Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm

Anthony G. Greenwald
University of Washington

Brian A. Nosek
University of Virginia

Mahzarin R. Banaji
Harvard University

In reporting Implicit Association Test (IAT) results, researchers have most often used scoring conventions described in the first publication of the IAT (A. G. Greenwald, D. E. McGhee, & J. L. K. Schwartz, 1998). Demonstration IATs available on the Internet have produced large data sets that were used in the current article to evaluate alternative scoring procedures. Candidate new algorithms were examined in terms of their (a) correlations with parallel self-report measures, (b) resistance to an artifact associated with speed of responding, (c) internal consistency, (d) sensitivity to known influences on IAT measures, and (e) resistance to known procedural influences. The best-performing measure incorporates data from the IAT's practice trials, uses a metric that is calibrated by each respondent's latency variability, and includes a latency penalty for errors. This new algorithm strongly outperforms the earlier (conventional) procedure.

The Implicit Association Test (IAT) provides a measure of strengths of automatic associations. This measure is computed from performance speeds at two classification tasks in which association strengths influence performance. The apparent usefulness of the IAT may be due to its combination of apparent resistance to self-presentation artifact (Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim & Greenwald, 1998), its lack of dependence on introspective access to the association strengths being measured (Greenwald et al., 2002), and its ease of adaptation to assess a broad variety of socially significant associations (see overview in Greenwald & Nosek, 2001).

The IAT's measure, often referred to as the *IAT effect*, is based on latencies for two tasks that differ in instructions for using two response keys to classify four categories of stimuli. Table 1 describes the seven steps (*blocks*) of a typical IAT procedure.

The first IAT publication (Greenwald, McGhee, & Schwartz, 1998) introduced a scoring procedure that has been used in the majority of subsequently published studies. The features of this *conventional algorithm* (see Table 4 later in the article) include (a) dropping the first two trials of test trial blocks for the IAT's two classification tasks (Blocks 4 and 7 in Table 1), (b) recoding latencies outside of lower (300 ms) and upper (3,000 ms) boundaries to those boundary values, (c) log-transforming latencies before averaging them, (d) including error-trial latencies in the analyzed data, and (e) not using data from respondents for whom average latencies or error rates appear to be unusually high for the sample being investigated. The main justification for originally using these conventional procedures was that, compared with several alternative procedures often used with latency data, the conventional procedures typically yielded the largest statistical effect sizes.

Previous theoretical and methodological analyses have provided methods of dealing with problems that occur in latency measures

Anthony G. Greenwald, Department of Psychology, University of Washington; Brian A. Nosek, Department of Psychology, University of Virginia; Mahzarin R. Banaji, Department of Psychology, Harvard University.

The revised scoring procedures described in this report are hereby made freely available for use in research investigations. SPSS syntax for computing Implicit Association Test measures using the improved algorithm can be obtained at the University of Washington Web site (http://faculty.washington.edu/agg/iat_materials.htm). However, the improved scoring procedures described in this report (patent pending) should not be used for commercial applications nor should they or the contents of this report be distributed for commercial purposes without written permission of the authors.

This research was supported by three grants from National Institute of Mental Health: MH-41328, MH-01533, and MH-57672. The authors are grateful to Mary Lee Hummert, Kristin Lane, and Deborah S. Mellott for helpful comments on an earlier version, and also to Laurie A. Rudman and Eliot R. Smith, who commented as colleagues rather than as consulting editors for this journal.

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, Washington 98195-1525. E-mail: agg@u.washington.edu

Journal of Personality and Social Psychology, 2003, Vol. 85, No. 2, 197–216
 0022-3514/03/$12.00 DOI: 10.1037/0022-3514.85.2.197

Table 1
Sequence of Trial Blocks in the Standard Election 2000 (Bush vs. Gore) IAT

Block	No. of trials	Function	Items assigned to left-key response	Items assigned to right-key response
1	20	Practice	George Bush images	Al Gore images
2	20	Practice	Pleasant words	Unpleasant words
3	20	Practice	Pleasant words + Bush items	Unpleasant words + Gore items
4	40	Test	Pleasant words + Bush items	Unpleasant words + Gore items
5	20	Practice	Al Gore images	George Bush images
6	20	Practice	Pleasant words + Gore images	Unpleasant words + Bush images
7	40	Test	Pleasant words + Gore images	Unpleasant words + Bush images

Note. For half the subjects, the positions of Blocks 1, 3, and 4 are switched with those of Blocks 5, 6, and 7, respectively. The procedure in Blocks 3, 4, 6, and 7 is to alternate trials that present either a pleasant or an unpleasant word with trials that presented either a Bush or Gore image. The procedure used for the Election 2000 IAT reported in this article differed from this standard procedure by including 40 practice trials in Block 6. The procedure for the race IAT reported in this article differed from the standard procedure by using 40 practice trials in Block 5. These strategies were used successfully to reduce the typical effect of order in which the two combined tasks are performed. IAT = Implicit Association Test.

in the form of speed–accuracy tradeoffs (e.g., Wickelgren, 1977; Yellott, 1971), age-related slowing (e.g., Faust, Balota, Spieler, & Ferraro, 1999; Ratcliff, Spieler, & McKoon, 2000), and spurious responses that appear as extreme values (or *outliers*; Miller, 1994; Ratcliff, 1993). Remarkably, research practice in cognitive and social psychology has been no more than mildly influenced by this methodological work. That limited influence may be explained by three practical considerations: First, some of the methodological recommendations are costly to use—for example, several hours of data collection with each subject may be needed to obtain data sets from which individual-subject speed–accuracy tradeoff functions can be constructed. Second, journal editors and reviewers rarely insist on the more painstaking methods. Third, researchers who use the more sophisticated (and painstaking) methods are rarely rewarded for their extra work—conclusions based on the more effortful methods often diverge little from those based on simpler methods.

The conventional scoring procedure for the IAT has not previously been subject to systematic investigations of psychometric properties. Additionally, the conventional scoring procedure lacks any theoretical rationale that distinguishes it from other scoring methods (Greenwald, 2001). Consequently, the authors welcomed a fortuitous opportunity to compare the conventional procedure with alternatives. This opportunity arose through the operation of an educational Web site (http://www.yale.edu/implicit/) at which several IAT procedures had been made available for demonstration use by drop-in visitors.

This article first describes the IAT Web site and then presents a series of studies that were designed to evaluate candidate alternative scoring procedures for IATs that operated on the Web site. The investigated scoring methods included (a) transformations of latency measures, (b) procedures for dealing with extreme (slow and fast) responses, (c) replacement (penalty) schemes for error trials, and (d) criteria for identifying a respondent's data as unfit for computing IAT measures. The article concludes by recommending a replacement for the conventional IAT-scoring algorithm.

General Method

The Yale IAT Web Site

The Yale IAT Web site was intended to function as the Internet equivalent of an interactive exhibit at a science museum. The site was designed to allow Web visitors to experience what the authors and many laboratory subjects have experienced: inability to control the manifestations of automatic associations that are elicited by the IAT method. Drop-in visitors could take demonstration versions of IATs that had been in laboratory use for 2–4 years. Within 5–10 min, a visitor to the Web site could complete a measure of implicit attitude or stereotype, after optionally responding to some items that requested demographic information and explicit (self-report) measures of the target attitude or stereotype.[1]

Unlike laboratory IATs, the Web site IATs provided respondents with a summary interpretation of their test performance by characterizing it as showing "strong," "medium," "slight," or "little or no" association of the type measured by each test. Respondents could also inspect distributions of summary results for large numbers of previous respondents. Amplifying the usual debriefing procedure of an experiment, the Web site also provided answers to numerous questions concerning the IAT's methods and interpretations, including a discussion of the distinction between the implicit prejudice that the IAT sometimes measures and the more ordinary meaning of (explicit) prejudice.[2] Approximately 1.2 million tests were completed at the Yale IAT Web site between October 1998 and May 2002, when the present analyses were begun.

[1] The rationale for interpreting the IAT's association strength measures as indicators of social cognitive constructs such as implicit attitude or implicit stereotype rests on theoretical definition of those constructs in terms of concept–attribute associations. This theoretical conception has been described by Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellott (2002).

[2] This distinction is described, on a Web page of answers to frequently asked questions for an IAT designed to measure implicit race attitudes, as follows: "Social psychologists use the word 'prejudiced' to describe people who endorse or approve of negative attitudes and discriminatory behavior toward various out-groups. Many people who show automatic White preference on the Black–White IAT are not prejudiced by this definition.

Respondents

Recruitment. Recruitment occurred via media coverage, links from other sites, links provided by search engines, and word of mouth. Media coverage may have been the most significant influence on response rate. For example, over 150,000 visits to the Yale IAT site were recorded in the 5 days following televised programs that described the IAT on the National Broadcasting Company (NBC) television program, *Dateline* (March 19, 2000) and on a Discovery Channel program titled *How Biased Are You?* (March 20, 2000). The data analyzed in this report were provided by respondents in a 9-month period between July 2000 and March 2001.

Characteristics of respondents. The IAT Web site included a prominent assurance that anonymity of visitors would be protected. Because of this anonymity, the Web site data provided no opportunity to track characteristics of respondents beyond their optional responses to some self-report questions that appeared on the site. Approximately 90% of respondents did, however, respond to some or all of the demographic questions. Of these respondents, 61% were female and 39% were male; 60% were below 24 years of age, 36% were between 24 and 50, and 4.6% were over 50; 0.7% were Native American, 6.4% were Asian, 5.0% were Black, 3.8% were Hispanic, 76.0% were White, 1.0% were biracial (Black–White), 3.3% were multiracial, and 4.0% reported "other" for ethnicity; 18% reported having a high school diploma or less education, 47% had some college experience, 21% had a bachelor's degree, and 14% had a postbaccalaureate degree; 80% of the respondents reported being from the United States and, of the 20% non-U.S. respondents, about half came from Canada, Australia, or Britain (evenly distributed), and the remainder from other countries.

Procedure

Materials and apparatus. Web site IATs were presented using Java Applet and Common Gateway Interface (CGI) technology. After it was downloaded via the respondent's browser, the program used the respondent's computer to present stimuli and to measure response latencies. The respondent's browser program returned the respondent's data to the Web server. The server then analyzed the data and reported a test result within several seconds. Test results were reported as showing "strong," "medium," "slight," or "little or no" strength of one of the association contrasts measured by the test.[3] For example, for the Race IAT, the results indicated the strength of respondents' automatic preferences for Black relative to White race—that is, differential association of Black and White with pleasant. Precision in measuring individual latencies was limited by the clock rate of the operating system that supported the respondent's Web browser (e.g., 18.2 Hz for Windows systems). This was not a debilitating limitation because of the nonsystematic nature of the resulting noise and the substantial reduction of its magnitude produced by averaging data over approximately 40 trials.

Self-report measures and demographic data. Before each IAT, respondents received an optional survey page that included items to measure explicit attitudes or beliefs regarding the IAT's target categories along with some demographic items. Respondents were informed that the self-report and demographic items were optional—respondents could proceed to the IAT demonstration without responding to the items.

IAT measures. Nine IAT measures were available at the Yale IAT site at various times starting in late September 1998: implicit race attitude, using either (a) African American and European American first names or (b) morphed racially classifiable faces and the attributes of good and bad; implicit age attitude, using either (c) first names or (d) morphed age-classifiable faces and the good–bad attribute contrast; (e) implicit gender–career stereotype, measuring association of female and male with career and family; (f) implicit gender–science stereotype, measuring association of female and male with science and liberal arts; (g) implicit self-esteem, measuring associations of self and other with good and bad; (h) implicit math–arts attitude, measuring associations of math and arts with good and bad; and (i) Election 2000 implicit candidate preference, measuring associations contrasting pairs of major candidates in the U.S. presidential primaries of 2000 with good and bad. More detailed descriptions of these IATs are available in Nosek, Banaji, and Greenwald (2002a). Four of the nine IATs (b, d, f, and i in the preceding list) provided the data for the present analyses.

Sequence of tasks. Respondents first saw preliminary information that described what they might experience in taking an IAT. They were then offered the opportunity to continue if they wished to do so. Those who continued then chose one IAT from a list of four to six that were currently available on the Web site. Third, respondents optionally reported their attitudes or beliefs in response to one or more self-report items that were worded to capture the comparison of concepts (e.g., preference for young vs. old) used in the upcoming IAT measure. Fourth, respondents optionally responded to demographic items. Fifth, respondents read instructions for the Web-administered IAT and proceeded to complete it. Completion of an IAT typically required 5–10 min. Preliminary information advised respondents (a) about possible discomforts that might be produced by the test's speed stress and its use of visual stimuli, (b) that the reported results of the test were not guaranteed to be valid, and (c) that there was no obligation to complete the IAT after starting it.

