items, beyond the moderate level needed to assure that the exemplars could be easily classified (see 1–A2), was not a contributing factor to IAT measures.

(b) *Differential familiarity of the contrasted categories* themselves. Relevant to this, Greenwald and Nosek (2001) concluded that the IAT does not work well when category

[11] Prior to the IAT's existence, the phrase "unconscious prejudice" was used to describe results of studies in which racial stimuli were used as evaluative primes. In September 1997, the *Journal of Experimental Social Psychology* published a special issue of five articles on "unconscious stereotyping and prejudice". Similar references to 'prejudice' appeared in a few of the earliest publications using the IAT. The phrases 'implicit racism' and 'unconscious forms of prejudice' were each used once by Greenwald et al., 1998, see pp. 1475, 1476). In that article, Greenwald et al. also concluded that they had demonstrated "evidence for divergence of the constructs represented by implicit versus explicit attitude measures" (p. 1477). A not-yet-published manuscript that was cited in the initial IAT publication had the working title of "Measuring implicit racism using the Implicit Association Test". However, by the time that article was published, 'implicit racism' had been replaced by "automatic preference for White Americans" (Dasgupta, McGhee, Greenwald, & Banaji, 2000).

exemplars consist of entirely unfamiliar stimuli such as nonsense words or when they have no correspondence to familiar categories (see also 1–A1).

(c) *Figure–ground asymmetries between stimulus items* in contrasted categories or (alternately stated) *greater salience of stimulus items* in one of the two contrasted categories than the other. Greenwald, Nosek, Banaji, and Klauer (2005) evaluated this alternative interpretation in existing literature and in two experiments, finding that manipulated figure–ground asymmetries did not have the effects expected by Rothermund and Wentura.

(d) *Strategic recoding* of the IAT's combined-task instructions. In their 2004 article, Rothermund and Wentura observed that "any feature that helps to distinguish between the two groups of stimuli that are assigned to the different responses can be used for a strategic recoding that simplifies the task". "[I]n some published IATs . . . the two target categories and . . . the two attribute categories are easily distinguishable on the basis of just one dimension, for example, valence" (p. 158). Greenwald et al. (2005) observed that this strategic recoding interpretation could alternatively be stated in terms of association strengths (see 3–4 and Figure 1). Rothermund et al. (2009) have suggested a similar interpretation, writing that "recoding [can] occur automatically—that is, without a conscious plan or strategy (i.e., recoding can result from an implicit learning of covariations between features and responses)".

3–9. Can the IAT measure multiple psychological processes?

Proposals for non-associative contributions to IAT measures have evolved into multi-component theories. First of these was the multinomial *quadruple process* (Quad) model (Conrey et al., 2005). Conrey et al. (p. 471) described the influence of Jacoby's (1991) process-dissociation procedure (PDP) on their theory. PDP had previously been used to tease apart the two components of various dual-construct theories that were becoming influential in both cognitive and social psychology (see 3–5). The Quad model's four processes were labeled *activation of association*, *stimulus discrimination*, *overcoming associations*, and *guessing*. Klauer, Voss, Schmitz, & Teige-Mocigemba (2007) used an existing model of reaction times in 2-choice tasks (Ratcliff, Gomez, & McKoon, 2004) to explain response latencies in the combined tasks of an IAT measure. The three processes in their *diffusion model* were identified as *information accumulation* (also called drift rate), *threshold setting*, and *nondecision components*. Meissner and Rothermund's (2013) multinomial *ReAL model* had three processes, which they identified as *recoding* (Re), *association* (A), and *label-based* (category) *identification* (L).

The three multi-component models each included an associative component (called information accumulation in the diffusion model) and a decision component (called discrimination in later statements of the Quad model, called threshold setting in the diffusion model, and called identification in the ReAL model). Their agreement in having association and decision components notwithstanding, the three models differ substantially (described in the publications cited in the preceding paragraph) in how the latency and error data obtained in IAT measures are used to provide measures of the theorized component processes . These models have potential to improve usefulness of IAT data either (a) by separating the non-associative components to purify the measure of an associative component of the IAT or (b) by

demonstrating that the non-associative components are interesting and useful in their own right. As an example of the latter, the Quad model has been used to predict self-regulatory behavior (e.g., Sherman et al., 2008). There have been only a few attempts to produce a purified associative measure that can be compared with the IAT's *D* measure (e.g., Klauer et al., 2007; Klauer et al., 2010; Wrzus, Egloff, & Riediger, 2017).

3–10. Is the IAT's measure contaminated by individual differences in executive function?

The executive function of *task switching* was first identified as a possibly unwelcome influence on IAT measures by Mierke and Klauer (2001). Mierke and Klauer obtained this result when the IAT measure was computed in either untransformed or logarithmically transformed millisecond units (which were the standard forms for reporting IAT results before 2003). Mierke and Klauer (2003) and Cai, Sriram, Greenwald, and McFarland (2004) replicated Mierke and Klauer's (2001) finding for IAT measures in millisecond units, but both also found that use of the IAT's *D* measure with the same data substantially reduced the correlation between task-switching ability and IAT measures, rendering it a non-significant correlate of IAT measures. In a recent article that tested contamination of latency-based implicit measures of attitudes by individual differences in executive function, Ito et al. (2015) also found that one individual difference measure of executive function (task switching) had a small correlation with the IAT's *D* measure (although it was statistically significant with their $N \approx 500$ sample size).

In its initial publications, the Quad model was tested using an IAT procedure that imposed a response deadline to increase error rates (latencies are not used in computing the Quad model's parameters). Later Quad model publications used the IAT in its standard form (e.g., Gonsalkorale, Sherman, Allen, Klauer, & Amodio, 2011), enabling comparison of the Quad model's association parameters with the IAT's standard *D* measure. Some informative data relating diffusion model parameters to the IAT measures are presented in Klauer et al.'s (2007) article. They found that the IAT's *D* measure only slightly outperformed the diffusion model's drift (information accumulation) parameter in magnitude of correlation with a political attitude measure in two experiments. Their findings also described possibilities for using the diffusion model's non-associative parameters to identify predictable ("method") variance of IAT measures that should be unrelated to the attitude constructs. For the ReAL model, comparisons with IAT measures are not possible because of the substantial variations from the standard IAT procedure used in empirical tests of the ReAL model. Such comparisons require obtaining standard *D* measures alongside the modified IAT used to collect data to test the ReAL model.

Among the findings that might emerge from the types of model-comparison investigations just suggested are (a) confirmatory factor analysis results showing whether the IAT's D measure and the associative components of the various theory load on a common latent variable, (b) psychometric and correlational validity tests of associative components of the multi-component models alongside the IAT's *D* measure, and (c) tests determining whether statistical combinations of (associative and non-associative) parameters of the multi-component models can improve on the psychometrics or predictive validity of the IAT's *D* measure.

3–11. Is the IAT a measure of traits or of situations?

Responding to the typically moderate test–retest reliability of IAT measures, Payne, Vuletich, and Lundberg (2017) theorized that "most of the systematic variance in implicit biases appears to operate at the level of situations" and that "measures of implicit bias . . . are meaningful, valid, and reliable measures of situations rather than persons" (p. 236). Several commenters on their article in the same journal issue took issue with their proposition concerning "most of the systematic variance" and with the related assertion that the IAT measures "situations rather than persons". The moderate views preferred by these commenters distributed influence more equally between person effects and situation effects (see also Jost, 2019). Payne et al.'s critique focuses attention on two questions: First, to understand how situations contribute to IAT effects. Second, to understand how individual differences contribute to IAT effects.

Effects of a wide variety of experimental interventions are now understood as non-durable influences on IAT measures (see 2–18). As described in 2–6, blood pressure measures similarly have multiple sources of non-durable situational influences, which are listed here with arrows indicating the established direction of the short-term changes in blood pressure that they produce: recent meal (↓), arm above heart level (↓), long rest prior to measurement (↓), talking during measurement (↑), pain (↑), anxiety (↑), recent smoking (↑), recent coffee (↑), arm below heart level (↑), and physical activity prior to measurement (↑) (Smith, 2014; see other references in 2–6). These multiple situational influences notwithstanding, very useful measures of health-relevant individual differences in blood pressure are routinely obtained in research studies by averaging multiple administrations of the measurement procedure.

To determine whether individual-subject-aggregated IAT measures can similarly produce highly reliable measures, it was possible to use data from Lindgren et al.'s (2018) recent publication of a 2-year study in which up to eight IAT measures on each of three alcohol-use-related IATs were provided by each of about 500 subjects. An unpublished analysis of these data (made available by Kristen Lindgren) showed that test–retest reliability of IAT measures increased from *r* values near .50 (.47, .49, and .50 for the three IATs, using just the first two measures of each IAT) to values of .75, .77, and .79 when the aggregate of IATs 1, 3, 5, and 7 was correlated with the aggregate of IATs 2, 4, 6, and 8 (Ns of 134, 149, and 150). The average of all 8 measures obtained for these subjects had (in theory) a test–retest reliability of $r = .89$, applying the Spearman–Brown formula given in 2–6.

There is much evidence that IAT measures are sensitive to individual differences. This includes (a) many published reports of known group differences in IAT-measured racial, ethnic, political, consumer brand, and many more attitudes, (b) strong known-groups correlations of many implicitly measured associations with self, including associations of self with gender, sexual orientation, religion, smoking, drinking, universities (and more), (c) demonstrations of predictive validity of IAT measures, which correlate individual differences in IAT measures with measures of judgment and behavior, (d) correlations of many self-report measures of attitudes with parallel IAT measures (see Section 2.8), and (e) individual differences measures of motivation to act in non-prejudiced fashion (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Olson & Fazio, 2000), ethnocentrism (Cunningham, Nezlek, & Banaji, 2004) and political ideology (Jost, Banaji, & Nosek, 2004; Nosek et al., 2007) are all found to correlate with race attitude IAT measures. By comparison with the many available demonstrations of person

differences associated with implicitly measured attitudes and identities, it is difficult to think of any clearly established demonstrations of IAT-measured attitudes toward situations or IAT-measured associations of self with situations.

3–12. Does the IAT have a rational zero point?