Limitations of the Web Site Data

Self-selection. The respondent samples for this research cannot be treated as representative of any definable population. At the same time, the sample was considerably more diverse than typical research samples (see *Characteristics of respondents*). An important feature of the samples was their large size, which afforded the statistical power to discriminate small, but possibly consequential, differences in properties of alternative scoring procedures.

Possible multiple participations by respondents. Because participation at the IAT Web site was anonymous, Web site visitors could complete as many IATs as they wished and could take the same IAT multiple times. Multiple data points from single respondents pose obvious problems for statistical analysis. However, the overall large number of respondents reduces the potential impact of this problem: Few, if any, single respondents could plausibly have provided as much as 0.1% (e.g., 10 in 10,000 observations) of any of the data sets. For additional discussion of multiple data points from single respondents see Nosek et al. (2002a). One of the preliminary optional questions given to respondents asked how many IATs they had previously completed. That measure was available to assess the effect of prior participation.

Criteria for Evaluating Candidate IAT Measures

Each of the following criteria for evaluating IAT measures was used in one or more of the present series of studies. The first two criteria, IAT correlations with explicit measures (high correlations desired) and correlation with average latency (low correlations desired) are the most important ones of the following six criteria.

IAT correlations with explicit measures. Three self-report items were available for comparison with each IAT. One was a Likert-type measure that requested a comparative appraisal of the two opposed target concepts (e.g., young vs. old for the Age IAT) on the IAT's attribute dimension

These people are apparently able to function in non-prejudiced fashion partly by making active efforts to prevent their automatic White preference from producing discriminatory behavior" (https://implicit.harvard.edu/implicit/demo/racefaqs.html).

[3] The *slight*, *medium*, and *strong* labels corresponded to results meeting the conventional criteria for small, medium, and large effect sizes of Cohen's (1977) *d* measure.

(positive vs. negative valence for three of the IATs; science vs. arts for the fourth). The second and third self-report items were in thermometer format, requesting separate judgments for the IAT's two target concepts on an 11-point scale for the IAT's attribute dimension. (The thermometer scales had just 5 points for the Gender–Science IAT. See the Appendix for wordings of all explicit-measure items.)

By subtraction, the two thermometer items were combined into a thermometer difference score. The Likert measure and the thermometer difference measures were then combined into an overall explicit measure by standardizing each and averaging the two resulting scores.

Correlations of the overall explicit measure with the various IAT measures were computed. Although values for implicit–explicit correlations varied widely for the four data sets, all were positive, consistent with previous observations (Nosek et al., 2002a). Using the conventional algorithm for scoring the IAT, implicit–explicit correlations were .11, .20, .29, and .69, respectively, for the Age, Gender–Science, Race, and Election 2000 IATs. Variations among these correlations are assumed to result from variations in the extent to which IAT and self-report share in measuring the associations that the IAT is intended to measure (e.g., for the Age IAT, associations of young or old with pleasant or unpleasant).

A central assumption for analyses in this article is that higher implicit–explicit correlations for a modified IAT measure can indicate greater construct validity of the modified measure as a measure of association strengths. This central assumption depends on a further assumption that association strength is a latent component of both the implicit and explicit measures. The importance of the shared–latent component assumption can be illustrated by analogy to the way in which a superior measure of height should increase the correlation between height and weight. In the case of height and weight, the shared latent component is height, in the sense that weight can be understood as having contributions due to height, girth, and density. In this circumstance, an improved height measure (e.g., a ruler that can be read to the nearest half inch rather than to the nearest half foot) should yield higher correlations with weight.

Just as for implicit–explicit correlations, the correlation between height and weight can vary considerably for different samples. For example, height and weight may be correlated almost perfectly when other determinants of weight (girth and density) are either kept constant or are correlated with height, as might be the case for a sample of newborn infants. By contrast, for a sample of American professional football players, the height–weight correlation may be much lower because heights may vary little and girths may vary considerably. Nevertheless, in either sample (newborns or football players), the height–weight correlation should be larger for a more sensitive measure of height. The interpretation of implicit–explicit correlations as indicators of construct validity of IAT measures is considered further in the General Discussion.

Correlations of IAT with response latency. Research on cognitive aging has established that effects of experimental treatments on response latency are generally larger for elderly than for young subjects. This age difference is known to be associated with greater average latency for elderly subjects (*age-related slowing*; e.g., Brinley, 1965; Faust et al., 1999; Ratcliff et al., 2000). Consequently, it is expected that IAT effects will be artifactually larger for any subjects who respond slowly, not just the elderly. This artifact should take the form of a positive correlation of extremity of IAT effects with response latency.[4] It is desirable for an IAT measure to minimize this undesired artifactual correlation with response speed.

Internal consistency. For each candidate scoring algorithm, two part-measures were created by applying the scoring algorithm separately to two mutually exclusive subsets of the IAT's combined-task trials. The correlation between these two part-measures, across respondents, provided a measure of internal consistency.

Sensitivity to known influences. Three of the IATs included in this research were known to be sensitive to implicit attitudes and stereotypes that are pervasive in (at least) American society. The Age IAT typically indicates strong implicit preference[5] for young relative to old, and the Gender–Science IAT typically indicates strong male–science and female–arts associations. For the Race (Black–White) IAT, the typical pattern is implicit preference for White relative to Black. Sensitivity to these known modal response tendencies was used as an indicator of performance for the alternative scoring algorithms. Use of this criterion is based on the assumption that the modal response tendencies reflect population differences in association strengths. That assumption is consistent with much research evidence (e.g., Asendorpf, Banse, & Mücke, 2002; Ashburn-Nardo, Voils, & Monteith, 2001; Egloff & Schmukle, 2002; Gawronski, 2002; Greenwald et al., 2002; Greenwald & Nosek, 2001; McConnell & Leibold, 2001; Nosek, Banaji, & Greenwald, 2002b; Rudman, Feinberg, & Fairchild, 2002), although some alternative interpretations have been suggested (e.g., Brendl, Markman, & Messner, 2001; Rothermund & Wentura, 2001).

Resistance to undesired influence of order of combined tasks. Analyses of Web site IAT data by Nosek et al. (2002a) confirmed, in Web site IATs, a finding originally reported by Greenwald et al. (1998): IAT measures tend to indicate that associations have greater strength when they are tested in the first combined task (see Table 1, Blocks 3 and 4) than in the second combined task (Blocks 6 and 7). On the assumption that association strengths are not altered by the order of combined tasks, an IAT measure that minimizes this procedural effect is desirable.

Resistance to effect of prior experience taking an IAT. Analysis of Web site IATs by Nosek et al. (2002a) indicated that IAT measures are reduced in extremity for respondents who have prior experience taking one or more IATs. On the assumption that taking the IAT does not alter the association strengths being measured, an IAT measure that minimizes this procedural effect is desirable.

Candidate Measures

The IAT measure has conventionally been computed as the difference between central tendency measures obtained from its two test blocks, which are Blocks 4 and 7 in Table 1. The present research started by selecting five candidate methods of computing this difference.

Median. The median of each test block was used as the block's summary measure. The difference between the two medians provided the IAT measure. The median is used relatively infrequently with latency dependent measures. It was included here mainly because of curiosity about its performance in comparison with other measures.

Mean. The arithmetic mean latency was computed for each test block. The resulting IAT measure was the difference between the two means. This measure is typically used for graphic or tabular presentation of results in IAT research, but has been inferior to the conventional (log) measure in statistical tests.

Log. The measure for each test block was the mean of natural logarithm transformations of individual-trial latencies. The IAT measure was the difference between these means. This is the transformation that has been conventionally used in statistical tests of IAT measures (e.g., analyses of variance, correlations, regressions, and effect size computations). The rationale for the log transformation is provided by the typically extended upper tails of latency distributions. The log transformation improves the symmetry of latency distributions by shrinking the upper tail and is thereby expected to improve central tendency estimates.

Reciprocal. The measure for each block is the mean of reciprocal latencies (computed as 1,000 ÷ latency). The IAT measure is the differ-

[4] This article provides clear evidence for the existence of this artifact (see Figure 2).

[5] The IAT measures relative strengths of associations. "Implicit preference" is a shorthand for stronger association of one of the two target concepts with positive valence, and/or weaker association of that concept with negative valence.

ence between these means. Like the log transform, the reciprocal improves the symmetry of distributions with extended upper tails. To keep directionality of measures the same for all IAT measures, the difference score for the reciprocal measure was reversed by subtracting it from zero.

D. This measure divides the difference between test block means by the standard deviation of all the latencies in the two test blocks. Part of the rationale for this adjustment is that magnitudes of differences between experimental treatment means are often correlated with variability of the data from which the means are computed. Using the standard deviation as a divisor adjusts differences between means for this effect of underlying variability. A related adjustment has been recommended for use in cognitive aging studies, in which treatment effects on latencies are often greater for elderly subjects, who show both higher means and greater variability of latencies than young subjects. (For discussions of the variability problem in cognitive aging studies see, e.g., Brinley, 1965; Faust et al., 1999; Ratcliff et al., 2000). A successful exploratory attempt to use this type of individual-variability calibrated measure was recently reported by Hummert, Garstka, O'Brien, Greenwald, and Mellott (2002).

Division of a difference between means by a standard deviation is quite similar to the well-known effect-size measure, *d* (Cohen, 1977). The difference between the present *D* measure and the *d* measure of effect size is that the standard deviation in the denominator of *D* is computed from the scores in both conditions, ignoring the condition membership of each score. By contrast, the standard deviation used in computing the effect size *d* is a pooled within-treatment standard deviation. To acknowledge both this measure's similarity to *d* and its difference, the present measure is identified with an italicized uppercase letter (*D*) rather than an italicized lowercase letter.[6]

Analysis and Reporting Strategy

The present series of studies examined alternative policies for retaining trials, including practice trials and error trials, in the data set (Study 1); alternative data transformations (Study 2); use of criteria based on speed or accuracy of responding as the basis for discarding respondents from the data set (Study 3); applying time penalties for the occurrence of errors (Study 4); and deleting extreme (fast or slow) latencies or recoding them to upper and lower boundary values (Study 5).

To keep the task of exploring alternative scoring procedures manageable, Studies 1–5 focused on the two most important performance criteria: magnitude of implicit–explicit correlation of IAT scores with self-report and resistance to covariation of the IAT measure with latency differences among respondents. Study 6 examined combinations of the best-performing procedures identified in Studies 1–5 and used the full set of performance criteria that were available to compare alternative scoring algorithms.

The series of six studies, conducted in parallel for four large data sets, generated many more analyses than can be described in this article. For Studies 1–5, results are presented in some detail for the data set that had largest values of implicit–explicit correlations (Election 2000 candidate preference).[7] Results of Studies 1–5 for the other three data sets (Age, Race, and Gender–Science) are mentioned in passing when they shed additional light. Results from all four data sets are presented for Study 6.

Study 1: Usefulness of Practice Trials and Error Trials

The conventional IAT algorithm discards the first two trials of each test block (Blocks 4 and 7 in Table 1) because of their typically lengthened latencies. Additionally, the conventional algorithm treats as practice (and excludes from measure computations) the two combined-task blocks that precede the two test blocks (Blocks 3 and 6 in Table 1). The conventional algorithm also differs from many other analyses of latency data by retaining latencies from trials on which errors occurred. Study 1 examined these exclusions and inclusions to determine whether they could be justified in terms of their impact on performance of the IAT measure.

Method

Data set. All four data sets were analyzed. However, only the results for the Election 2000 data set are described here in detail. Respondents could choose any two of the actively competing candidates for the nominations of the Republican and Democrat parties. (The most prominent candidates were George W. Bush, Al Gore, John McCain, and Bill Bradley.) Analyses were limited to the pair that was most often selected, George W. Bush and Al Gore.

Respondents. The U.S. Presidential Election took place on November 7, 2000. The analyzed data were obtained between October 3, 2000 and March 20, 2001. Of 11,956 who chose to contrast Bush and Gore in the IAT, slightly over a quarter (26.7%) took the IAT on or before Election Day. Another 31.1% took the IAT on or before December 13, the day on which the election officially concluded with the victory of George W. Bush. Complete IAT data were available for 8,891 respondents (3,065 did not complete the IAT). Of these, complete self-report data (one Likert item and two thermometer items) were available for 8,218 (92.4% of those who completed the IAT).