On self-report attitude measures, higher numbers typically indicate greater liking or favorableness toward the attitude's object. For example, the numerically high end of a thermometer-format measure of attitude toward a political candidate indicates maximum warmth (i.e., favorability) toward the candidate while the low end indicates maximum coldness (unfavorability). If the measure is scored from –5 to 5, the middle value (0) may be labeled 'neither warm nor cold". This mid-point can be understood as a rational zero-point, dividing responses into favorable (>0) and unfavorable (<0) to the candidate. Similarly, the mid-point on the widely used Rosenberg (1965) self-esteem inventory, achieved by agreeing equally with self-praising and self-critical statements, is assumed to separate those who have positive vs. negative self-directed attitudes.

The zero-point of an IAT measure is different in type. One obtains a zero score on an IAT attitude measure by responding equally rapidly in the two combined tasks. The IAT differs from the single-object thermometer measure described in the preceding paragraph because it includes two attitude objects. Consider a political IAT that compares association of positive valence with Candidate A vs. Candidate B. In this, zero divides respondents into those having more positivity toward A and those having more positivity toward B. This zero-point is comparable to that for a thermometer-difference measure, in which one responds to a thermometer measure separately for each candidate. The thermometer difference indicates relative preference for the candidate with the higher thermometer score. This type of difference score typically correlates more strongly with vote choice than does either of its component individual thermometer scores.

Blanton and Jaccard (2006) proposed that zero point of IAT measures is "arbitrary" and that "the assumption that the zero point on the IAT measure maps directly onto the true neutral preference [e.g.,] for Whites over Blacks is dubious" (p. 34). Blanton, Jaccard, Mitchell, Strauts, and Tetlock (2015) went further to say that the zero point of the race attitude IAT should be placed at a numerically positive value of the *D* measure. Based on analyses presented in their article, they concluded that there is an average "right bias" (e.g., p. 1468) of the race attitude IAT's zero point of 1.5 standard deviations above the IAT measure's $D = 0$ value. Their calculated correction for the estimated 'right bias' would decrease the proportion of persons estimated as showing more than slight implicit White preference in the studies they reviewed (pp. 1472–1473) from an average of 83% (using an unaltered IAT *D* measure) to an average of 28%.

To empirically assess validity of the IAT's zero value as a rational zero point, Blanton et al. (2015) offered a regression analysis method in which race IAT scores were regressed onto other measures that Blanton et al. were confident had (on average) rational zero points. They expected these analyses to reveal "the mean IAT score one expects to observe among individuals who exhibit no behavioral preference for Whites versus Blacks". An average value of zero for the

intercept using their regression method should indicate lack of racial preference, meaning that "behavioral neutrality map[s] onto IAT neutrality" (p. 1471).

The "logic model" (p.1471) on which Blanton et al. (2015) based their regression method can be unpacked by (a) starting from the formula for the intercept of a bivariate regression, expressing both the IAT measure and its presumed-rational-zero-value predictor (X) in standard deviation (SD) units, then (b) using the formula to describe intercepts for regressions in both the direction tested by Blanton et al. and in the reverse direction:

$$\text{Intercept}_{\text{IAT}} = M_{\text{IAT}} - r_{\text{X-IAT}} \times M_{\text{X}} \quad (1)$$

where M_{IAT}, M_{X}, and $r_{\text{X-IAT}}$ are (respectively) mean of IAT, mean of predictor X, and the X–IAT correlation coefficient (see, e.g., Cohen, Cohen, West, & Aiken, 2003, p. 44). Reversing X and Y produces:

$$\text{Intercept}_{\text{X}} = M_{\text{X}} - r_{\text{X-IAT}} \times M_{\text{IAT}} \quad (2)$$

Equations (1) and (2) can be solved for values of M_{X} and M_{IAT} that would allow both intercepts to be zero, starting by setting the left sides of the two equations to 0, setting $r_{\text{X-IAT}}$ in both equations to the weighted mean value of that correlation observed in the data sets analyzed by Blanton et al. (2015). The solution must yield values for M_{IAT} and M_{X} that will produce the desired zero values of intercepts in both directions of regression. To obtain a value of $r_{\text{X-IAT}}$ for use with the two equations, weighted average values of $r_{\text{X-IAT}}$ were first computed from the two sets of studies for which Blanton et al. reported analyses (their Tables 4 and 6). These were, respectively, $r_{\text{X-IAT}} = .10$ (from their Table 4) and $r_{\text{X-IAT}} = .25$ (from their Table 6). Using either of those values, the simultaneous-equation solution for Equations (1) and (2) is that both M_{X} and M_{IAT} should equal zero—values of zero for both M_{X} and M_{IAT} will allow zero intercepts to be observed in both directions. When $r_{\text{X-IAT}} = 1.0$, zero intercepts in both directions can also be observed when $M_{\text{X}} = M_{\text{IAT}}$.

Data (generously provided by Hart Blanton) for the 37 regression analyses summarized in Blanton et al.'s (2015) Table 6 were used to compute individual-study intercepts for the regression of IAT on predictor , the weighted average of which was 0.51. Applying Blanton et al.'s logic, 0.51 is the mean IAT score (in SD units, corresponding approximately to $D = .20$) that one expects to observe among individuals who exhibit no explicit-attitude preference for Whites versus Blacks.

The regression method could also be applied in the reverse direction, leading to a weighted average intercept of −0.01, which calls for interpretation (equally applying Blanton et al.'s logic model) as the mean explicit race attitude that one expects to observe among individuals who exhibit no IAT preference for Whites versus Blacks. A result so close to zero indicates that the IAT's zero point *is* located at an appropriate rational-zero value.

Applying Blanton et al.'s logic in both directions of regression thus produces two mutually inconsistent conclusions. However, this is not actually paradoxical. Statistical understanding of regression intercepts obliges that, unless a regression involves two perfect measures (test–retest reliability = 1.0) that are perfectly correlated ($r = 1.0$), the intercepts will not be identical when

the direction of aggression is reversed. The data chosen by Blanton et al. were very far from meeting either the reliability or correlation criteria of perfection. The only reasonable conclusion was that their method was inappropriate.

A different theoretical basis for testing validity of zero points of IAT measures was available using balanced identity theory (BIT; see 2–15). BIT's *balance–congruity principle* makes two predictions that are expected to be confirmed using measures (of attitudes, stereotypes, identities, and self-esteem) for which zero values validly indicate absence of difference in complementary association strengths (cf. Greenwald et al., 2002, pp. 9–12). These predictions are testable in a 'balanced identity design' in which each subject completes a set of three measures for which BIT predicts these interrelations. These trios can be measures either of (a) identity, attitude, and self-esteem or of (b) identity, stereotype, and self-concept. Although multiple tests of BIT's two predictions using IAT measures have been consistent with the rational-zero assumption, single-study confirmations do not provide compellingly strong tests of the validity of the IAT's zero point.

Cvencek et al. (submitted) gathered 36 available tests of BIT's balance–congruity principle into a meta-analysis having adequate statistical power for tests of the IAT's rational-zero assumption. Their analytic strategy was to test the two zero-point-sensitive predictions either (a) with unaltered IAT measures or (b) with IAT measures altered by adding constants to displace their zero points. If the unaltered measures have valid zero values, tests using them should conform well to BIT's predictions, whereas tests conducted with displaced zero values should show poorer fit, increasingly so as the magnitude of displacement is increased.

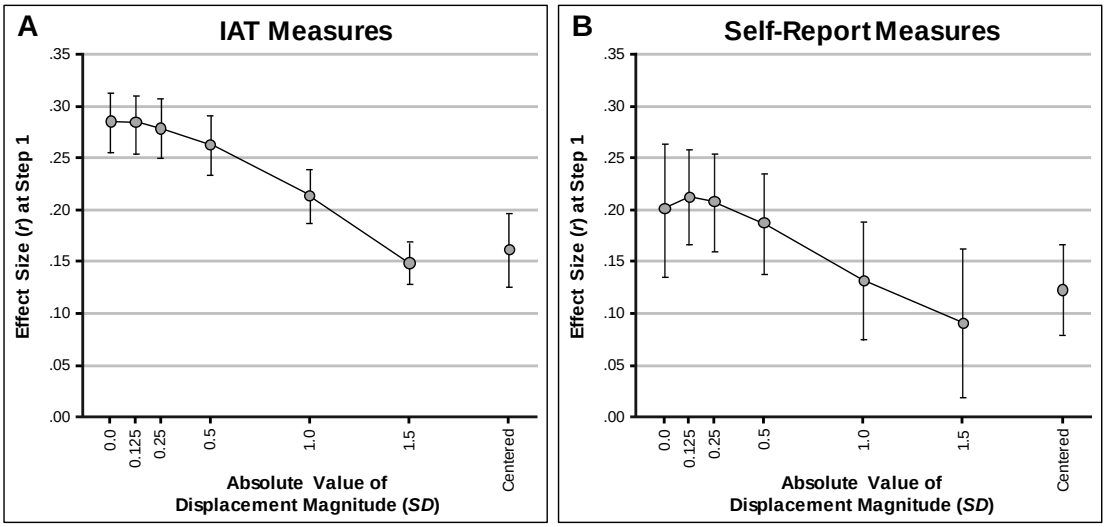

Figure 2. Weighted averages of correlational effect sizes for the prediction of each measure in a balanced identity trio as the product of the other two measures. A: for 36 studies done with IAT measures. B: for 16 studies that used parallel self-report measures. Data are presented for absolute values of displacements of the measures in each correlation, computed as a weighted average of effect sizes for positive and negative displacements. Error bars = 95% confidence intervals.

The first test is provided by an unusual prediction from BIT's balance–congruity principle—the prediction is that each IAT measure in a balanced identity trio of measures should be positively correlated with the *multiplicative product* of the other two measures (derived in

Greenwald et al., 2002, pp. 9–12). Figure 2A plots results of tests of this prediction, meta-analytically combined for the 36 studies. The plotted value over 0.0 on the X-axis (indicating no displacement of the predictor measures) shows that the weighted average of the predicted correlations for the 108 tests (3 in each study) was .285. The set of 108 tests was then repeated by displacing each IAT measure in positive or negative direction by values ranging up to 1.5 standard deviations. The figure shows that magnitude of the predicted correlation declined increasingly with increasing displacements—the deterioration was clearly noticeable with displacements [illegible] SD.[12] This finding provides confirmation of the validity of the location of the [illegible] similar pa[illegible] when it was applied [illegible] e available studies (see F[illegible] a, the mea[illegible] displacement [illegible] ticeable w[illegible] SD.

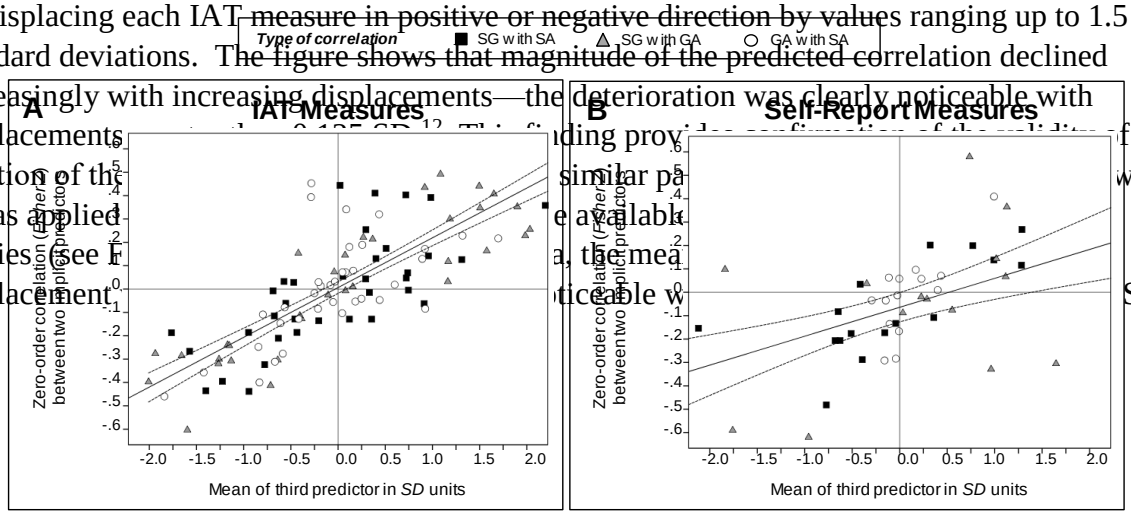

Figure 3. Plots of Fisher Z-transformed correlations between pairs of association strength measures in balanced identity studies, plotted as a function of the mean of the third measure in the design. Plots include regression slopes and their 95% confidence intervals. Distinct data point markers identify the type of correlation between two of the three association measures in each balanced identity designs: self–group (SG, identity), self–attribute (SA, self-esteem or self-concept), and group–attribute (GA, attitude or stereotype). For points representing each correlation, the X-axis gives the value (in SD units) of the mean of predictor variable in the design. Data are presented for IAT measures (panel A) and self-report measures (panel B).

The second prediction testing validity of IAT measures' zero points is a corollary of the first prediction: The correlation between any two of the three measures in a balanced identity trio is predicted by the mean of the third measure (see Greenwald et al., 2002, p. 10, for the derivation of this prediction). The prediction is that, if one of the three measures has a mean value of zero, the correlation between the other two measures should be zero; and if the mean of the predictor is numerically positive (alternately, negative), the correlation between the other two variables

[12] This one-eighth SD margin is consistent both with the rational zero assumption and with the observation (see 2–2) about effects of order of administration of the IAT's two combined tasks on IAT scores. Order effects have magnitudes averaging approximately 0.125 SD (see Greenwald et al., 2003, Tables 2 and 3); this expected individual-subject variation in the IAT's zero point fits with confirmation not deteriorating until displacements exceed that magnitude.

should be positive (alternately, negative), and increasingly so as the mean of the predictor is numerically larger. It follows that the regression of correlations between pairs of the three measures on means of the third measures should have a positive slope and should pass through the regression plot's origin. Figure 3A shows this regression plot for the 108 correlations between pairs of IAT measures in the 36 studies, each predicted by the mean of the third measure in the trio. The 95% confidence interval (CI) of the slope includes, as predicted, the plot's origin. The slope crosses the X-axis less than .05 SD from the origin, from which it can be concluded that displacement of the mean in either direction by more than a very small amount will decrease conformity to the zero-intercept prediction, and increasingly so with greater displacements. Figure 3B shows the same regression scatterplot for the 48 available correlations from the 16 studies that included self-report measures. For these, conformity to prediction is weaker than in Figure 3A—the 95% CI does not include the origin and the slope crosses the X-axis about 0.50 SD from the origin.

Accompanying their claim that the IAT's zero (= no preference) point is 'arbitrary', Blanton and Jaccard (2006) wrote: "If a researcher is interested in identifying the measured value corresponding to the true zero, one must . . . develop a theory that makes predictions about how data for other variables should pattern themselves as one moves across the dimension of interest and through the true zero point" (p. 34). That is what was done in the analyses of Figures 2 and 3, confirming the zero interpretation. It is unclear why a similar strategy was not used by Blanton et al. (2015) in their evaluation of the IAT's zero point. In their 2006 article, Blanton and Jaccard also recommended another strategy: "[R]esearchers could identify the IAT score that acts as a psychological dividing line between a behavioral preference for Blacks and a behavioral preference for Whites". However, this strategy can be appropriate only if one assumes both (a) that IAT-measured preference is the only cause of the behavioral preference (ignoring impression management and social influences operating in the behavioral observation situation) and (b) the IAT measure has no influences other than those that affect the observed behavior.[13]

3–13. Can IAT measures be treated as diagnostic of individual persons?

Concerns have been expressed (by both scientists and non-scientists) about the scientific and ethical bases for using IAT measures to describe characteristics of individual persons. The first of two scientific concerns is with reliability (i.e., accuracy) of the measures. As used most often in research, the IAT has insufficient precision for accurate characterization of individuals. A comparison with blood pressure measurement (see also 2–6 and 3–12) is relevant. Both IAT and blood pressure measures have only moderate test–retest reliability. A single blood pressure measure is not a trustworthy diagnostic measure for the person on whom it is taken. However, a trustworthy description of individuals can be provided by averages of multiple blood pressure administrations. The same is true for IAT measures, as was described in 3–12 using data from a

[13] Prepublication comments on this article by expert researchers in substance abuse expressed the view that one should expect alignment of behavioral indifference with IAT-measured indifference in substance use (e.g., between smoking and not-smoking). This is one of the problematic assumptions on which Blanton et al. rested their flawed regression method. In the substance abuse case, this argument does not consider the likelihood that IAT-measured preferences for smoking, even for addicted smokers, are formed in part by repeated encounter information that associates smoking with lung disease and early death.

2-year longitudinal study by Lindgren et al. (2018). A second scientific concern is with the weak relationship between IAT measures and discriminatory judgment or behavior measures. Even though these correlations magnitudes are large enough to be predictive of societally significant discrimination (see 3–3), they do not warrant a conclusion that an individual who is identified to have an implicit race preference (*even* if this assessed with high reliability using averaged IATs) will act in biased fashion due to the measured associations. This second scientific provides the prime basis for ethical concern. Should people be denied consideration for employment, for jury duty, for service as police officers, or for hiring as a manager or an executive based on an IAT measure that is weakly or moderately predictive of discriminatory judgment or behavior? Jost (2019) recently provided an overview of this and related ethical considerations.

IAT research studies most often report only analyses of group means or correlations based on groups of subjects. Most research studies therefore provide little cause for concern that IAT scores might (inappropriately) be interpreted as diagnostic of individuals. Nevertheless, the concern about diagnostic use for individuals does apply to some educational uses of IAT measures that are publicly, anonymously, and freely available via Project Implicit (https://implicit.harvard.edu/implicit/). The limited reliability of IAT measures may not be of great concern in this context, because information available on this public site explains that individual administrations should not be treated as definitive. One of the site's responses to frequently asked questions advises repetition of a test for which a result is doubted, and to average results of repeated takings of the same IAT. In the context of repeating tests and averaging results, the accuracy of the interpretation of zero values of IAT measures assumes importance. If the zero point is more than slightly mis-located, averages of repeated IAT scores can be misleading—therefore possibly misleading the site's visitors to believe that they have a directional automatic preference (e.g., an automatic preference for racial White relative to racial Black) that they don't have. In this respect the available evidence for validity of the IAT's zero-point (see 3–13) provides useful assurance.

4. QUESTIONS AWAITING ANSWERS

Much has been learned about the IAT in the 20 years since its first publication, and much remains to be learned. The questions in this section are the known unknowns. Description of each is limited mainly to indicating how the question relates to what is already known.

4–1. What is the (so far) hidden secret to producing measures of association strengths that are more effective than the IAT?

In "A comparative investigation of seven indirect attitude measures", Bar-Anan and Nosek (2014) summarized psychometric evidence for usefulness of the IAT and the six next most frequently used latency-based indirect attitude measures. Of the other six, only evaluative priming (Fazio et al., 1986) predated the IAT. Four of the other measures were ones that had been created as hoped-for improvements on the IAT. The remaining measure (Brief IAT) deliberately resembled the IAT. Bar-Anan and Nosek concluded, "The Implicit Association Test (IAT) and Brief IAT (BIAT) showed the best overall psychometric quality" (p. 668). Multiple other hoped-to-be-superior measures have been offered in publications, and there must be multiple other attempts that have never seen the light of publication. The task of developing a superior measure may become easier when the next two questions have answers.

4–2. What processes determine performance latency in the IAT's two combined tasks?

In the article "Correlated operations in searching stored semantic categories", David Meyer (1973) offered a serial information processing stage model to account for the article's interesting finding: "Ss judged whether or not a stimulus word belonged in either of two distinct semantic categories. Both positive and negative decisions were faster when the categories were close in meaning than when they were separated by a large semantic distance" (p. 124). Meyer offered an interpretation: "Perhaps searching a particular semantic category produces "excitation" that spreads to other nearby categories, so that after the shift from the first category to the second, the subsequent rate of searching the second category may be greater if it is close to the first one" (p. 129). This explanation did not mention 'associations' as being involved, but the reference to 'spread' of 'excitation' used language that was often used in the 1970s to describe operation of semantic associative networks.

Quoting Meyer (1973) further: "When Ss classify stimulus words with respect to pairs of categories whose semantic distance is either large or small, it is possible to observe the influence of one retrieval operation on another, as reflected through changes in RT. This outcome suggests that the strategy of investigation could be modified to study the organization of memory as well" (p. 132). Perhaps the IAT is such a 'modified' strategy 'to study the organization of memory'. A schematic model of the organization of a portion of memory that might contribute to such performance in the IAT's combined tasks was given in present Figure 1 (in 3–4). Although many research users of the IAT are content to describe the IAT as a measure of association strengths, the possible involvement of other processes, such as those described in 3–8 to 3–12, should also be considered.