Preliminary exclusions of very long latencies. The data set contained occasional extremely long latencies—some in excess of 10^6 ms, which is more than a quarter of an hour. These extravagant latencies could have been produced when respondents temporarily abandoned the IAT in favor of some other activity. Such extreme values are not generally tolerated in analyses of latency data. Had they been retained in the present data sets, they would have impaired some of the candidate measures much more than others. At the same time, it seemed desirable to keep initial cleansing to a minimum. Somewhat arbitrarily, then, latencies above 10,000 ms were excluded before any further computations.

IAT measure computations. Each of the five measures (median, mean, log, reciprocal, and *D*) involved computing, first, a central tendency measure for each of the two combined tasks and, second, a difference between these central tendency measures. All IAT measures were computed such that higher numbers indicated implicit preference for George W. Bush relative to Al Gore. The different measures were compared in terms of correlations of IAT measures both with self-report (i.e., explicit) measures and with respondent average latencies. Respondents were classified as self-reported Bush or Gore supporters on the basis of their responses to the 5-point Likert item that assessed relative preference for Bush and Gore. Before computing correlations with average latency, IAT measures for self-reported Gore supporters were reversed (subtracted from zero) so that the expected correlation of IAT scores with average latencies would be positive. The correlations with average latency were computed using the data only for respondents whose self-described support for either candidate was strong. The sample contained 5,202 self-characterized strong supporters, of whom 3,373 (64.8%) favored Gore.

---

[6] The authors conducted numerous analyses to compare the *D* and *d* transformations as IAT effect measures. The *D* transformation was observed consistently to be superior and, accordingly, only results for *D* are presented in this report.

[7] Part of the reason for focusing on this data set is as a useful contrast to the low implicit–explicit correlations that have been reported in most previous publications concerning the IAT. Although such low correlations are typical for attitudes and stereotypes involving stigmatized groups, there are important domains for which correlations are higher—not only attitudes toward political candidates, but also attitudes toward academic subjects (Nosek et al., 2002b) and consumer attitudes (Maison, Greenwald, & Bruin, 2001).

Results and Discussion

First two trials of combined-task blocks. The first analysis examined effects of the conventional algorithm's preliminary discard of the first two trials of combined-task blocks (Blocks 4 and 7 in Table 1). This practice was originally based on the observation that the first two trials' latencies were, on average, substantially slower than the remainder of trials in the same blocks. However, the slowness of these latencies does not necessarily mean that their inclusion will contaminate measures. To determine the usefulness of data from the first two trials, two data sets were prepared that differed in inclusion versus exclusion of the first two trials of combined-task blocks.

Correlations with self-report measures were slightly higher for the data set that retained the first two trials. In addition, correlations of IAT extremity with respondents' average latencies on combined-task blocks were slightly lower with inclusion of the first two trials. Both of these results indicated that the first two trials of combined-task blocks were useful, despite their relatively high latencies. This pattern occurred similarly in the data sets for the Race, Age, and Gender–Science IATs. Accordingly, all of the following analyses included the data from the first two trials of combined-task blocks.

Data from Blocks 3 and 6. The conventional algorithm excludes trials from Blocks 3 and 6, treating them as practice. To assess the usefulness of these data, separate IAT measures were computed from Blocks 3 and 6 (practice) and from Blocks 4 and 7 (test). Remarkably, for all five pairs of IAT measures (median, mean, log, reciprocal, and *D*), correlations with explicit measures were higher for the measure based on Blocks 3 and 6 than for the measure based on Blocks 4 and 7. Further, the difference was more than trivial. The largest difference was for the reciprocal measure (practice $r = .635$; test $r = .478$). This discovery that practice blocks provided a good IAT measure was confirmed in the data sets for the Race, Age, and Gender–Science IATs.

To make use of the data from practice blocks, new IAT measures were computed as equal-weight averages of practice and test block measures for all five transformations. With the exception of the reciprocal measure, these *practice+test* measures yielded higher correlations with self-report than did either the practice measure or the test measure alone. For example, for the *D* measure, practice $r = .748$, test $r = .700$, and practice+test $r = .773$. Correlations of IAT measures with respondent average latency tended to be higher for the practice measure than for the test measure. For practice+test measures, the correlations with average latency tended to be similar to those for practice alone. Again using *D* for the illustration, practice $r = .073$, test $r = .048$, and practice+test $r = .070$.

Error latencies. It is common practice in studies with latency measures to analyze latencies only for correct responses. By contrast, the conventional IAT algorithm uses error latencies together with those for correct responses. Study 1 included analyses to compare the value of including versus excluding error latencies.

A preliminary analysis of the Election 2000 IAT data was limited to respondents ($n = 1{,}904$) who had at least two errors in each of Blocks 3, 4, 6, and 7. The analysis indicated that error latencies ($M = 1{,}292$ ms; $SD = 343$) were about 500 ms slower than correct response latencies ($M = 790$ ms; $SD = 301$). The increased latency of error trials is explained by the Web IAT's procedural requirement that respondents give a correct response on each trial. (Error feedback in the form of a red letter *X* indicated that the initial response was incorrect. Respondents' instructions were to give the correct response as soon as possible after seeing the red *X*.) Latencies on error trials therefore always included the added time required for subjects to make a second response.

A second preliminary analysis, which was limited to respondents who had self-characterized strong preference for either Gore or Bush, showed that error rates were higher when respondents were required to give the same response to their preferred candidate and unpleasant words ($M = 12.4\%$) than when giving the same response to their preferred candidate and pleasant words ($M = 5.5\%$).

Together, these two preliminary analyses suggested that inclusion of error latencies should enhance IAT effects. This enhancement should occur because errors were both (a) slower than correct responses and (b) more frequent when the task required giving the same response to nonassociated target–attribute pairs (e.g., the preferred candidate and unpleasant-meaning words). In a test for correlation of IAT measures with the combined self-report measure, the *D* measure performed better ($r = .753$) when error latencies were included than when they were excluded ($r = .730$). At the same time, the correlation with average response latency was only very slightly greater (which is undesirable) when error latencies were included ($r = .070$) than when they were excluded ($r = .063$). The increase in correlation with self-report amounts to a 3.0% increase in variance explained compared with an increase in variance explained of only 0.1% in the correlation of IAT with average latency. For this reason, it appeared very reasonable to retain error latencies in the IAT measures. Further alternatives for treating data from error trials are considered in Study 4.

In several ways, Study 1 demonstrated that inclusion of data is a generally good policy for the IAT. Improvements in performance were apparent in data sets that retained (a) the first two trials of combined-task blocks, (b) error latencies, and (c) data previously treated as practice (Blocks 3 and 6 in the IAT schema of Table 1). The greatest of these improvements of performance resulted from including data from Blocks 3 and 6 in addition to those from Blocks 4 and 7.

Study 2: Comparing Five Transformations of Latencies

Method

Results of Study 1 were applied in constructing data sets used for all of the remaining studies. The data sets for Studies 2–6 therefore used all trials from Blocks 3, 4, 6, and 7, including trials on which errors occurred. With this inclusive data set, the five measures described above under *Candidate Measures* were evaluated in terms of their correlation with explicit measures and their resistance to contamination by latency variations among respondents. These two performance criteria could be evaluated by examining *latency operating characteristic* (LOC) functions, which are plots of measures as a function of the latencies of the responses on which they are based (e.g., Lappin & Disch, 1972).

Results of Study 2 are shown in Figures 1 and 2 in the form of LOC plots for the implicit–explicit correlation and for the mean value of the IAT measure. The explicit measure used in the correlations for Figure 1 was (as described above) the average, for each respondent, of standardized values of a Likert-type measure of candidate preference and a difference measure created from thermometer-type measures of liking for each candidate (Bush and Gore). As a preliminary to constructing any LOC plots, an

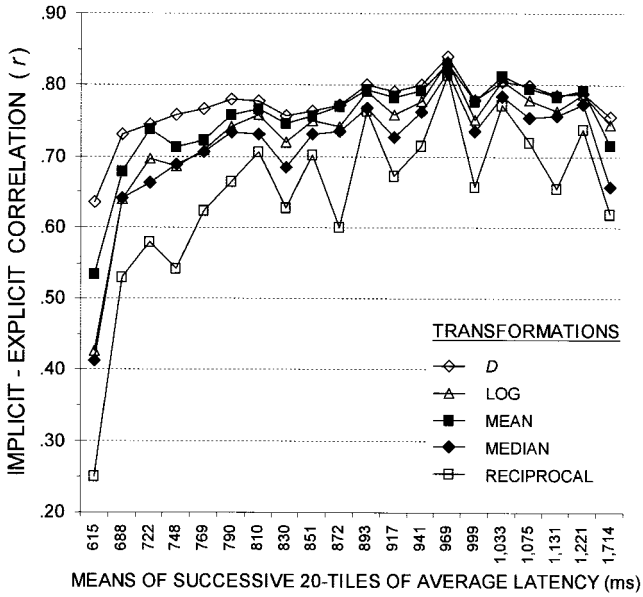

Figure 1. Latency operating characteristics (LOCs) for correlations with self-report for five Implicit Association Test (IAT) scoring algorithms. Higher correlations and flatter LOC curves indicate better performance. Data points are correlations for 20 groups of respondents, sorted by their response speed. Data are from Study 2, Election 2000 IAT data set. For each correlation, *n* ranges between 396 and 420.

average latency measure was computed for each respondent as an equal-weight average of mean latencies computed from each of the four data blocks (involving a total of 140 trials). In the sample of 8,891 respondents for whom this measure was available, average latencies had a mean of 929 ms (SD = 776) and ranged from 215 ms to 69,814 ms. (Such a high value was possible because these averages were computed before deleting latencies greater than 10,000 ms from the data set.) Using this measure, 20-tiles of the distribution were identified. The first 20-tile consisted of the 5% of the sample with fastest average latencies, and the last consisted of the 5% with slowest average latencies.

Results and Discussion

Figure 1 displays correlation LOCs for the median, mean, log, reciprocal, and *D* measures. These LOCs indicate better performance of the IAT measure to the extent that they are (a) high in elevation (higher correlations indicate better performance) and (b) level (i.e., flat), indicating consistency of the correlation across the wide range of respondent speeds. On both of these criteria, the *D* measure performed best of the five investigated transformations, and the reciprocal measure performed worst. That is, the LOC for the *D* measure was both higher and more level than the LOCs for the other four measures (see Figure 1). Differences among the measures are most noticeable at the fast (left) end of the LOCs. The measure using the mean was the second-best performer on both of the two desirable characteristics and is quite close to the best-performing *D* measure in the slower (right) half of the LOC.

Figure 2 displays LOCs for the means of the five measures, using data for the 5,202 respondents who indicated strong preference for either Gore or Bush on the Likert self-report measure. For this analysis, IAT values for Gore supporters were subtracted from zero so that all mean values were expected to be positive. For Figure 2's LOC, elevation is not a critical indicator because the several measures used four different numeric scales that are not directly comparable. (Only the median and mean share a metric.) On the basis of assuming that extremity of implicit candidate preferences of slow responders should not differ on average from that of fast responders, levelness of the LOC functions in Figure 2 is very desirable. For the LOCs shown in Figure 2, the mean and median measures performed quite poorly. For the median, the data suggested that implicit favorableness toward the preferred candidate of the slowest responders was over seven times that of the fastest responders (ratio = 7.09:1). For the mean, the corresponding figure was an almost equally poor 5.96:1. For the log, *D*, and reciprocal measures, the corresponding values were, respectively, 2.82:1, 1.42:1, and 1.26:1. Thus, all of the measures produced larger values of IAT measures for slow than fast responders, but the measures varied considerably in the extent to which their values were correlated with (i.e., contaminated by) response speed.

A simple summary of Figure 2's data is provided by the correlation of each IAT measure with response speed for the entire subsample of strong supporters. These correlations ranged from a low value of r = .050 for the reciprocal measure to a high of r = .344 for the mean. The other values were: *D* (r = .070), log (r = .226), and median (r = .309).