4–3. How do the representations measured by the IAT influence behavior?

As described in 3–5, one possibility is that automatically activated associations directly cause correlated behaviors; another is that IAT-measured representations and correlated behaviors have no direct causal connection; rather they have the same or similar causes; and a third is that the mental representations assessed by IAT measures activate processes that mediate performance of the correlated judgments and behaviors. Greenwald and Banaji (2017) offered an undeveloped outline of an explanation in this last form, suggesting that multiple associations, acting simultaneously, might combined to shape conscious mental content that in turn directs judgment and behavior.

4–4. How should correlations between average IAT scores in one subset of a community and behavior of others in the community (in quasi-multilevel designs) be explained?

A result recently reported by Hehman, Flake, and Calanchini (2018) illustrates this question. Hehman et al. studied lethal shootings by police within 135 localized metropolitan regions in the U.S. Their analysis revealed that disproportionate killing of African Americans by police (relative to their proportion in the regional population) was correlated with average implicit race preference and average implicit race–weapons stereotype measured for *White* residents in the same geographic regions. The authors credited the implicit attitudes and stereotypes measured in regional White residents with an explanatory role in the disproportionate killings of African Americans by the region's police.

Because this finding has been reported only once it should gain support of replication before interpretations are offered confidently. If the finding proves replicable, some choices for its explanation are: (1) the finding is a consequence of police implicit biases shared with the community's White residents, or (2) in a region in which Whites have relatively strong automatic White preferences, police officers may be more explicitly biased against Blacks, or (3) any of many other possibilities. One other possibility is based on the known positive correlation of IAT scores shown by a region's White population and the proportion of Black persons in the region's population (Rae, Newheiser, & Olson, 2015). The greater Black population concentration might be associated with greater police presence and patrolling in predominantly Black neighborhoods. Because data in the form analyzed by Hehman et al. are not difficult to obtain—by combining the public archive provided by Project Implicit with other public (including Census) data sources—the choice among these three types of interpretations may arise frequently—there are not yet established methods for choosing among them. In another study using archival race IAT data (and similarly having multiple interpretive possibilities), Price and Orchard (2017) found that a Black–White difference in adverse birth outcomes (lower birth weight of babies and more preterm births for Black American mothers) was greater in U.S. counties with higher levels of implicit White race preference. Because Orchard and Price used race-related covariates that removed the effect of Black population percentage on their county-level dependent measures ("unemployment rate, the fraction of the population that are college graduates, the total population, the fraction of the population that is black, and the black poverty rate", p. 193) their findings do not actually show a correlation of the adverse birth outcomes with county *mean* IAT. Rather, they show correlation with a *corrected mean* that removes the contribution of African Americans' (low) White-preference IAT scores to the county mean. This observation indicates

the necessity for development of standards for how to report results from such quasi-multilevel research designs.

Nosek et al. (2009) reported results from aggregated national responses to a gender-stereotype IAT that typically reveals greater association of male (than female) with science (contrasted with arts). These national gender-stereotype measures correlated with male–female differences in 8th-grade children's math and science achievement for the respective countries. Other findings obtained by combining data from IAT responders archived by Project Implicit (available at http://osf.io)with measures available for other persons in the same regions can be expected as researchers mine public data sources to obtain measures plausibly related to the various IAT measures that are now publicly available. Methods for establishing interpretations of such findings remain to be developed.

4–5. In what order do implicit attitudes, identities, stereotypes, and self-esteem develop?

It is now possible to use the IAT to obtain implicit measures of self-esteem and identities for 5-year-old children (Cvencek, Greenwald, & Meltzoff, 2016). Cvencek et al. concluded that "By preschool age, children display fundamental properties of adult implicit social cognition that relate to maintenance and functioning of group identities" (p. 50). It would be very desirable to establish the earliest age at which self-esteem and identities can be revealed by implicit measures. This in turn could help to identify the experiences that establish the attitudes, stereotypes, and identities that are measurable by the IAT at age 5. Little can be said with great confidence about the order in which these social cognitions emerge until indirect measures are available for younger ages. Nevertheless, a useful indication is provided by observing the relative strengths of implicit identities, attitudes, and stereotypes of the youngest children who can be studied. Cvencek, Greenwald, and Meltzoff (2011) found that implicitly measured gender identities were stronger than implicitly measured gender stereotypes, suggesting that the gender identities had emerged earlier. Cvencek, Greenwald and Meltzoff (2016) found that that implicit self-esteem was at least as strong as implicit gender identity in preschool children, suggesting that self-esteem possibly develops earlier than gender identity.

4–6. Can implicit measures be predictively useful in longitudinal studies?

Although IAT measures are subject to multiple situational influences (see 3–12), these situational influences appear not to produce durable influences (see 2–18). The situational influences presumably contribute to IAT measures' relatively low test–retest reliability (see 2–5), but this is improvable by aggregating multiple IAT administrations (see 2–6). In combination, these observations suggest that IAT measures of attitudes, stereotypes, or identities are sufficiently stable to represent durable personal characteristics that may be useful as predictors in longitudinal studies. There have not yet been any published longitudinal investigations using IAT or other indirect measures of social cognitive constructs.

4–7. How can implicit biases be durably altered?

This question has been actively investigated for nearly two decades, using experimental studies of suspected implicit-bias-reducing intervention strategies. As was described in

considering this research (see 2–18), numerous interventions were inappropriately assumed to be successful in producing durable impacts on implicit biases. Lai et al. (2016; see 2–18), who found that these interventions were not effective when tested after delays of 24 hours or more, provided a useful review of research on modifiability of implicit biases. They appropriately mentioned some important starts at testing longer durability of intervention effects in studies by Devine and colleagues (e.g., Devine, Forscher, Austin, & Cox (2012) and by Dasgupta and colleagues (Dasgupta, 2013), but also concluded that these studies fell short of establishing methods that could be relied on to produce durable changes in implicit biases.

4–8. Are there still-to-be-discovered moderators of predictive validity of IAT measures?

Predictive validities of IAT measures (implicit–criterion correlations: ICCs) are positively predicted by the correlation between IAT and parallel self-report measures (Greenwald et al., 2009; Kurdi et al., 2019; see 2–12). A second known moderator is based on form of the judgment or behavioral criterion measure (see 2–11): ICCs are higher if the criterion measure is in relative form, such as an indicator of favoring members of one of two groups over members of the other (this was found both by Kurdi et al. and by Oswald et al., 2013).[14] A third moderator, 'social sensitivity', was based on a judgment by raters of "the extent to which self-reporting the construct assessed by the measure might activate concerns about the impression that the response would make on others" (Greenwald et al., p. 19). ICCs were higher when social sensitivity was low, which in part reflected the contrast between studies involving intergroup discrimination (high in social sensitivity, relatively small ICCs) and ones dealing with consumer behavior and politics (low in social sensitivity, relatively large ICCs).

Individual differences in personality variables may yet be found to moderate predictive validity of IAT measures. However, there has not yet been extensive use of personality measures as moderators in IAT predictive validity studies.

4–9. Will implicit self-esteem be established as a valid and useful individual difference measure?

Self-esteem continues to be widely assumed to play an important role in normal, healthy, social functioning, despite there being little consensus among theorists about how self-esteem functions in the normal personality. Theorized functions of self-esteem in the normal personality include (a) self-protection (ego-defense), (b) self-promotion (self-enhancement), and (c) identity formation and maintenance (Greenwald & Cvencek, in press). Measurement methods for self-esteem include open-ended, checklist, and Likert response formats along with much less widely used indirect methods (Bosson, Swann, & Pennebaker, 2000). Usefulness of the set of available measures has been disputed for both explicit self-esteem (e.g., Baumeister, Campbell, Krueger, & Vohs, 2003) and implicit self-esteem (e.g., Bosson et al.; Buhrmester, Blanton, & Swann, 2011). The strongest present indicator of usefulness of implicit self-esteem measures is their

[14] The measurement-form moderator was not coded by Greenwald et al. (2009), but it is related to a statistically significant moderator that Greenwald et al. labeled "complementarity", defined as a judgment of "the extent to which liking one of the two IAT target categories in a measure implied disliking the other" (p. 21). Studies high in complementary were also ones that were likely to have a criterion measure scored in relative form.

success in confirming theoretical predictions of balanced identity theory, (described in present sections 2–15 and 3–12).

4–10. What are the most important childhood experiences that create the attitudes, stereotypes, identities, and self-esteem revealed by IAT and other indirect measures?

This question lacks other-than-speculative answers. Individual-subject indirect measures of attitudes and stereotypes are not yet available for the ages (presumably 2–4 years) during which these attitudes and stereotypes are likely formed. Until flexibly usable indirect measures are available for these young ages, attempts to identify important formative experiences might use either naturalistic observation of toddlers or retrospective surveys seeking to identify differences in characteristics of childhood environments of persons who as adults differ in IAT-measured attitudes and stereotypes. Such studies have not yet been done.

4–11. What are the long-term trajectories of implicitly measured attitudes and stereotypes?

Three types of multi-year trajectories might be described if the longitudinal studies suggested in 4–6 can be conducted: First, one can examine *age trends between early childhood and early adulthood*. In addition to identifying the age at which indirectly measurable attitudes and stereotypes begin to form, it will be useful to discover the age range in which there is greatest modifiability of implicit measures—this would pinpoint an age range in which the social environment is likely having maximum impact. Second, tracking *age trends between early adulthood and later adulthood* may identify persons who have experienced durable reductions in implicit biases, and this in turn may suggest hypotheses about how to modify implicit biases. Third, *societal time trends* can be tracked using data obtained at the Project Implicit web site (at https://implicit.harvard.edu/implicit). Archiving of data for the many IATs provided by visitors to the Project Implicit site started in December 2002. Those data have been used in published reports since 2003. The archive was made publicly available (at https://osf.io/y9hiq/) in 2014 and was recently used by Charlesworth and Banaji (2019) for analysis of time trends of six IATs between 2007 and 2016. The most substantial change observed by Charlesworth and Banaji was that implicit sexual orientation attitudes (favoring straight relative to gay) became less pro-straight over the 10-year period, confirming a previous observation by Westgate, Riskind, and Nosek (2015).

4–12. What are the effects of possessing implicitly measurable stereotypes tied to one's own identities?