The brief summary of Study 2 is that overall, the *D* measure performed best. It showed clearly the best performance on the criterion of implicit–explicit correlation and was second best in

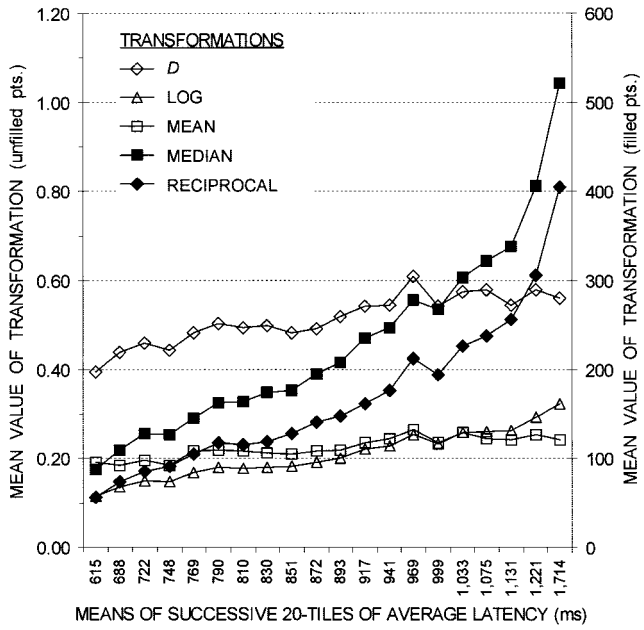

Figure 2. Latency operating characteristics (LOCs) for mean values of Implicit Association Test (IAT) measures for five scoring algorithms. More level LOC curves indicate better performance. Data points are means for 20 groups of respondents, sorted by their response speed. Data are from Study 2, Election 2000 IAT data set. Analyses were limited to respondents who indicated strong preference for either Bush or Gore on a self-report item; IAT scores for Gore supporters were reversed. For each mean, *n* ranges between 210 and 297. pts. = points.

having a low correlation with average latency. The reciprocal measure, which was best on the criterion of low correlation with average latency, performed so poorly on both elevation and levelness of the implicit–explicit correlation LOC (see Figure 1) as to remove it from competition for designation as the best-performing measure.

Study 3: Possible Respondent-Exclusion Criteria

In studies that use latency measures, it is routine to consider excluding subjects for either excessive slowness or excessive error rates. For the present data, it was appropriate also to consider exclusions for excessive speed, possibly produced by Web site visitors who were responding to the stimuli as rapidly as possible without even trying to classify them. Some such protocols might actually have been contributed by the researchers or their associates, who might have been proceeding rapidly through a Web IAT procedure only for the purpose of checking its operation.

Method

For each respondent in the Election 2000 data set, an overall measure of percent errors was computed, along with three summary measures based on response speed—average latency, percentage of "fast" (< 300 ms) responses, and percentage of "slow" (> 3,000 ms) responses. All measures were computed as unweighted averages of averages that were first computed separately for Blocks 3, 4, 6, and 7.[8]

Each of the four measures was initially examined to locate cut points that would exclude 0.25%, 0.5%, 0.75%, 1.0%, 2.5%, 5.0%, and 10.0% of respondents. The percentages excluded by the chosen cut points differed slightly from these target percentages because of the large numbers of ties in the sample for all of the measures except average latency. The cut points were then applied (for each measure separately) in an attempt to identify criteria that would produce a noticeable gain in performance of one or more of the five IAT transformations while keeping low the percentage of respondents lost to analyses by exclusion.

Results and Discussion

Performances of the five IAT measures (*D*, mean, median, log, and reciprocal) were examined in terms of each measure's correlation with (a) its parallel explicit measure for the entire sample (high values are desired) and (b) average latency for the subsample of self-characterized strong supporters of Bush or Gore (values near zero are desired, indicating lack of contamination of the measure by slowness of responding).

Somewhat surprisingly, average percentage of fast responses was the only dimension for which a relatively small exclusion of respondents achieved a clearly useful result. Figure 3 presents the data for correlation of the five IAT measures with explicit candidate preference as a function of exclusion criteria that eliminated successively increasing numbers of respondents. The *D*, log, mean, and median measures were arrayed in that order. Each showed mild increases in correlations with self-report as the exclusion criterion varied between unlimited inclusion of fast responses (n = 8,218) and zero tolerance for fast responses (n = 7,488, eliminating 8.9% of the sample). By comparison with the other four measures, the reciprocal measure showed dramatic improvement as more fast responders were excluded, indicating that its performance was most impaired by the presence of fast responses in the data set.

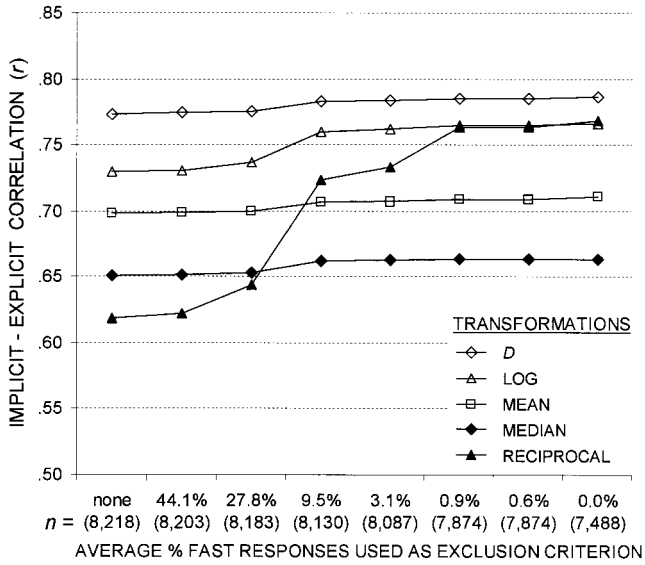

Figure 3. Effects of seven criteria for excluding respondents as a function of their proportion of fast (latency < 300 ms) responses on correlations with self-report for five Implicit Association Test (IAT) scoring algorithms. Higher correlations indicate better performance. The leftmost data point in each curve is for no exclusion of respondents. Both the exclusion criterion and the remaining sample size are indicated on the abscissa. Data are from Study 3, Election 2000 IAT data set. Maximum n = 8,218.

The *D* measure's maximum correlation with self-report (r = .787) was achieved in the analysis that was limited to respondents whose data contained no fast responses (right-most data point in Figure 3). However, this required eliminating 8.9% of respondents, which seemed overly costly in light of the small gain in implicit–explicit correlation beyond that achieved in the analysis that included respondents with up to 9.5% fast responses (r = .783, n = 8,130, eliminating only 1.1% of respondents).

Exclusions based on average error rates also produced some improvement in the implicit–explicit correlation. However, it was necessary to eliminate 9.4% of respondents on the basis of error rates in order to obtain the same improvement achieved by eliminating just 1.1% of respondents on the basis of average percentage of fast responses. Excluding 9.4% of respondents (which excluded all those with more than 17.5% errors) seemed an unacceptably large loss of data. Additional analyses that considered exclusions on the basis of the combination of average percent of fast responses and average error rates also provided insufficient gain to justify the additional losses of data.

The increase in implicit–explicit correlation for the best-performing *D* measure—from r = .773 (with no exclusion) to r = .783 (excluding respondents with more than 9.5% fast responses)—is not large. At the same time, the 1.5% increase in variance explained (from 59.8% = $.773^2$ to 61.3% = $.783^2$) is not trivial.

Figure 4 shows the effects of exclusions based on average percent of fast responses on the correlations of the five IAT

[8] Three additional measures were based on the maximum percentages of errors, slow responses, and fast responses observed in any single block. None of these maximum measures proved useful as a criterion on which to base exclusions. Consequently, they are not mentioned further.

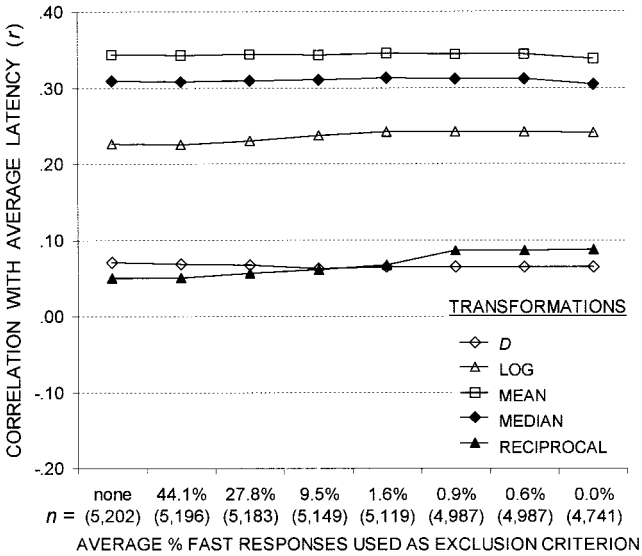

Figure 4. Effects of seven criteria for excluding respondents as a function of their proportion of fast (latency < 300 ms) responses on correlations with average response latency for five Implicit Association Test (IAT) scoring algorithms. Lower correlations indicate better performance. The leftmost data point in each curve is for no exclusion of respondents. Both the exclusion criterion and the remaining sample size are indicated on the abscissa. Data are from Study 3, Election 2000 IAT data set. Analyses were limited to respondents who indicated strong preference for either Bush or Gore on a self-report item; IAT scores for Gore supporters were reversed. Maximum n = 5,202.

measures with average latency. This is a correlation for which the desired result is close to zero—showing little or no contamination of the IAT measure by response speed. The reciprocal and *D* measures were the best performers, with correlations uniformly below r = .10 for all levels of exclusion. By comparison, the log, median, and mean measures performed poorly, all having correlations above r = .20 at all levels of exclusion. Interestingly, the exclusion policy based on average percent of fast responses that worked well for the criterion of implicit–explicit correlation simultaneously improved performance slightly for the *D* measure (i.e., lowering the correlation with average latency) while slightly impairing performance for the reciprocal measure (see Figure 4).

On the basis of Study 3, the remaining studies analyzed data both using all respondents and eliminating those with more than 10% fast responses. The criterion of 10% was selected arbitrarily as a rounded value of the 9.5% criterion that was successfully used for the Election 2000 data set in Study 3.

Study 4: Treatment of Trials With Error Responses

The most widely used method of dealing with latencies from trials with incorrect responses is simply not to use those latencies. Research reports often describe the proportion of trials on which errors occurred and then exclude those trials from analyses of latencies. This strategy seems quite satisfactory when, as often happens, independent variables have similar effects on latencies and error rates. That is, when treatments that produce higher response latencies also produce higher error rates, analyses of latencies and error rates will support the same conclusions. Furthermore, because effects on error rates are often weaker than those on latencies, the strategy of discarding error latencies is also considered satisfactory when effects on error rates are weak or nonsignificant. (However, cf. Wickelgren, 1977, who questioned the wisdom of treating nonsignificant error rate differences as ignorable.)

Study 1's results call into question the practice of routinely discarding error latencies. The relevant finding from Study 1 is that IAT measures showed higher implicit–explicit correlations when error latencies were included in analyses than when they were discarded. Study 4 was designed to consider, as strategies for error trials, procedures more elaborate than simply retaining or discarding error latencies. These alternatives involved replacing error latencies with values that functioned as error penalties.

Method

Analyses were conducted both on the full Election 2000 data set and on a data set that was reduced by eliminating the respondents for whom more than 10% of trials were faster than 300 ms (i.e., based on the results of Study 3). Because the previous studies had clearly established that the *D* measure was superior to other transformations (viz., mean, median, log, and reciprocal), the analyses in Study 4 and later studies were limited to variations of the *D* measure.

Five types of error treatments were evaluated in Study 3: (a) no treatment—latencies of error responses were used in the same fashion as those of correct responses; (b) deletion of error trials from the data set; (c) replacement of errors with the block mean of correct responses plus a constant (*penalty*; five penalties were used—200, 400, 600, 800, or 1,000 ms); (d) replacement of errors with the block mean of correct responses plus a penalty computed as the block's standard deviation of correct responses multiplied by a constant of 1.0, 1.5, 2.0, 2.5, or 3.0; and (e) replacement of errors with the block mean of correct responses plus a value computed as the block mean multiplied by 0.2, 0.4, 0.6, 0.8, or 1.0.

The various strategies used in Study 4 ranged from no penalty for errors (i.e., discarding error latencies) to penalties that were considerably larger than the built-in penalty provided by retaining error latencies. Study 1 had shown that the mean of correct responses averaged 790 ms (SD = 301), and error latencies averaged 502 ms slower than correct response latencies. Accordingly, the strategy of retaining error latencies was approximately equal to using a penalty in the middle of each of the three sets of five penalty computations.

Results and Discussion

Figure 5 shows the effect of 15 error-penalty strategies on correlation of the *D* measure with self-reported candidate preference. For comparison, values for two other strategies—error latencies used without alteration and error trials discarded—are shown. Three conclusions are apparent from the plotted results. First, and confirming a finding of Study 1, discarding error trials was an inferior strategy—indeed, inferior to all 16 other strategies plotted in Figure 5. Second, the most successful strategy was using unaltered error latencies. Third, among the 15 error-penalty formulas, most successful were ones that provided penalties that in average value were close to the average approximate 500-ms penalty that resulted from the procedural requirement to provide a correct response after making an error.