Many stereotypes associate people with traits they have no desire to be associated with. The following half-dozen examples are just the start of an extremely long list: *Asians are shy. Jews are greedy. Cheerleaders lack intelligence. Women are weak. Police use excessive force. White men can't jump.* Stereotypes are considered problematic primarily because they can prompt unjustified judgments about individual outgroup members. Relatively little attention (theoretical or empirical) has been given to problems that follow from applying stereotypes to oneself. The balance–congruity principle of balanced identity theory (BIT; Greenwald et al., 2002) predicts that well-established stereotypes are likely to be self-applied. An example served

as the title of an article by Nosek, Banaji, and Greenwald (2002): "Math = Male, Me = Female, Therefore Math ≠ Me". Generalizing, a stereotype associated with an identity favors development of a self-concept that associates one's self with the stereotyped trait.

There has been little investigation of consequences of self-applied stereotypes. Preceding BIT, Jost, Pelham, and Carvallo (2002) reported findings that high self-esteem (implicitly measured) could favor self-application of positive stereotypes by those identified with an elite academic institution. Considering just the example in Nosek et al.'s (2002) title, one can ask: (a) Does the female ≠ math stereotype function as a self-fulfilling prophecy that impairs women's math performance? (b) Does the female ≠ math stereotype prompt young girls and women not to start on paths that could lead to a career involving math? (c) Is the female ≠ math stereotype a source of anxiety for women when they approach situations that may require math ability? These questions presently lack research-based answers. In a study of Singaporean children in Grades 1, 3, and 5, Cvencek, Kapur, and Meltzoff (2015) found both that the implicit stereotype associating math more with male than female gender increased with grade level and that association of math with self (implicit math self-concept) was positively correlated with math achievement (see also Block, Hall, Schmader, Innes, & Croft, 2018).

4–13. What are the effects of discrepancies between implicitly and explicitly measured attitudes?

The relatively weak correlations that are often observed between parallel implicit and explicit measures have led researchers to investigate the variations, across subjects, in magnitude and direction of difference ('discrepancy') between these parallel measures. Closest attention has been given to these *implicit–explicit discrepancies* in studies of self-esteem. Jordan, Spencer, Zanna, Hoshino-Browne, and Correll (2003) identified two self-esteem categories based on direction and magnitude of difference between implicit self-esteem (ISE) and explicit self-esteem (ESE). *Secure high self-esteem* is indicated by high scores on both ESE and ISE. *Fragile (or defensive) self-esteem* is indicated by positive ESE that is substantially higher than ISE. Multiple publications have adopted this nomenclature (see, e.g., citations in the introduction section of Kernis, Lakey, & Heppner, 2008). As Kernis et al. noted, the evidence remains less than convincing both because of variations in findings among published results and the likelihood that relevant data in either published or unpublished studies have remained unreported. This evidence is not yet collected in a meta-analytic review.

Studies of Black–White racial attitudes also reveal implicit–explicit discrepancies that remain insufficiently understood. On explicit measures of attitude the average of responses provided by African American respondents indicates substantial Black preference, while their IAT scores average near neutral, representing a mixture of respondents showing racial Black preference, racial White preference, and neither. White American subjects show a reverse pattern, with IAT measures revealing strong White preference and explicit measures being much closer to neutrality. The pattern for White respondents of explicit egalitarianism accompanied by subtle, indirect indicators of racial bias was given the label *aversive racism* well before implicit measures existed (Gaertner & Dovidio, 1986). The 'aversive' label reflects Gaertner and Dovidio's understanding that these egalitarian subjects experience interracial interactions

uncomfortably, resulting in their *avoiding* interactions and thereby also avoiding the self-perception that they may harbor racial biases. The IAT's frequent finding of implicit White preference for explicitly egalitarian White respondents has added to prior empirical support for the concept of aversive racism. Banaji and Greenwald (2013) used the empirical observation of discomfort in interracial interactions for egalitarians to suggest that phenomena of aversive racism might also be labeled 'uncomfortable egalitarianism'. This work is having substantial impact in understanding healthcare disparities (e.g., Penner et al., 2010).

Even though implicit–explicit discrepancy has been described and discussed for self-esteem and racial attitude, both theory and method for understanding the effects of these discrepancies remain underdeveloped. There remains wide variety among researchers on preferred practice for identifying and measuring implicit–explicit discrepancies. Improvement of theory about effects of these discrepancies may be needed to motivate further development of measurement procedures.

5. APPLICATION POTENTIAL TO REMEDIATE IMPLICIT BIASES

In recent years, managers and staff of many businesses have been convened in group meetings that include a 1-hour or 2-hour presentation on implicit bias. These presentations often explain that many people possess implicit biases of which they are unaware and that implicit biases can produce unintentionally biased behavior that can disadvantage women, minorities, and others. In many hospitals and medical schools, caregivers and administrators learn that implicit biases can produce disparities that disadvantage racial and ethnic minorities and impoverished persons. In public school systems, teachers and administrators learn at professional meetings that implicit biases can intrude on their judgments, resulting in inequitable administration of suspensions and expulsions, also affecting evaluations of student performances and, thereby, student progress and likelihood of graduation. In institutions of higher learning, administrators and faculty members receive presentations that describe the effects of implicit bias on student admissions, treatment of students in classrooms, and faculty hiring. In meetings convened by their court systems or professional organizations, judges and lawyers learn that implicit biases can adversely affect outcomes to minorities at multiple post-arrest stages of criminal justice—bail setting, arraignment, indictment, trial, and sentencing. In police departments, officers and administrators may receive instruction on how implicit biases can affect decisions made in traffic stops, pedestrian stops, uses of force, and decisions to arrest or not. These educational programs are often characterized as "implicit bias training" or "diversity training".[15]

5–1. Have the methods of diversity or implicit bias training been established as succeeding (a) in providing education about implicit bias, (b) in reducing implicit biases, or (c) in improving organizational diversity statistics?

The concept of implicit bias has achieved broad recognition outside of social psychology, even while theoretical understanding of implicit biases needs further development. The limited theoretical development is noticeable especially in the lack of established explanations for correlations between implicit biases and behavior (described in Parts 3 and 4 of this article). It is therefore understandable that efforts to incorporate scientific understanding of implicit bias into group-administered diversity training remain underdeveloped.

Stated goals of group-administered diversity training are almost always in part educational—to explain what implicit bias is and what consequences it can have. Many offerings of implicit bias training are successful in producing some education. A frequent second goal is therapeutic—reducing audience members' implicit biases or reducing the likelihood that those in the audience will, in the future, unintentionally discriminate against others. For this second goal, there is no reason to expect that diversity 'trainers' (who might more properly be called 'diversity educators') can achieve what researchers cannot produce empirically (see 2–18 and 4–7). A further obstacle to achieving therapeutic goals is diversity educators' typical lack of access to organizational personnel records that could allow them to determine whether their training activity produces advertised consequences. Almost all efforts at diversity education remain

[15] Treatment of applications in this section focuses on diversity training offered to groups in organizational settings. These should be distinguished from procedures investigated in laboratory studies (reviewed by Bezrukova, Spell, Perry, & Jehn, 2016).

unaccompanied by efforts to appraise achievement of either stated training goals or improvement of organizational diversity.

Title VII of the United States's Civil Rights Act of 1964 identified 'protected classes'—which include most prominently race, religion, national origin, age, sex, and disability status—enables persons to sue employers if they receive adverse treatment based on any of these characteristics. The 1964 law authorized creation of the Equal Employment Opportunity Commission (EEOC), which requires employers annually to submit data describing the status of their personnel in the various protected classes. Kalev, Dobbin, and Kelly (2006) and Dobbin and Kalev (2016) used EEOC data to evaluate effects of corporate diversity activities on hiring of women and minorities, from which they concluded that most corporate diversity training is ineffective.

Reviewing a broad mixture of observational, experimental, and field research studies, Paluck and Green (2009) concluded that "a small fraction [of the 985 studies they located] speak convincingly to the questions of whether, why, and under what conditions a given type of [prejudice reduction] intervention works" (p. 339). They concluded "that the causal effects of many widespread prejudice-reduction interventions, such as workplace diversity training and media campaigns, remain unknown" (p. 339). A recent meta-analysis by Bezrukova et al. (2016) concluded that the 260 studies they reviewed had an average training effect size of Hedges g = 0.38 (95% CI=[.33,.42]), which is between conventionally recognized 'small' (= 0.20) and 'moderate' (= 0.50) levels. Although this appears to be a conclusion that diversity training is typically effective, Bezrukova et al. carefully avoided stating that conclusion. They did conclude that "many of the diversity training programs fell short in demonstrating effectiveness on some training characteristics" (p. 1227). They did not offer conclusions about how to construct a successful diversity training program for administration in corporate settings.

To those who seek effective methods of implicit bias reduction, it must be distressing that Lai et al. (2016) concluded that there is not yet any established method to produce durable implicit bias reductions. Even so, Lai et al.'s findings provide no reason to abandon research on possibly effective methods for bias reduction. However, they do provide a reason for practitioners not to claim that they can provide training experiences that will reduce implicit biases. Until evidence for durable implicit bias reduction exists, those who offer diversity education should not claim ability to produce this outcome.

5–2. Can decision makers in possibly discrimination-prone positions (a) detect and interrupt implicit biases as they are operating, or (b) adopt mental strategies to suppress activation of implicit biases?

Computer software that influences physical processes as those processes are occurring are said to operate 'in real time'. In this sense, implicit biases operate in real time, influencing social interactions as those interactions occur. This section's question asks whether remedies intended to counteract implicit biases can act in real time to disrupt implicit bias.

If, while they are making judgments that could produce unintended disparities, decision makers can be aware of signals that implicit bias is operating, they should be able to intercept

implicit bias and avoid its undesired effects. The associative–propositional evaluation (APE) theory of Gawronski and Bodenhausen (2006; see 3–5) explicitly supposes existence of this *conscious override* possibility. Gawronski and Bodenhausen proposed that an attitude's associative representation can produce propositional implications that people may judge to be inconsistent with their endorsed attitudes: "If . . . the propositional implication of an automatic affective reaction is inconsistent with other relevant propositions, it may be considered invalid" (p. 694). In what may be the only empirical result testing whether decision makers can detect their own implicit biases, Hahn et al. (2014; see 3–1) found moderate accuracy of subjects in predicting scores on their own IAT measures. Hahn et al. interpreted their result as "suggest[ing] that people can sense their internal spontaneous reactions"—i.e. , suggesting that decision makers can be aware of their implicit attitudes in real time—i.e., while in the process of making decisions. However, Hahn et al. also mentioned alternative interpretations for accuracy of self-predictions of IAT results, including using one's explicit attitude (see 2–8) or one's knowledge of culturally pervasive biases as the basis for prediction.