Figure 6 shows effects of the 15 error penalties and the two comparison conditions on correlations of the *D* measure with

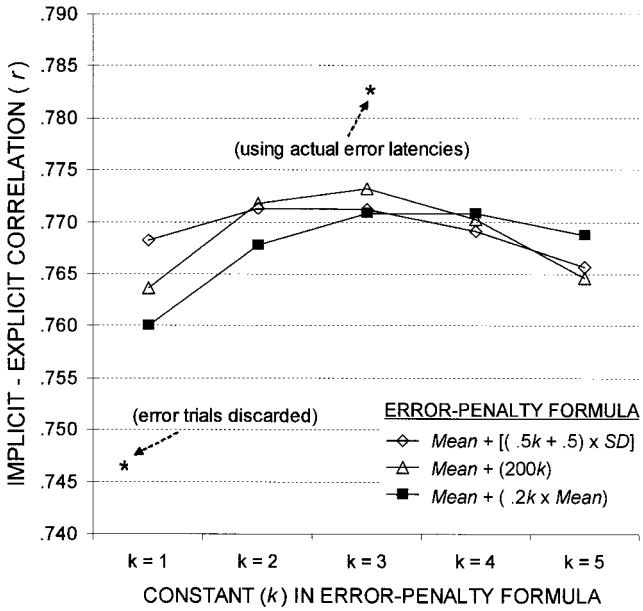

Figure 5. Effects of 15 strategies for error penalties on correlations with self-report for the *D* algorithm. Effects of using error latencies as is and of deleting error trials are shown as labeled asterisks. Higher correlations indicate better performance. Data are from Study 4, Election 2000 Implicit Association Test data set, excluding respondents who had more than 10% fast (< 300 ms) responses. $N = 8{,}132$.

average latency. For this measure, correlations close to zero are desired. The best results (i.e., smallest correlations) were obtained with error penalties that added a constant to the mean of correct responses. Use of unaltered error latencies produced a result that was near to the results of discarding error trials and using penalties computed as a constant proportion of the mean of correct responses (filled black squares in Figure 5).

Study 4 establishes that it is satisfactory to use unaltered error latencies in the Web IAT. This conclusion must be qualified by noting that in the Web IAT procedure, error latencies included the time required to produce a second response—in effect, they contained a built-in error penalty. The conclusion from Study 4, therefore, cannot be extended either to (a) procedures that do not require a correct response on each trial or (b) procedures that record the latency to the initial response (whether or not the error correction is required). For procedures with no built-in error penalty, Study 4 indicates that use of an error penalty is likely to produce better results than will be obtained with either unaltered error latencies or deletion of error trials. However, because several error-penalty formulas worked reasonably well, the results of Study 4 do not establish the clear superiority of any specific form of error penalty. The question of best form of error penalty is therefore deferred to Study 6, where results from all four data sets are jointly considered.

Study 5: Treatments of Trials With Extreme (Fast or Slow) Latencies

In addition to transformations such as logarithm and reciprocal, remedies for problems due to misshapen tails of latency distributions include (a) setting lower and/or upper bounds beyond which latencies are deleted from the data set and (b) similarly, using lower and/or upper bounds as values to which more extreme values are recoded (for simulation analyses of methods for dealing with extreme latency values, see Ratcliff, 1993; Miller, 1994). Study 5 examined both deletion and recoding-to-boundary strategies. As in Studies 3 and 4, performance of IAT measures was evaluated in terms of implicit–explicit correlations (higher values desirable) and correlations of the IAT measure with average latency (lower values desirable). As for Study 4, Study 5 was limited to the *D* measure because of its superior performance in Studies 1–3.

Method

Study 5 was conducted as three substudies. The first substudy examined deletion and recoding-to-boundary for the lower tail of the distribution, using boundaries of 300, 350, 400, 450, 500, or 550 ms. The second substudy examined deletion and recoding-to-boundary for the upper tail, using 6,000, 4,000, 3,000, 2,500, 2,250, and 2,000 ms as boundaries. The final substudy explored selected combinations of lower and upper boundaries.

Results and Discussion

Figure 7 presents the effects of the 36 extreme-value treatments on correlations of the *D* measure with the two-item measure of explicit candidate preference, Figure 8 presents the corresponding results for correlations with average latency. All of these correla-

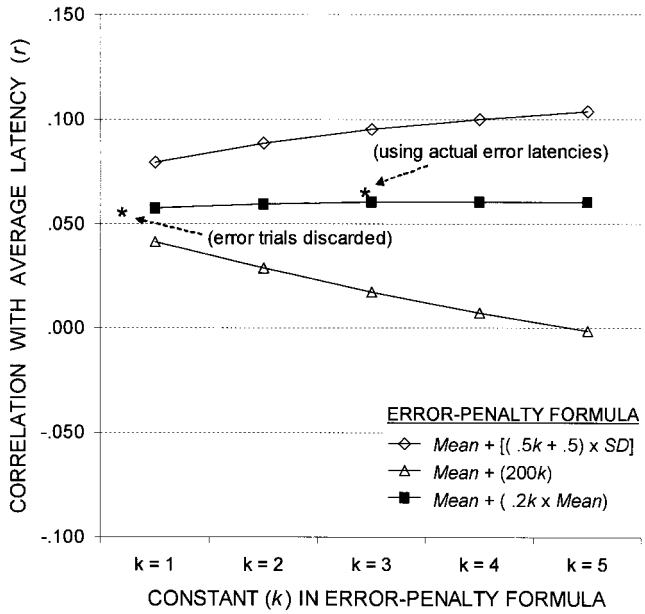

Figure 6. Effects of 15 strategies for error penalties on correlations with average response latency for the *D* algorithm. Effects of using error latencies as is and of deleting error trials are shown as labeled asterisks. Lower correlations indicate better performance. Data are from Study 4, Election 2000 Implicit Association Test (IAT) data set, excluding respondents who had more than 10% fast (< 300 ms) responses. Analyses were limited to respondents who indicated strong preference for either Bush or Gore on a self-report item; IAT scores for Gore supporters were reversed. $N = 5{,}151$.

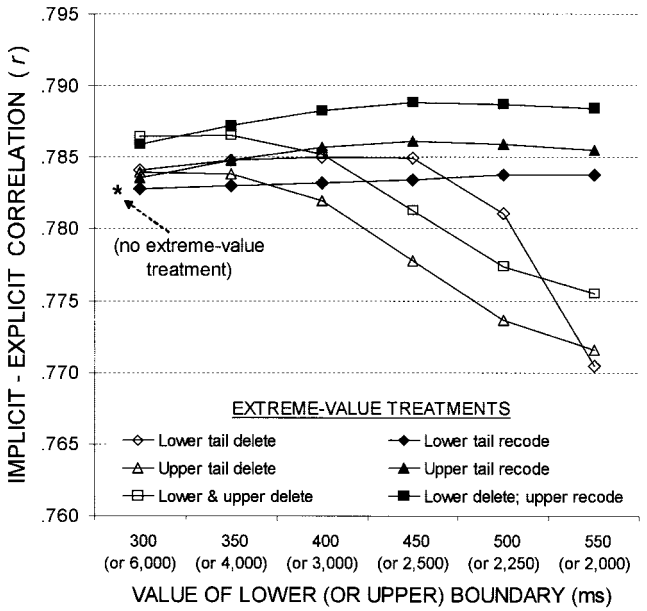

Figure 7. Effects of 36 strategies for treating low and high extreme latencies on correlations with self-report for the *D* algorithm. The correlation for data using no extreme-value treatment is shown as a labeled asterisk. Higher correlations indicate better performance. Data points to the left involve less severe extreme-value treatments than those to the right. Data are from Study 5, Election 2000 Implicit Association Test data set, excluding respondents who had more than 10% fast (< 300 ms) responses. $N = 8{,}132$.

tions were computed using the sample that was reduced (from $N = 8{,}218$ to $N = 8{,}132$) by excluding respondents who had more than 10% of responses faster than 300 ms (on the basis of Study 3). In both figures, an asterisk shows the result obtained when no extreme value treatment (beyond the initial deletion of latencies over 10,000 ms) was applied.

Lower tail treatments. In Figures 7 and 8, the curves with open and filled diamonds show, respectively, correlations involving IAT measures that used lower tail deletion and lower tail recoding with boundaries ranging from 300 to 550 ms. Figure 7 shows that lower tail deletion (open diamonds in Figure 7) produced small increases in the implicit–explicit correlation for lower boundary values up to 450 ms, above which performance was inferior to no lower bound treatment. Lower bound recoding produced virtually no change in the implicit–explicit correlation for all six boundary values that were examined. Figure 8 shows that the effect of lower bound treatments on correlations with average latency was nil for the lowest two boundary values for both deletion and recoding. At lower boundaries of 400 ms and above, contamination of measures by response speed increased for the lower bound deletion strategy, but not for lower bound recoding. All of these effects were small.

Upper tail treatments. Upper tail deletion (curves with open triangles in Figures 7 and 8) produced a very slight improvement in implicit–explicit correlation for the two highest boundary values (6,000 ms and 4,000 ms) and deterioration (relative to no upper boundary) at lower values of the upper bound (see Figure 7). For all six upper boundary values, the recoding-to-boundary strategy (filled triangles) produced a very small improvement in the implicit–explicit correlation. For the criterion of correlation of the *D* measure with average latency, both strategies (deletion and recoding-to-boundary) yielded inferior performance (i.e., higher values) compared to no upper tail treatment (see Figure 8).

Combined lower and upper tail treatments. The curves marked by open and filled squares in Figures 7 and 8 show results for the combination of deletion of values below 400 ms with all of the upper bound treatments. In Figure 7, both deletion (open squares) and recoding (filled squares) yielded improvements relative to the 400-ms lower bound deletion alone for the two widest upper boundary values (6,000 ms and 4,000 ms). At narrower values, this mild improvement was retained for the recoding strategy but not for the deletion strategy. The results for the criterion of correlation with average latency (Figure 8) were very similar to those for upper tail treatments without any lower tail treatment (see preceding paragraph). That is, these results were consistently inferior to using no deletion or recoding (marked by the asterisk in Figure 8).

In summary, performance of the *D* measure was virtually unaltered by lower bound recoding at any value (filled diamonds in Figures 7 and 8). Upper tail recoding modestly improved implicit–explicit correlations at all upper bound values but consistently increased contamination by average response latency, as did upper tail deletion. The highest value of the implicit–explicit correlation ($r = .789$) occurred for the combination of deletion below 400 ms and recoding values above 2,500 ms to 2,500 ms. However, all of

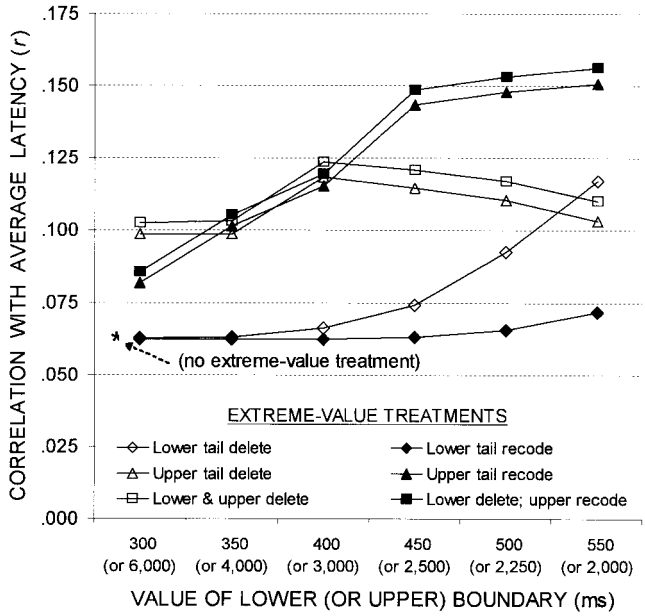

Figure 8. Effects of 36 strategies for treating low and high extreme latencies on correlations with average response latency for the *D* algorithm. The correlation for data using no extreme-value treatment is shown as a labeled asterisk. Lower correlations indicate better performance. Data points to the left involve less severe extreme-value treatments than those to the right. Data are from Study 5, Election 2000 Implicit Association Test (IAT) data set, excluding respondents who had more than 10% fast (< 300 ms) responses. Analyses were limited to respondents who indicated strong preference for either Bush or Gore on a self-report item; IAT scores for Gore supporters were reversed. $N = 5{,}151$.

the strategies involving upper tail treatments had the undesirable effect of increasing contamination by average response latency (see Figure 8). By contrast, the strategy of lower tail deletion at 350 or 400 ms produced a small improvement in implicit–explicit correlation ($r = .785$) without increasing (or decreasing) contamination by average latency.

In summary of Study 5, gains in implicit–explicit correlation resulting from deletion or recoding of extreme values were small. Some of these small increases were accompanied by (undesired) increases in the correlation of IAT scores with average latency. As a consequence of these observations, judgment about the value (if any) of extreme-value treatments should await consideration of results from Study 6, which used additional performance criteria.

Study 6: Additional Performance Criteria and Additional Data Sets

Summary of Studies 1–5

Using the criterion of implicit–explicit correlation, Study 1 found that IAT measures were improved (a) slightly, by including the first two trials of combined-task blocks (which had previously been deleted from analyses), and (b) substantially, by incorporating the data from two blocks that had previously been treated as practice trials. Study 2 established that the *D* measure was superior to other transformations (mean, median, log, and reciprocal) both in magnitude of the implicit–explicit correlation and in minimizing variations in that correlation across variations in respondents' average speed of responding. Study 2 also showed that the *D* measure was satisfactory in having a low correlation of the IAT measure with average response latency. Study 3 demonstrated the value of excluding a small proportion of respondents for whom 10% or more of responses were faster than 300 ms. Study 4 extended Study 1's finding that it was useful to retain latencies from error trials. That is, Study 4 showed gains, relative to deletion of error trials, achieved by replacing error latencies with values that functioned as error penalties. Study 5 found a very small improvement in the *D* measure when responses faster than 400 ms were deleted from respondents' data sets.

The goal of Study 6 was to evaluate all of the scoring strategies that appeared promising in Studies 1–5. Seven performance criteria were used to evaluate these finalists. The first two of these were the two important criteria that had been used in Studies 1–5: (a) implicit–explicit correlation and (b) resistance to contamination related to speed of responding. The additional five performance criteria were (c) internal consistency, measured by the correlation between one IAT measure based on Blocks 3 and 6 and another based on Blocks 4 and 7 (see Table 1); (d) resistance to the often-observed *order effect* (i.e., associations appear stronger when they are tested in Blocks 3 and 4 rather than in Blocks 6 and 7); (e) resistance to the reduction in IAT scores that is typically observed among those who have previously completed one or more IATs; (f) sensitivity to modal response tendencies (e.g., the Age IAT typically shows considerably stronger association of young than old with pleasant); and (g) magnitude of the standardized coefficient for the path between latent implicit and explicit variables in a confirmatory factor analysis (CFA).

Method

It was necessary first to choose measures for inclusion in Study 6. Study 2 had made clear that the *D* measure, which uses each respondent's latency variability to provide the unit for the IAT measure, decisively outperformed the four measures that were not so calibrated—that is, the measures based on the median latencies in each block, means of untransformed latencies, or means of logarithm or reciprocal transformations. This superiority of the *D* transformation was as apparent in the other three IAT data sets (Race, Age, and Gender–Science) as it was in the Election 2000 data set. Accordingly, Studies 3–6 focused on variations of the *D* measure.

Study 4 examined 17 strategies for dealing with error trials. Study 5 examined 13 strategies for dealing with extreme latencies at each of the upper and lower tails of latency distributions. There were 2,873 ($= 17 \times 13 \times 13$) possible combinations of these error and extreme-value treatments. In addition, Study 3 evaluated eight cut points on each of four dimensions as bases for excluding subjects, along with an additional eight that combined two criteria, for a total of 40. Adding the four additional combinations of including or excluding the first two trials of each block and using or not using Blocks 3 and 6, the number of available combinations of the variations on the *D* measure that were examined in Studies 1–5 approached half a million.

Because of the huge number of possible strategy combinations for the *D* measure, it was necessary to select for Study 6 a severely restricted subset. To do that, the authors conducted Studies 1–5 on the remaining three IAT data sets (Age attitude, Race attitude, and Gender–Science stereotype). The hope was that the different data sets would reinforce each other to identify just a few successful strategies from each study. Study 6 would then examine these individually and in combination, with the hope that the combined results for Study 6's seven performance criteria (described in the third paragraph above) would allow settling on one, or at most a very few, variations of the *D* measure as an improved scoring algorithm for the IAT.

On the basis of a review of results from the four IAT data sets, six variations of the *D* measure were selected for Study 6, identified as D_1–D_6. D_1 was the simplest, involving no adjustment beyond the preliminary deletion of latencies over 10,000 ms that was done for all measures. D_2 additionally deleted latencies below 400 ms (on the basis of Study 5). The remaining four *D* variations included error penalties (on the basis of Study 4). D_3 replaced error trials with the mean of correct responses in the block in which the error occurred plus a penalty of twice the standard deviation of correct responses in the block in which the error occurred. D_4 replaced error trials with the mean of correct responses plus 600 ms. D_5 and D_6 used the same error penalties as D_3 and D_4 and additionally deleted latencies below 400 ms.

For purposes of comparison, Study 6 included four variations of the conventional IAT measure, identified as C_1–C_4. C_1 was the measure originally recommended by Greenwald et al. (1998) for use in statistical tests. This measure used data only from Blocks 4 and 7 (excluding their first two trials), recoded latencies outside boundaries of 300 ms and 3,000 ms to those boundary values, and log-transformed the resulting values before taking the difference between means for the two blocks. C_2 differed from C_1 only by omitting the log transformation; this was the measure used by Greenwald et al. (1998) for graphic or tabular presentation of results (because its millisecond units are more understandable than the log-transformed units). C_3 used the same computational procedures as C_1, but paralleled the *D* measures by (a) retaining the first two trials of combined-task blocks, (b) computing an additional measure on the basis of Blocks 3 and 6, and then (c) averaging the two resulting scores. C_4 was the same as C_3 but omitted the log transformation so that it had millisecond units (like C_2).

Performances of the 10 measures, D_1–D_6 and C_1–C_4, were evaluated on the seven criteria (see the paragraph just before this *Method* section) for all four IATs. The procedure used to measure each criterion is described together with the presentation of its results.

Results and Discussion

Table 2 summarizes results for the Election, Gender–Science, Race, and Age IATs on the seven performance criteria. Entries in Table 3 are averages of the corresponding entries in Table 2, computed using Fisher's *r*-to-*Z* transformation.

Implicit–explicit correlation. For all four IATs, the explicit measure was the previously described one, an average based on one Likert-type item and two thermometer-format items. As in Studies 1–5, the two thermometer items were combined into a single score by taking their difference. Standardized transforms of this difference score and the Likert item score were averaged into the explicit (self-report) measure that was correlated with the 10 variants of the IAT measure. The first data rows of Table 2, Sections A–D, and Table 3 show results for these implicit–explicit correlations. With the lone exception of Measure C_3 in Table 2, Section B, the 6 *D* measures outperformed all of the conventional measures in every analysis. Table 3 shows that D_6, which combined deletion of values below 400 ms with a 600-ms error penalty, slightly outperformed the other error-penalty formulas, D_3, D_4, and D_5. At the same time, the 2 measures that used unaltered error latencies, D_1 and D_2, outperformed the 4 *D* measures that used error penalties.

Resistance to contamination related to speed of responding. Each respondent's overall latency of response was summarized by computing an unweighted average of the mean latencies for the four combined-task blocks (Blocks 3, 4, 6, and 7). The possible contamination of IAT measures by response-speed differences among respondents was examined by using this average latency measure to construct LOCs, of the types reported previously for Study 2 (see Figures 1 and 2). The results of those LOC analyses are well represented by the correlations of the 10 IAT measures with overall latency as presented in the second data rows of Table 2, Sections A–D, and Table 3. For all of these correlations, the desired result is $r = 0$, which would indicate absence of contamination of the IAT measure by differences in overall response speed. The *D* measures were uniformly superior to all of the conventional measures (i.e., closer to $r = 0$) for all four IATs individually as well as for their average, which is shown in Table 3. In this case, superior performance was provided by two of the *D* measures that incorporated error penalties, D_4 and D_6. Their values averaged very close to zero. By comparison, for the four conventional measures, average correlations ranged between .157 and .296, revealing a substantial level of contamination by individual differences in response speed. The log-transform versions of the conventional procedure (C_1 and C_3) had noticeably less contamination by response speed than did the two measures that used millisecond units (C_2 and C_4).

Internal consistency. For 8 of the 10 measures included in the data columns in Table 2, an internal consistency measure was provided by the correlation between a measure based on Blocks 3 and 6 and one based on Blocks 4 and 7. This strategy was not available for the 2 conventional measures that used data only from Blocks 4 and 7. For those 2 measures (C_1 and C_2), the internal consistency correlation was computed as the correlation between an IAT measure based on Trials 3–20 in each of Blocks 4 and 7 and a second IAT measure based on Trials 21–40 in those same blocks. Overall (see Table 6) the best-performing measure was C_3, which applied the conventional IAT scoring procedures to data from four blocks of trials. Among the *D* measures, the two that did not use error penalties, D_1 and D_2, produced higher internal consistency correlations than did the four that used error penalties.

Somewhat surprisingly, these results indicated that measures with relatively poor performance on the major criteria (implicit–explicit correlation and resistance to contamination by response speed) had superior performance on internal consistency correlations. This result suggests the possibility that artifactual variance contributed to internal consistency of the conventional measures. For example, the artifact associated with average response latency (see row 2 of Table 3) accounted for between 2.5% and 8.8% of variance in the conventional measures. To the extent that the conventional measures assess this artifact reliably the artifact will contribute to their internal consistency, but the resulting increase in internal consistency does not indicate an increase in validity of the measure as a measure of association strength. For this reason, it may be appropriate to treat internal consistency as an uncertain guide to construct validity.

Order effect. The very first IAT studies (Greenwald et al., 1998) observed effects of the order in which the two possible task combinations of each IAT were administered. For example, when the first task in a flower–insect attitude IAT was to respond with one response key to flower names and pleasant words and with the other key to insect names and unpleasant words, performance of that task was faster than when it was done second. This may involve the familiar phenomenon of negative transfer (e.g., Woodworth & Schlosberg, 1954), whereby practice at one task interferes with performance at a second task that requires giving different responses to the first task's stimuli. The result of this negative transfer is that the strength of flower–pleasant associations appears greater when the task that uses this association—the task requiring the same response to flower names and pleasant words—comes first.

The order effect just described has been observed frequently but not invariably. Ideally, an IAT measure should be free of this order effect. Table 2 summarizes magnitudes of observed order effects in the four data sets. These are shown as correlations of each IAT measure with a dichotomous measure of the order in which the two tasks were performed. The dichotomous measure was always scored so that the order effect would appear as a positive value of this correlation.

The magnitudes of order effects varied considerably across the four IATs. The effects were noticeably lower for the Election IAT (average *r* in row 4 of Table 2, Section A = .056) and the Race IAT (Table 2, Section C, average $r = .024$) than for the Gender–Science and Age IATs (average *r*s = .278 and .173, respectively, in Table 2, Sections B and D). These varying magnitudes of the order effect were almost certainly due to differences in procedures among the four IATs. The two IATs with small order effects incorporated extra trials in either Block 5 or Block 6 of the IAT procedure (see Table 1 note). These extra trials for the second combined task likely overcame some of the negative transfer resulting from tasks performed in Blocks 1, 3, and 4.