Another set of possibilities for using one's own resources to intercept implicit biases follows from an assumption that may be widely held—that undesired biases can be overridden by pausing to think deliberately or by meditating before making decisions that might adversely affect others. To determine whether deliberating or meditating can mitigate implicit biases requires research that has not yet been done. Needed studies would contrast deliberation or meditation with suitable control conditions, then observe expressions of bias.

In their meta-analyses, both Greenwald et al. (2009) and Kurdi et al. (2018) evaluated whether conscious controllability of performances on criterion measures of discriminatory judgment or behavior was associated with reduced correlations between implicit biases and discriminatory judgment or behavior. Kurdi et al. additionally coded subjects' awareness that the criterion behavior involved discrimination. If subjects are (a) aware of having implicit biases and (b) desire not to discriminate, the correlation between IAT bias measures and discriminatory behavior should be reduced when the behavior is controllable, and subjects know that the behavior would express discrimination. Contrary to that expectation, Greenwald et al. found that judged controllability of performance was not a significant moderator of correlation between IAT measures and discriminatory judgment or behavior. Kurdi et al. (2018) found, contrary to the conscious override hypothesis, that both the controllability and awareness moderators were slightly negatively related to correlations of IAT measures with discriminatory judgment and behavior. Evidence supporting the conscious override process is, at present, lacking.

Likening the operation of implicit attitudes and stereotypes to visual illusions, Greenwald and Banaji (2017) proposed that "ordinary social perceivers have no easy way to judge . . . whether their stereotype-influenced perceptions may be invalid" (p. 867). The Oxford English Dictionary defines 'intuition' as "the immediate apprehension of an object by the mind without the intervention of any reasoning process". Visual illusions have intuition's characteristic of immediate apprehension without aid of reasoning. Just as those who experience visual illusions cannot correct their conscious experience to remove the illusion, Greenwald and Banaji proposed that implicit-stereotype-influenced judgments are *social illusions* that cannot immediately be identified as errors, nor can they be corrected by introspective efforts. This view is decidedly

less friendly than is Gawronski and Bodenhausen's (2006) APE model to the possibility of online detection and conscious override of implicit biases.

5–3. Blinding and discretion-elimination (pre-commitment to decision criteria) can prevent implicit biases from causing adverse impacts

An analogy: When disability or illness is due to an infectious agent, treatments of choice are to destroy the agent with an antibiotic or to mobilize the body's natural defenses with immunization. Parallel strategies would be desirable for implicit biases. Unfortunately, no such curative treatment for implicit biases has yet been developed (see 2–18). Infections transmitted by persons (carriers) who are unaware of being contagious provide a more apt medical analogy for how implicit bias can unknowingly produce discrimination. Protection against unknowingly transmitted infections can be deployed either by a potential transmitter or by a potential receiver of the infection. In the medical case, the carrier who is unaware of the infection lacks any reason to deploy a protection strategy. In the implicit bias case, the unaware transmitter is often a person in a superior power position—a person who might reasonably suspect that unintended discrimination is possible. The potential victim's strategies are more defensive than protective—to ignore adverse treatment, to protest after the fact, or to respond with counter-action. None of these is likely to be either effective or constructive in the case of implicit bias. The best opportunities for effective countermeasures are therefore ones that can be deployed by the decision maker.

Because of its simplicity and effectiveness, a preferred decision-maker strategy is *blinding*. A model for this strategy is the orchestral blind audition, in which candidates for instrumental positions perform behind a screen, allowing each performer to be heard but not seen. With blinding, bias based on demographic characteristics is not possible. Adoption of this strategy by major American symphony orchestras in the 1970s led to a substantial increase in their hiring of women instrumentalists (Goldin & Rouse, 1980).

When blinding is not possible, a second strategy, *discretion-elimination*, is available. In U.S. court decisions involving employment discrimination, decision-maker discretion in evaluating job applicants and employees has been identified as a policy that enables discriminatory personnel decisions (e.g., Hart, 2005; Heilman & Haynes, 2008). Discretion can be sharply reduced when decision makers pre-commit to valid decision criteria before they conduct evaluations (Uhlmann & Cohen, 2005). When decisions are conscientiously based on valid criteria that decision makers will not revise as they are deliberating, implicit biases should have less chance of influencing decisions.

A limitation of the discretion-elimination strategy is that validated decision procedures are difficult to construct and therefore are often unavailable. In discussing remedies for flawed discretionary judgments, which they understood as failures of 'intuitive expertise', Kahneman and Klein (2009) described both the values and the challenges of using mechanical decision procedures (algorithms) to replace human judgment: "[T]he conditions necessary for the construction and use of an algorithm . . . include (a) confidence in the adequacy of the list of variables that will be used [and] (b) a reliable and measurable criterion" (p.524). They further noted that "[T]he introduction of algorithms and other formal decision aids in organizations will

often encounter opposition and unexpected problems of implementation. Few people enjoy being replaced by mechanical devices or by mathematical algorithms" (p. 524). Kahneman and Klein's analysis directly suggests the challenges of developing and implementing non-subjective decision procedures ('mechanical devices or algorithms') to replace the use of discretion in judgments such as hiring and evaluation of employees.

A non-mechanical alternative to blinding that is widely recommended by personnel psychologists because of its potential to minimize discretion is the 'structured interview'. Structured interviews provide a fixed list of questions asked to all job applicants. At the (high) quality end, one asks each applicant a set of questions validated for the relevance of their answers to qualification for the position being sought; these should be administered identically (perhaps mechanically) to all applicants, then scored using a scoring scheme known from previous validation research to provide a measure predictive of job performance. At the low-quality end might be an unresearched set of questions, administered by multiple interviewers and having no validated scoring scheme.

5–4. Reasons for tracking effects of actions on those whose outcomes are affected by the actions

Officers and managers in many large organizations (corporations, governments, court systems, police departments, hospitals, and universities) regularly make decisions that affect the organization's employees, job applicants, customers, clients, and charges. Should these decision makers be content not knowing whether their decisions produce unintended adversity to some of those who are affected by their decisions? Employers obviously can act unfairly when they have no way of knowing whether their actions create disparities or unfair disadvantages. This could explain why many decision makers are in no hurry to conduct the types of 'self-critical analysis' (Pollard, 1999) of their employee data that may reveal unintended discrimination. At the same time, their organizations very likely possess personnel data that reveal discrimination. Evidence that is easily available to the employer is rarely open to scrutiny by those who have been disadvantaged. Possible victims must hire a lawyer who, in turn, must persuade a judge that evidence for possible discrimination justifies the court's authorizing 'discovery' of the employer's personnel records.

Two established research findings suggest that most large organizations' personnel data are likely to reveal that their employees who are members of protected classes have suffered adverse impacts that could be due to implicit bias. The two findings are that IAT-measured implicit biases (a) are pervasive (see 2–7) and (b) are consistently correlated with discriminatory intergroup behavior at small to moderate levels (see 2–9 and 2–10). These findings lead to expectation of adverse impacts that can be societally significant as described in 3–3.[16] In the

[16] This potential is indicated by numerous audit studies or analyses of public data that have revealed disparities in hiring by businesses, renting and selling by realtors, lending by banks, medical care by hospitals, and management of justice in prisons, courts, and police departments. Perhaps because of the difficulty of obtaining cooperation to conduct the needed types of studies in businesses, hospitals, and court systems, there has been almost no published empirical research testing hypotheses about causes and consequences of implicit bias in such institutions. For some of the rare exceptions, see the studies of hiring by Rooth and colleagues (Agerström & Rooth, 2011; Rooth, 2010) and of medical care by Penner and colleagues (Hagiwara et al., 2013; Penner et al., 2010.)

United States, large organizations are obliged to file annual EEOC reports on their employment of women and minorities. These reports do not require information that could locate disparities in treatment of employees beyond hiring, nor can they reveal racial or other disparities in service deliveries to clients and customers. For the leader of an organization that has a priority of providing equal opportunity to employees and equal service to customers and clients, the substantial probability of unintended adverse impacts to legally protected classes might justify undertaking 'self-critical analysis' either to discover that there are no pockets of disadvantage or to locate them as a preliminary to fixing them. Such action by leadership requires initiative, perhaps even courage.

AFTERWORD

Using psychological understanding of sensory illusions as a model, Greenwald and Banaji (2017) proposed that stored associative knowledge, accumulated through long experience, controls how new sensory information constructs conscious perceptions that guide judgments and decisions. Lacking a well-formed theory of this process, they offered two metaphors:

> "[Experience-based] associations might be understood as mental pigments that operate in combination to construct rich mental images . . . [A] mass of associative knowledge acts as a *cultural filter* that elaborates perception and judgment, in ways that can vary across persons when cultural environments have constructed the associative mass idiosyncratically" (p. 868).

These metaphors suggest the possibility of a theory to explain how a lifetime's accumulation of stored associative knowledge might produce conscious figments that can guide behavior in often useful, but also unintentionally biased, ways. The next decade of research with the Implicit Association Test (or perhaps with a hoped-for superior successor) should help to develop that theory.

APPENDIX A

"Standard" (7-Block) IAT Procedure

As most frequently used in research, an IAT consists of seven sets (blocks) of trials in which stimuli from four categories are classified. Any IAT is completely specified by the labels to be used for the four categories and the stimulus items (exemplars) used to represent each of the four categories. The subject's task in each of the seven blocks is to provide correct classifications of stimulus items (generally by pressing an assigned left- or right-positioned key on a computer keyboard—for example "E" and "I" (alternately, "D" and "K") on a QWERTY keyboard—into their categories. Typically, two of the categories are called *target* categories. The first reported IAT (Experiment 1 in Greenwald, McGhee, & Schwartz, 1998) used *flower* and *insect* as the labels for its two target categories . The other two categories are *attribute* categories. These were *pleasant* and *unpleasant* (valence) in the Flower–Insect attitude IAT.