On average, the order effects were similar in magnitude for the *D* measures and the conventional measures (see Table 3). However, it is appropriate to look at the data just for the two IATs (Gender–Science and Age) for which noticeable order effects were observed. For these (see Table 2, Sections B and D), the *D* measures unexpectedly showed somewhat larger order effects than

Table 2
Performance of 10 Measures on Seven Criteria

Characteristics of measures	Variations of *D* measure						Conventional measures		Conventional measures with added trials	
	D_1	D_2	D_3	D_4	D_5	D_6	C_1	C_2	C_3	C_4
Included trials	All trials of Blocks 3, 4, 6, and 7						Trials 3–40 of Blocks 4 and 7		All trials of Blocks 3, 4, 6, and 7	
Lower tail treatment	None	Delete if < 400 ms	None		Delete if < 400 ms		Recode latencies < 300 ms to 300 ms			
Upper tail treatment	Delete if latency > 10,000 ms						Recode latencies > 3,000 ms to 3,000 ms			
Error treatment	Include error latencies in analyses		Replace errors: mean(C) + 2 *SD*	Replace errors: mean(C) + 600 ms	Replace errors: mean(C) + 2 *SD*	Replace errors: mean(C) + 600 ms	Include error latencies in analyses			
Latency transformation	Modified effect size computation (see text)						Logarithm	None	Logarithm	None
	Variations of *D* measure						Conventional measures		Conventional measures with added trials	
Seven performance criteria	D_1	D_2	D_3	D_4	D_5	D_6	C_1	C_2	C_3	C_4
A: Election 2000 IAT data[a]										
1. Implicit–explicit corr.	.783	.785	.771	.773	.767	.773	.687	.663	.758	.733
2. Corr. with average latency	.063	.066	.095	.017	.097	.019	.176	.289	.229	.364
3. Internal consistency corr.	.764	.767	.740	.747	.728	.743	.665	.636	.763	.743
4. Order effect corr.	.091	.086	.049	.048	.041	.039	.052	.044	.051	.043
5. Corr. with IAT experience	−.023	−.027	−.062	−.030	−.069	−.036	−.014	−.034	−.082	−.113
6. IAT effect size	1.54	1.55	1.44	1.46	1.41	1.45	1.10	1.01	1.35	1.21
7. Implicit–explicit path in CFA	.858	.860	.853	.854	.850	.854	.787	.770	.831	.810
B: Gender–Science IAT data[b]										
1. Implicit–explicit corr.	.251	.254	.239	.239	.239	.241	.196	.186	.240	.227
2. Corr. with average latency	.064	.065	.056	.024	.055	.023	.158	.247	.168	.281
3. Internal consistency corr.	.594	.598	.579	.589	.572	.587	.598	.566	.624	.603
4. Order effect corr.	.251	.257	.296	.288	.302	.297	.279	.270	.261	.253
5. Corr. with IAT experience	−.094	−.097	−.100	−.096	−.102	−.100	−.095	−.102	−.117	−.138
6. IAT effect size	1.04	1.05	1.00	1.00	0.99	1.00	0.81	0.75	0.98	0.93
7. Implicit–explicit path in CFA	.326	.328	.311	.311	.316	.313	.256	.246	.304	.291
C: Race IAT data[c]										
1. Implicit–explicit corr.	.359	.361	.359	.357	.360	.358	.292	.271	.343	.322
2. Corr. with average latency	−.018	−.017	−.010	−.058	−.010	−.059	.090	.176	.105	.211
3. Internal consistency corr.	.564	.566	.556	.558	.546	.548	.579	.562	.593	.580
4. Order effect corr.	−.023	−.017	.039	.030	.045	.040	.054	.052	−.003	−.002
5. Corr. with IAT experience	−.089	−.090	−.096	−.084	−.095	−.085	−.123	−.135	−.151	−.174
6. IAT effect size	1.00	1.00	0.99	1.00	0.98	1.00	0.82	0.75	0.91	0.84
7. Implicit–explicit path in CFA	.465	.468	.467	.464	.470	.467	.374	.351	.436	.411
D: Age IAT data[d]										
1. Implicit–explicit corr.	.170	.172	.172	.175	.174	.178	.106	.091	.137	.113
2. Corr. with average latency	.051	.051	.042	−.001	.039	−.004	.203	.300	.204	.325
3. Internal consistency corr.	.521	.523	.524	.527	.512	.520	.574	.567	.571	.566
4. Order effect corr.	.127	.134	.197	.181	.204	.191	.189	.183	.150	.141
5. Corr. with IAT experience	−.204	−.208	−.200	−.188	−.203	−.192	−.205	−.210	−.250	−.266
6. IAT effect size	1.38	1.39	1.33	1.34	1.32	1.33	1.08	0.99	1.25	1.14
7. Implicit–explicit path in CFA	.227	.230	.231	.233	.236	.239	.139	.119	.177	.147

Note. Abbreviations for the 10 measures (D_1–D_6 and C_1–C_4) are explained in the *Method* section of Study 6. The seven performance criteria are described in detail in the *Results* section of Study 6. On the basis of Study 3, samples excluded respondents for whom more than 10% of IAT responses were faster than 300 ms. mean(C) = block mean of correct-response latencies; *SD* = block standard deviation of correct-response latencies; IAT = Implicit Association Test; corr. = correlation; CFA = confirmatory factor analysis.

[a] N = 8,132 for Criteria 1 and 7; 5,151 for Criteria 2 and 6; 8,784 for Criteria 3 and 4; and 4,908 for Criterion 5. [b] N = 10,475 for Criteria 1 and 7; 11,549 for Criteria 2, 3, 4, and 6; and 10,509 for Criterion 5. [c] N = 6,811 for Criteria 1 and 7; 7,734 for Criteria 2, 3, 4, and 6; and 6,307 for Criterion 5. [d] N = 10,537 for Criteria 1 and 7; 11,384 for Criteria 2, 3, 4, and 6; and 7,194 for Criterion 5.

Table 3
Performance of 10 Measures on Seven Criteria (Average of Four IATs)

Characteristics of measure	Variations of *D* measure						Conventional measures		Conventional measures with added trials	
	D_1	D_2	D_3	D_4	D_5	D_6	C_1	C_2	C_3	C_4
Included trials	All trials of Blocks 3, 4, 6, and 7						Trials 3–40 of Blocks 4 and 7		All trials of Blocks 3, 4, 6, and 7	
Lower tail treatment	None	Delete if < 400 ms	None		Delete if < 400 ms		Recode latencies < 300 ms to 300 ms			
Upper tail treatment	Delete if latency > 10,000 ms						Recode latencies > 3,000 ms to 3,000 ms			
Error treatment	Include error latencies in analyses		Replace errors: mean(C) + 2 *SD*	Replace errors: mean(C) + 600 ms	Replace errors: mean(C) + 2 *SD*	Replace errors: mean(C) + 600 ms	Include error latencies in analyses			
Latency transformation	Modified effect size computation (see text)						Logarithm	None	Logarithm	None
	Variations of *D* measure						Conventional measures		Conventional measures with added trials	
Seven performance criteria	D_1	D_2	D_3	D_4	D_5	D_6	C_1	C_2	C_3	C_4
1. Implicit–explicit corr.	.434	.436	.425	.426	.424	.428	.347	.326	.408	.383
2. Corr. with average latency	.040	.041	.046	−.005	.045	−.005	.157	.254	.177	.296
3. Internal consistency corr.	.621	.624	.608	.614	.597	.608	.605	.584	.645	.629
4. Order effect corr.	.113	.116	.147	.138	.150	.144	.145	.139	.116	.110
5. Corr. with IAT experience	−.103	−.106	−.115	−.100	−.118	−.104	−.110	−.121	−.151	−.173
6. IAT effect size	1.240	1.248	1.190	1.200	1.175	1.195	.953	.875	1.123	1.030
7. Implicit–explicit path in CFA	.530	.533	.525	.525	.525	.527	.434	.413	.491	.464

Note. Abbreviations for the 10 measures (D_1–D_6 and C_1–C_4) are explained in the *Method* section of Study 6. The seven performance criteria are described in detail in the *Results* section of Study 6. For performance criterion 6, entries in this table are averages of the four corresponding entries in Tables 2–5. For the remaining (correlational) criteria, entries in this table were computed by first converting the entries in Tables 2–5 to Fisher's *Z* and then reconverting the averaged *Z*s to *r*. mean(C) = block mean of correct-response latencies; *SD* = block standard deviation of correct-response latencies; IAT = Implicit Association Test; corr. = correlation; CFA = confirmatory factor analysis.

the C measures; the *D* measures that used computed error penalties (D_3–D_6) showed larger order effects than those that had built-in error penalties (D_1 and D_2). These observations are considered further in the General Discussion.

Resistance to the effect of prior IAT experience. One of the optional self-report questions on the IAT Web site asked about the respondent's number of prior completed IATs. There were five reporting options: 0, 1, 2, 3–5, and 6 or more. It was known from previous analyses that prior experience with the IAT was associated with a reduction in IAT scores for those who reported one or more prior uses, compared with those reporting zero prior uses (see Greenwald & Nosek, 2001). Little or no further reduction in IAT scores occurred for two or more previous uses. Accordingly, the five-choice measure of prior IAT experience was reduced to a dichotomy that distinguished zero from one or more prior uses.

It is desirable for an IAT measure not to be affected by previous experience taking the IAT. The effect of prior experience means that scores of IAT novices cannot be compared directly with those of non-novices and, for the same reason, posttests cannot be compared directly with pretests (when the pretest is the first IAT taken). The desired correlation of an IAT measure with the dichotomous prior experience measure is therefore zero. However, the expectation based on previous observations is that this correlation will be negative—that is, numerically less extreme IAT scores will be observed for those with prior IAT experience.

The fifth data rows of Table 2, Sections A–D, and Table 3 report correlations with the prior experience measure. These correlations were uniformly negative, as expected. The six *D* measures varied little and performed noticeably better (i.e., had lower correlations) than the two conventional measures that used data from all four blocks (C_3 and C_4).

Sensitivity to modal response tendencies. The Age, Race, and Gender–Science IATs typically show, respectively, stronger association of young than old with pleasant, stronger association of European American than African American with pleasant, and stronger associations of female with arts and male with science than of female with science and male with arts. For the Election 2000 IAT there was no similar modal tendency in the population of respondents. However, there was a strong difference in IAT scores between self-identified (on the 5-point Likert item) strong supporters of Bush and Gore. That difference was used in the test for modal response tendencies.

The sixth data rows of Table 2, Sections A–D, and Table 3 report these modal tendencies as *d* effect sizes. For the Election 2000 IAT, the *d* measure derives from the two-group comparison of strong Bush and strong Gore supporters. For the other three IATs, it is the one-sample effect size of the entire sample's grand mean difference from zero. The computational procedure for the Election 2000 IAT made the *d* measure partly redundant with the implicit–explicit correlation that appears in the first data row of

each table. However, there was no such redundancy for the other three IATs. Tables 2 and 3 show that the six *D* measures were consistently more sensitive to modal response tendencies than were the four conventional measures. Among the *D* measures, the two that involved only the built-in error penalty (D_1 and D_2) were slightly superior, on average, to the four that used a computed error penalty.

Magnitude of implicit–explicit path in CFA. Two explicit measures (Likert and thermometer difference) were available for each IAT data set, and two submeasures of each IAT (the two used in the internal consistency correlations) were also available. These four measures were sufficient to permit a confirmatory factor analysis (CFA) that used two measures to identify a latent explicit factor and two measures to identify a latent implicit factor. Goodness-of-fit statistics obtained in the various CFAs indicated, without exception, that this two-factor model fit all of the data very well. The seventh data row of each table shows the standardized coefficients for the path between latent implicit and explicit factors obtained from each of the CFAs. This path coefficient can be understood as an estimate of the correlation that might be observed between error-free implicit and explicit measures. Consistently higher values of this path were obtained for the six *D* measures than for the four conventional measures. Although D_1 and D_2 were slightly superior to the other four *D* measures, it can be seen in Table 3 that, on average, there was very little difference among the six *D* measures.

Comparison of the six D measures. The main purpose of Study 6 was to identify one or more superior variations of the *D* measure. The six *D* variations selected for use in Study 6 varied along two dimensions: (a) treatment of fast responses (deletion of latencies below 400 ms vs. no deletion) and (b) treatment of error trials (use of error latencies unaltered vs. replacement of errors with the mean plus twice the standard deviation of correct latencies in the block in which the error occurred vs. error replacement by block mean of correct latencies plus 600 ms). These variations had been selected on the basis of their superiority over other strategies for treating extreme latencies and errors in Studies 4 and 5.

Tables 2 and 3 show that the differences among the six finalist *D* measures were neither large nor fully consistent across performance criteria or data sets. Because no single *D* variation clearly separated itself from the other five in Study 6, conclusions about the features that should be included in a revised IAT scoring algorithm are deferred to the General Discussion.

General Discussion

The present findings call strongly for replacing the IAT's conventional scoring procedure. The conventional IAT algorithm was decisively outperformed by all six *D* measures selected for Study 6. This superiority of the *D* measures was evident on five performance criteria: (a) magnitude of implicit–explicit correlation, (b) resistance to contamination by response speed differences, (c) resistance to the IAT-score-reducing effect of prior experience with the IAT, (d) sensitivity to known effects on IAT measures, and (e) latent implicit–explicit path in CFAs.

This discussion focuses first on the possibility of an alternative interpretation of the important criterion of magnitude of implicit–explicit correlations; next on the two performance criteria that diverged from the other five—internal consistency and the effect of order of combined tasks; and then on practical issues of applying the present results to research uses of the IAT.

Further Consideration of Performance Criteria

Implicit–explicit correlations. Implicit–explicit correlations were higher for the *D* measures than for all other algorithms. This result was observed consistently in all four IAT domains. As developed in the introduction, these higher implicit–explicit correlations can indicate greater construct validity of an IAT measure if association strengths are a component of both the implicit and explicit measures. This was illustrated in the introduction by analogy to the way in which an improved measure of height can produce a larger correlation between height and weight. The height–weight relation was proposed as an appropriate example because, conceptually, height is a component of both measures.