The standard order of seven blocks (typical trial numbers [totaling 190] in parentheses), is

1. Classify the items for the two target categories (20)

2. Classify the items for the two attribute categories (20)

3. Classify items for all four categories, one attribute and one target category assigned to each of the two keys, using the assignment of categories to left and right keys as in Blocks 1 and 2 (20).

4. Same as Block 3 (40).

5. Classify the two target categories, reversing the key assignments of Block 1 and having more trials than in Block 1 (30).

6. Classify items for all four categories, using the reversed key assignments of the target categories as in Block 5 (20).

7. Same as Block 6 (40).

The number of trials for reversed 2-category practice in Block 5 can affect the magnitude of effect on the IAT of the order in which the two combined tasks are encountered. After several years of experience, an increase from 20 to 30 trials in Block 5 was adopted as a procedure that often keeps the effect of order of combined tasks to a minimum (see 2–2).

For the four combined-task blocks (3, 4, 6, and 7), which present exemplar items from all four categories, there is a *strict alternation* between presenting an item from one of the two target categories on odd-numbered trials and an item from one of the two attribute categories on even-numbered trials (see 1–B3). Determination of which target category is assigned a left (vs. right) key response in Block 1 and which attribute category is assigned to the left key in Block 2 are typically counterbalanced across subjects. There are typically between four and six items in each of the four categories. The number of trials in a block is often adjusted to allow each of the stimuli to appear equally often. With the same number of exemplars (*n*) for each of the four

categories, this can be done in the 2-category blocks (1, 2, and 5) by having trial counts that are integer multiples of 2*n*, and in the combined-task blocks (3, 4, 6, and 7) trial counts being an integer multiple of 4*n*. With 5 items per category, the numbers might be as shown in the 7-block listing above. With 4 items per category, the numbers of trials in the 7 blocks might by 16, 16, 32, 48, 24, 32, 48. For 6 items per category, these numbers might be 12, 12, 24, 48, 24, 24, 48.

As stated in this section's 1–B6 and 1–B7, however, exactly equating numbers of presentations for target or attribute exemplars should be subordinated to other considerations in determining the trial count for each block. As one example, the numbers for 4 items per category might be set at 16, 16, 24, 40, 24, 24, 40. The number of appearances of each item in combined tasks can then be equated because the sum of trials in each combined-task's pair of blocks is an integer multiple of 4*n* — e.g., for Blocks 3 and 4 the sum is 24+40 = 64 (= 4*4*n*). Other numbers of items per category, especially with different numbers of exemplars in attribute and target categories, might require inappropriately large numbers of trials to maintain equal appearances of each exemplar for target and/or attribute categories. The strict equality need not be treated as essential.

A procedure that records latency to occurrence of the correct response is typically used, with the IAT program recording occurrence of error responses but not registering the trial's latency as completed until the correct response occurs. The value of this method was shown by Greenwald, Nosek, and Banaji (2003).

APPENDIX B

Algorithms for the IAT's *D* Measure

Step	**Built-in error penalty procedure (preferred)** *Each trial's latency is recorded to occurrence of the trial's correct response; trials on which errors preceded the correct responses are included*	**Computed error penalty** *For IAT procedures that end a trial on the first keypress, recording the latency of that keypress and coding the response as correct or error*
1	Designate combined tasks as A (for which faster performance will produce a positive score) and B (for which faster performance will produce a negative score). With counterbalancing, half of subjects will encounter A in Blocks 3&4, half in Blocks 6&7.	same
2	Discard all trials in Blocks 1, 2, and 5	same
3	Identify blocks for combined task as A1 and A2; those for combined task B as B1 and B2. If task A is Blocks 3&4, Block 3 is A1, Block 4 is A2.	same
4	Eliminate from remaining data (Blocks 3, 4, 6, and 7) *only* trials with latencies > 10,000 ms	same
5	Eliminate all subjects for whom *more than* 10% of remaining trials have latencies less than 300 ms	same
6	Compute latency means (MnA1, MnA2, MnB1, MnB2) and SDs (SDA1, SDA2, SDB1, SDB2) for each of the four blocks for all remaining trials	Compute latency means for *correct responses* in each of the four blocks (separately) for remaining trials; also, replace each correct response with a score computed as the *mean of correct responses in the same block as the error, plus a penalty* (see below)
7	Compute two mean latency differences: B1–A1 = (MnB1 – MnA1) and B2–A2 = (MnB2 – MnA2)	Compute the two mean latency differences from all trials, including the error trials that were replaced in Step 6 using error penalties
8	Compute an *inclusive* (not pooled) SD1 using all latencies in Blocks A1 & B1; another (SD2) using all latencies for A2 & B2 (SD2). These can be computed from means and SDs from Step 6 as shown in the lines below this table	Compute the two inclusive SDs using all trials (using the error trials with their replaced latencies)
9	Compute (B1–A1) / SD1; and (B2–A2) / SD2	same
10	*D* = Average of two quotients computed in Step 9	same

SD1 = SQRT((((NA1-1)*SDA1^2+(NB1-1)*SDB1^2)+((NA1+NB1)*((MnA1-MnB1)^2)/4))/(NA1+NB1-1))
SD2 = SQRT((((NA2-1)*SDA2^2+(NB2-1)*SDB2^2)+((NA2+NB2)*((MnA2-MnB2)^2)/4))/(NA2+NB2-1))

In the above two lines, 'N', 'Mn', and 'SD' indicate numbers of trials, means, and standard deviations for the block indicated by the following 2 characters (A1, B1, A2, or B2); the caret (^) precedes an exponent.

Table 2 of Greenwald, Nosek, & Banaji (2003) suggested two options for the error penalty computation. One of these (D_3) used twice the block's SD (i.e., twice SDA1, SDA2, SDB1, or SDB2, depending on the block in which the error occurred). The other option (D_4) used a constant of 600 ms for all blocks. Greenwald et al. also noted the option of deleting responses faster than 400 ms, a procedure that typically affects the resulting measure very little.

REFERENCES

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*, 790–805. http://dx.doi.org/10.1037/a0021594

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*, 299–306. doi: Doi 10.1002/Acp.1691

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

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie*, *48*, 145–160.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*, 668–688.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, *4*, 1–44. doi:10.1111/1529- 1006.01431

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, 17, 53–58.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin*, 142, 1227–1274. http://dx.doi.org/10.1037/bul0000067

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, 242–261.

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

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*, 1468–1481. doi:10.1037/a0038379

Block, K., Hall, W. M., Schmader, T., Inness, M., & Croft, E. (2018). Should I stay or should I go? Women's implicit stereotypic associations predict their commitment and fit in STEM. *Social Psychology*, *49*(4), 243-251. http://dx.doi.org/10.1027/1864-9335/a000343

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. http://dx.doi.org/10.1002/ejsp.487

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, 79, 631–643. http://dx.doi.org/10.1037/0022-3514.79.4.631

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the implicit association test. *Journal of Personality and Social Psychology, 81*, 760–773. doi:10.1037/0022-3514.81.5.760

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

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*, 365–385. http://dx.doi.org/10.1037/a0021341

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (2004). The Implicit Association Test's D measure can minimize a cognitive skill confound: Comment on McFarland and Crouch (2002). *Social Cognition, 22*, 673–684.

Charlesworth, T. E. S., and Banaji, M. R. (2019, in press). Patterns of implicit and explicit attitudes I. Long-term change and stability from 2007–2016. *Psychological Science*.

Cofer, C. N., & Foley, J. P., Jr. (1942). Mediated generalization and the interpretation of verbal behavior: I. Prolegomena. *Psychological Review, 49*, 513–540. http://dx.doi.org/10.1037/h0060856

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The Quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469–487. doi: 10.1037/0022-3514.89.4.469

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cunningham, W.A., Nezlek, J.B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, *30*(10), 1332–1346.

Cvencek, D., & Greenwald, A. G. (in press). Self-esteem, expressions of. In B. J. Carducci (Editor-in-Chief) & A. Di Fabio, D. H. Saklofske, & C. Stough (Vol. Eds.), *The Wiley-Blackwell encyclopedia of personality and individual differences: Vol. III. Personality processes and individual differences*. Hoboken, NJ: John Wiley & Sons.

Cvencek, D., Greenwald, A. G., Brown, A., Snowden, R., & Gray, N. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology, 32*, 302–314.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4-year-olds: The Preschool Implicit Association Test. *Journal of Experimental Child Psychology*, *109*(2), 187–200.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2012). Balanced identity theory: Evidence for implicit consistency in social cognition. In Gawronski, B., & Strack, F. (Eds.), *Cognitive consistency: A unifying concept in social psychology* (pp. 157–177). New York: Guilford Press.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2016). Implicit measures for preschool children confirm self-esteem's role in maintaining a balanced identity. *Journal of Experimental Social Psychology*, *62*, 50–57.

Cvencek, D., Kapur, M., & Meltzoff, A. N. (2015). Math achievement, stereotypes, and math self-concepts among elementary-school children in Singapore. *Learning and Instruction*, *39*, 1–10.doi:10.1016/j.learninstruc.2015.04.002

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. http://dx.doi.org/10.1111/j.1467-8624.2010.01529.x

Cvencek, D. et al. (submitted). *Meta-analytic evaluation of IAT and self-report measures in testing balanced identity theory*. Manuscript submitted for publication. University of Washington.

Dobbin, F., & Kalev, A. (2016, July–August). Why Diversity Programs Fail. *Harvard Business Review*, *94*(7), 52–60.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology*, *47*, 233–279. http://dx.doi.org/10.1016/B978-0-12-407236-7.00005-X

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*, 800–814.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*, 316–328.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009a). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*, 347–368. http://dx.doi.org/10.1037/a0014211

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009b). Theoretical claims necessitate basic research: Reply to Gawronski, Lebel, Peters, and Banse (2009) and Nosek and Greenwald (2009). *Psychological Bulletin*, *135*, 377–379. http://dx.doi.org/10.1037/a0015328

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267–1278. http://dx.doi.org/10.1016/j.jesp.2012.06.003

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*, 835–848. http://dx.doi.org/10.1037/0022-3514.82.5.835

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology*, *54*, 297–327. http://dx.doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C.,& Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). Orlando, FL: Academic Press.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 59–127). San Diego, CA, US: Academic Press. http://dx.doi.org/10.1016/B978-0-12-385522-0.00002-0

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*, 107–119. http://dx.doi.org/10.1037/0003-066X.46.2.107

Glashouwer, K. A., Smulders, F. T., de Jong, P. J., Roefs, A., & Wiers, R. W. (2013). Measuring automatic associations: validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting. *Journal of Behavioral Therapy and Experimental Psychiatry*, *44*, 105–113.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*, 715–741.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72*, 861–871.

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*, 553–561.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25.

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*, 1022–1038.

Greenwald, A. G., & Lai, C. K. (2020[in press]). Implicit social cognition. *Annual Review of Psychology*.

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., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, 69, 669–684.

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

Hagiwara, N., Penner, L. A., Gonzalez, R., Eggly, S., Dovidio, J. F., Gaertner, S. L., …, Albrecht, T. L. (2013). Racial attitudes, social control, and adherence in racially discordant medical interactions. *Social Science and Medicine*, *87*, 123–131.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369-1392. http://dx.doi.org/10.1037/a0035028

Hart, M. (2005). Subjective decisionmaking and unconscious discrimination. *Alabama Law Review*, *56*, 741–791.

Hehman, E., Flake, J.K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*, 393–401.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heilman, M.E., & Haynes, M.C. (2008). Subjectivity in the appraisal process: A facilitator of gender bias in work settings. In E. Borgida & S.T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom*. (pp. 127–156). Oxford: Blackwell Publishing.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Horner, A. J., & Henson, R. N. (2008). Priming, response learning and repetition suppression. *Neuropsychologia*, *46*, 1979–1991. http://dx.doi.org/10.1016/j.neuropsychologia.2008.01.018

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive responding to IAT feedback. *Social Cognition, 35*(5), 520-562. http://dx.doi.org/10.1521/soco.2017.35.5.520

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology, 108*(2), 187-218. http://dx.doi.org/10.1037/a0038557

Jacoby, L. L. (1991). A process-dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, *30*, 513–541.

Jacoby, L. L., Kelley, C. M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*, 326–338.

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*, 10–19. http://dx.doi.org/10.1177/0963721418797309

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25*(6), 881-919. http://dx.doi.org/10.1111/j.1467-9221.2004.00402.x

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology, 38*(6), 586-602. http://dx.doi.org/10.1016/S0022-1031(02)00505-X

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*, 515–526. http://dx.doi.org/10.1037/a0016755

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*, 589–617. http://dx.doi.org/10.1177/000312240607100404

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*, 16–32.

Kernis, M. H., Lakey, C. E., & Heppner, W. L. (2008). Secure versus fragile high self-esteem as a predictor of verbal defensiveness: Converging findings across three different markers. *Journal of Personality*, *76*, 477–512. http://dx.doi.org/10.1111/j.1467-6494.2008.00493.x

Kim, D. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66*, 83–96.

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. *The Quarterly Journal of Experimental Psychology, 63*(3), 595-619. http://dx.doi.org/10.1080/17470210903076826

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93*(3), 353-368. http://dx.doi.org/10.1037/0022-3514.93.3.353

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (in press). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785. http://dx.doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016. http://dx.doi.org/10.1037/xge0000179

Lindgren, K. P., Baldwin, S. A., Olin, C. C., Wiers, R. W., Teachman, B. A., Norris, J., . . . Neighbors, C. (2018). Evaluating within-person change in implicit measures of alcohol associations: Increases in alcohol associations predict increases in drinking risk and vice versa. *Alcohol and Alcoholism*, 53, 386–393. http://dx.doi.org/10.1093/alcalc/agy012

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*(1), 45-69. http://dx.doi.org/10.1037/a0030734

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für Experimentelle Psychologie, 48*(2), 107-122. http://dx.doi.org/10.1026//0949-3946.48.2.107

Mierke, J., & Klauer, K. C. (2003). Method-Specific Variance in the Implicit Association Test. *Journal of Personality and Social Psychology*, *85,* 1180–1192. http://dx.doi.org/10.1037/0022-3514.85.6.1180

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469. http://dx.doi.org/10.1037/0096-3445.132.3.455

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565-584. http://dx.doi.org/10.1037/0096-3445.134.4.565 Nosek, B. A., & Greenwald, A. G. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). Psychological Bulletin, *135*, 373–376. http://dx.doi.org/10.1037/a0015047

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 (Pp. 265–292). In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior*. Psychology Press.

Nosek, B.A., & Hansen, J.J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*, 553−594.

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

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B., Wiers, R. W., Somogyi, M., Akrami, N., Ekehammar, B., Vianello, M., Banaji, M. R., & Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*, 10593–10597.

Olson, Michael A.; Fazio, Russell H. (May, 2000). Responses to the Implicit Association Test by individuals motivated to control prejudiced reactions. Midwestern Psychological Association (MPA). http://dx.doi.org/10.1037/e413792005-136

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*(5), 413-417. http://dx.doi.org/10.1111/1467-9280.00376 Orchard, J., & Price, J. (2017). County-level racial prejudice and the black-white gap in infant health outcomes. *Social Science & Medicine*, *181*, 191–198. http://dx.doi.org/10.1016/j.socscimed.2017.03.036

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192. http://dx.doi.org/10.1037/a0032734

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (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. http://dx.doi.org/10.1037/pspa0000023

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition*, *19*, 97–144. http://dx.doi.org/10.1521/soco.19.2.97.20706

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*, 241–263. http://dx.doi.org/10.1037/h0027272

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367. http://dx.doi.org/10.1146/annurev.psych.60.110707.163607

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277-293.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233-248. http://dx.doi.org/10.1080/1047840X.2017.1335568

Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., & Markova, T. (2010). Aversive racism and medical interactions with Black patients: A field study. *Journal of Experimental Social Psychology*, *46*, 436-440. http://dx.doi.org/10.1016/j.jesp.2009.11.004

Phelan, J. E., & Rudman, L. A. (2008). *Negations are Not Good for the IAT*. Unpublished manuscript, Rutgers University.

Pinter, B., & Greenwald, A. G. (2005). Clarifying the role of the "other" category in the self-esteem IAT. *Experimental Psychology*, *52*, 74–79.

Pollard, D. A. (1999). Unconscious bias and self-critical analysis: The case for a qualified evidentiary equal employment opportunity privilege. *Washington Law Review*, *74*, 913–1031.

Qian, K. M., Heyman, G. D., Quinn, P. C., Messi, F. A., Fu, G., & Lee, K. (2016). Implicit racial biases in preschool children and adults from Asia and Africa. *Child Development*, *87*, 285–296. https://doi.org/10.1111/cdev.12442

Rae, J. R., Newheiser, A.-K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science*, *6*, 535–543. http://dx.doi.org/10.1177/1948550614567357

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, *19*, 249–254. http://dx.doi.org/10.1111/j.1467-9280.2008.02076.x

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review*, *111*(1), 159-182. http://dx.doi.org/10.1037/0033-295X.111.1.159

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563-575. http://dx.doi.org/10.3758/BF03207490

Richardson-Klavehn A, Bjork RA. (1988). Measures of memory. *Annual Review of Psychology*, *39*, 475–543.

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling implicit association test data? test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PLoS ONE* 10: e0129601. https://doi.org/10.1371/journal.pone.0129601

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real-world evidence. *Labour Economics*, *17*, 523–534. doi:10.1016/j.labeco.2009.04.005

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.

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). *The Quarterly Journal of Experimental Psychology*, *62*, 84–98. doi: 10.1080/17470210701822975

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift für Experimentelle Psychologie*, *48*, 94–106. doi: 10.1026/0949- 3946.48.2.94

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*, 139–165. doi: 10.1037/0096-3445.133.2.139

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. doi: 10.1037/0096-3445.134.3.426

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin*, *27*, 1164–1178.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, *17*, 437–465.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772-780. http://dx.doi.org/10.1111/j.1467-9280.2008.02156.x

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115*(2), 314-335. http://dx.doi.org/10.1037/0033-295X.115.2.314

Smith, R. (2014). Blood pressure averaging methodology: Decreasing the rate of misdiagnosing hypertension. [Available online.]

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4,* 108–131.

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology, 56*(4), 283-294. http://dx.doi.org/10.1027/1618-3169.56.4.283

Sriram, N., Nosek, B. A., & Greenwald, A. G. (2007). Scale Invariant Contrasts of Response Latency Distributions. https://doi.org/10.31234/osf.io/erjh7

Stanovich, K. E., West, R. F., & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 80-91). New York, NY, US: Guilford Press.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie*, *48*, 123–134. http://dx.doi.org/10.1026//0949-3946.48.2.123

Stergiou, G. S., et al. (2002). Reproducibility of home, ambulatory, and clinic blood pressure: Implications for the design of trials for the assessment of antihypertensive drug efficacy. *American Journal of Hypertension, 15,* 101–104.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220–247.

Strack, F., & Deutsch, R. (2012). A theory of impulse and reflection. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 97–117). Thousand Oaks, CA, : Sage Publications Ltd. http://dx.doi.org/10.4135/9781446249215.n6

Swanson, J. E., Rudman, L. A., & Greenwald , A. G. (2001). Using the Implicit Association Test to investigate attitude-behavior consistency for stigmatized behavior. *Cognition and Emotion, 15,* 207-230.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*, 474–480.

Verschuere, B., & Kleinberg, B. (2017). Assessing autobiographical memory: the web-based autobiographical Implicit Association Test. *Memory*, *25*, 520–530.

von Helmholtz, H. (1925). *Handbook of physiological optics* (Vol. 3; J. P. C. Southall, Trans.). New York, NY: Optical Society of America. (Original work published 1867).

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*, 273–295.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, *11*(3), e0152719–22. http://doi.org/10.1371/journal.pone.0152719

Westgate, E. C., Riskind, R. G., & Nosek, B. A. (2015) Implicit preferences for straight people over lesbian women and gay men weakened from 2006 to 2013. *Collabra*, *1*(1): 1, pp. 1–10. http://dx.doi.org/10.1525/collabra.18

Williams, R. H., & Zimmermann, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262-274. http://dx.doi.org/10.1037/0022-3514.72.2.262

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*, 1193–1209.

Wrzus, C., Egloff, B., & Riediger, M. (2017). Using implicit association tests in age-heterogeneous samples: The importance of cognitive abilities and quad model processes. *Psychology and Aging*, *32*(5), 432-446. http://dx.doi.org/10.1037/pag0000176

Xu, K., Nosek, B., & Greenwald, A.G. (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data* 2(1):e3, DOI: http://dx.doi.org/10.5334/jopd.ac