There are also circumstances in which finding that a modified measure yields a larger correlation with another measure can indicate reduced construct validity for the modified measure. Suppose, for example, that modification of a measure of quantitative aptitude increases its correlation with a measure of verbal aptitude. This increased correlation could be due to the modified quantitative aptitude measure containing greater contamination with verbal aptitude. This state of affairs might plausibly occur if the modified quantitative measure has a higher proportion of word problems relative to problems represented more abstractly with numbers or symbols. In this verbal–quantitative example, the shared component that increases the correlation is not a construct-valid aspect of quantitative aptitudes.

This article's use of implicit–explicit correlations as positive indicators of construct validity rests on the belief that components of these correlations are better modeled by the height–weight example than by the verbal–quantitative example. In order for the verbal–quantitative example to provide the superior model, the *D* measure would have to exceed the other algorithms in capturing some nonassociative component of the self-report measures—for example, impression management. However, there is no plausible basis for that conclusion. Additional basis for the conclusion that the *D* transformation is superior in construct validity comes from unpublished analyses of other Web IAT data sets by the second author (Nosek, 2003) showing higher correlations of the *D* measure with several demographic and sociopolitical measures that were hypothesized to be related to the association strengths measured by the IAT.

Internal consistency. Highest internal consistency was unexpectedly observed for Measure C_3 (see the third data row of Table 3). On discovering this result in Study 6, the authors suggested that the higher internal consistency of C_3 might be due to its being more reliably sensitive than other measures to an artifact associated with latency differences among respondents. This effect of latency differences could increase internal consistency without contributing to construct validity. Unfortunately, the present data sets provide no decisive means of evaluating this speculation.

Resistance to the effect of order of combined tasks. For the criterion of resistance to the effect of order of administering the IAT's combined tasks, Study 6 found that the *D* measures showed less resistance to this undesired effect than did the conventional measures. This was apparent for the two IATs (Gender–Science and Age) for which substantial order effects were observed (see

the fourth data rows in Table 2, Sections B and D). The negative transfer interpretation of the order effect (described in Study 6) interprets the order effect as an influence of IAT procedures on the strengths of the associations being measured. With this interpretation, the *D* measure's greater order effects are consistent with the *D* measure's construct validity. Nevertheless, the order effect remains undesirable. Fortunately, variations in magnitude of order effects among the four IATs indicate that it is possible to avoid this undesired procedural influence on IAT scores by increasing the numbers of trials either in Block 5 or Block 6 of the IAT procedure (for more evidence of the success of this procedural adjustment, see Nosek, Greenwald, & Banaji, 2003).

Choice Among Variations of the D Measure

By small margins, best average performances on three of the five performance criteria that indicated superiority of the *D* measure were obtained when latencies lower than 400 ms were deleted (see Measure D_2 in Table 3, first, sixth, and seventh data rows). Although D_2 used no special treatment of errors, it had the substantial built-in error penalty created by the Web IAT's requirement to provide a correct response after any error. For the four *D* measures that replaced error latencies with computed penalties, there was virtually no difference between the two measures that deleted latencies below 400 ms (D_5 and D_6) and the two that did not (D_3 and D_4). These observations very slightly favor the strategy of deleting latencies below 400 ms. However, the gain appears so slight as to make the strategy questionable.

Study 4 had previously demonstrated that superior results were achieved on the criterion of implicit–explicit correlation for the strategy of retaining error latencies compared with deleting error trials. Furthermore, with the exception of the Gender–Science IAT, Study 4 also found that all error-penalty formulas yielded higher implicit–explicit correlations than did the strategy of deleting error trials. Study 6 examined in greater detail the two error-penalty formulas that had performed best among the larger number examined in Study 4. Study 6 provided no basis for concluding that either of these two error-penalty formulas was superior to the other or to the procedurally built-in error penalty. Rather, the built-in error penalty of D_1 and D_2 was slightly superior to the calculated error penalties.

The only confident conclusion about preferred form of the *D* measure to emerge from Study 4 was that the *D* measure should be used with an error penalty. The error penalty might be a built-in procedural penalty, as for Measures D_1 or D_2 in Study 6. Alternatively, for IAT procedures that contain no built-in penalty, either of the two penalty formulas used in Study 6 (the 600-ms penalty or the 2 $\times$ standard deviation penalty) should perform approximately equally.

Generalizing to Laboratory Uses of the IAT

The analyses summarized in Tables 2 and 3 used samples that omitted respondents for whom more than 10% of trials had latencies faster than 300 ms. The cut point of 10% fast responses was selected as a compromise among criteria that, in Study 3, were effective in the separate analyses of the four IATs. Over the four data sets, use of the 10%-fast-responses cut point eliminated an average of 1.74% of respondents, which is a smaller percentage of elimination than has been typical of most laboratory IAT studies. Examination of data for respondents who had more than 10% fast responses revealed that their error rates were often high. For example, in the Election 2000 data set, the average error rate for the 1.1% of respondents who exceeded the 10%-fast-responses criterion was 35.7%, compared with an average of only 8.7% errors for the remaining 98.9% of respondents.

The authors were surprised to discover that additional eliminations based on high error rates did not improve results more than slightly beyond what was achieved with the 10%-fast-responses criterion. The minor additional improvement that could be achieved seemed insufficient to justify discarding a relatively large proportion of additional respondents. Study 3 showed that discarding respondents on the basis of slow responding actually impaired performance of the various IAT measures.

The 10%-fast-responses exclusion criterion, which proved most useful in the present studies, may not be sufficient for laboratory studies. In laboratory studies there might be more reason to discard respondents on the basis of high error rates or slow responding. Also, in laboratory studies single aberrant cases may have greater impact than they do in very large data sets such as those of the present research. It therefore seems unwise to use the present results as the basis for a strong recommendation on data-discard policies for laboratory studies. The 10%-fast-responses criterion can be recommended as a minimum exclusion policy for laboratory studies. Laboratory users of the IAT should remain alert in the usual fashion for indications that individual protocols may be untrustworthy.

The authors have begun to use the *D* measure in laboratory investigations in which the conventional algorithm has also been included for comparison. These laboratory uses have most often, but not invariably, indicated larger effect sizes for the *D* measure. These variations in superiority of the *D* measure are consistent with the expected variability of results from small sample investigations. Others will no doubt likewise occasionally encounter samples in which the *D* measure is outperformed when the same data set is analyzed with multiple variations of IAT measures. For their own research, the authors' policy will be to report results for the *D* measure regardless of what has been found with other measures examined for comparison. To do otherwise—for example, by selecting the measure that yields the largest effect size on a test of interest—will inevitably bias effect size estimates.

The Improved Algorithm

The conventional scoring procedure and the improved algorithm that emerges from the present analyses are compared in Table 4. The improved algorithm has three substantial changes from the conventional procedure: (a) use of practice-block data (Step 1 in Table 4), (b) use of error penalties (computed in Steps 5 and 7), and (c) use of individual-respondent standard deviations to provide the measure's scale unit (computed in Step 6 and applied in Step 11).

One way to assess the value of the improved algorithm is to compute the percent savings in research resources that can be obtained due to its expected effect of increasing research power. For these computations, Measures D_2 and C_1 were used to represent the improved algorithm and the conventional algorithm, respectively. Sample sizes required for power of .80 to reject the null

Table 4
Conventional and Improved Implicit Association Test (IAT) Scoring Algorithms Compared

Step	Conventional algorithm	Improved algorithm	Approximately equivalent alternatives for improved algorithm
1	Use data from B4 & B7	Use data from B3, B4, B6, & B7	
2	Nonsystematic elimination of subjects for excessively slow responding and/or high error rates	Eliminate trials with latencies > 10,000 ms; eliminate subjects for whom more than 10% of trials have latency less than 300 ms	
3	Drop first two trials of each block	Use all trials	
4	Recode latencies outside 300/3,000 boundaries to the nearer boundary value	No extreme-value treatment (beyond Step 2)	Delete trials with latencies below 400 ms
5		Compute mean of correct latencies for each block	Also compute *SD* of correct latencies for each block
6		Compute one pooled *SD* for all trials in B3 & B6; another for B4 & B7	Compute these pooled *SD*s just for correct responses
7		Replace each error latency with block mean (computed in Step 5) + 600 ms	Replacement = block mean + 2 × block *SD* computed in Step 5; alternately, use latency to correct response in a procedure that requires a correct response after an error
8	Log-transform the resulting values	No transformation	
9	Average the resulting values for each of the two blocks	Average the resulting values for each of the four blocks	
10	Compute the difference: B7 − B4	Compute two differences: B6 − B3 and B7 − B4	Differences can be computed in the opposite direction
11		Divide each difference by its associated pooled-trials *SD* from Step 6	
12		Average the two quotients from Step 11	

Note. Block numbers (e.g., B1) refer to the procedure sequence shown in Table 1. The conventional algorithm has no procedures corresponding to Steps 5–7 or Steps 11–12 of the improved algorithm. *SD* = standard deviation. SPSS syntax for computing IAT measures using the improved algorithm can be obtained at http://faculty.washington.edu/agg/iat_materials.htm

hypothesis with two-tailed $\alpha = .05$ were computed for research designed to determine statistical significance of an implicit–explicit correlation. On the basis of the average correlations reported for Measures C_1 and D_2 in the first data row of Table 3, the effect sizes used for these power computations were $r = .347$ for the conventional algorithm and $r = .436$ for the improved algorithm. Cohen's (1977, p. 458) Formula 10.3.5 was used to compute required sample sizes.[9] These computations yielded required sample sizes of 63 for the conventional algorithm and 39 for the improved algorithm. The reduction in required sample size afforded by the improved algorithm is therefore 38.1%. This amount of savings can be very significant in research with high per-respondent costs—for example, studies that use individual-subject interviews or studies of difficult-to-locate populations. The savings would be larger in a study with lower expected correlations (e.g., it would be 62.1% using the estimates from the Age IAT as shown in Table 2, Section D).

In addition to the cost savings just illustrated, the improved algorithm offers a gain in construct purity. That is, the improved algorithm, compared with the conventional scoring procedure, is less contaminated by extraneous variables. One such contaminant is the conventional IAT measure's production of spuriously extreme IAT scores for slow responders (see Figure 2 and summary data for Criterion 2 in Table 3, Measures C_1–C_4). The new algorithm almost completely eliminates this artifact (Table 3, Criterion 2, Measures D_1–D_6). Resistance to the response-speed artifact should be useful in studies that compare IAT scores for groups, such as children versus adults, that differ in speed of responding. The new algorithm likewise should provide more valid correlations of IAT measures with individual difference measures, such as

[9] When doing this computation, Cohen's (1977) Formula 10.3.3 should be corrected to read: $z' = \text{arctanh}\ (r)$.

age or working memory capacity, that correlate with response speed. A second artifact for which the new algorithm affords some protection is prior IAT experience. Completion of one or more IATs tends to reduce magnitudes of subsequent IAT scores (see Table 3, data for Criterion 5). The new algorithm's reduced sensitivity to prior IAT experience should be useful in pretest–posttest designs or in studies with multiple IAT measures. Unfortunately, the effect of prior experience is not completely eliminated by the new algorithm (see Table 3, Criterion 5, Measures D_1–D_6). It therefore remains appropriate, when using the new algorithm, (a) to be cautious in interpreting pretest–posttest differences and (b) to counterbalance order of administration for multiple IAT measures.

The benefits of the new algorithm are not limited to the few situations just illustrated. Compared with the previous conventional procedure, the new IAT algorithm should generally (a) better reflect underlying association strengths, (b) more powerfully assess relations between association strengths and other variables of interest, (c) provide increased power to observe the effect of experimental manipulations on association strengths, and (d) better reveal individual differences that are due to association strengths rather than other variables. Accordingly, the new IAT-scoring algorithm can be recommended as a general replacement for the previous conventional procedure.

References

Asendorpf, J. B., Banse, R. & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83,* 380–393.

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology, 81,* 789–799.

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.

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.

Brinley, J. F. (1965). Cognitive sets, speed and accuracy of performance in the elderly. In A. T. Welford & J. E. Birren (Eds.), *Behavior, aging and the nervous system* (pp. 114–149). Springfield, IL: Charles C Thomas.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology, 83,* 1441–1455.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777–799.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology, 49,* 171–180.

Greenwald, A. G. (2001, October). *Top 10 list of things wrong with the IAT.* Presentation at the annual conference of the Society of Experimental Social Psychology, Spokane, WA.

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., 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. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48,* 85–93.

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging, 17,* 482–495.

Kim, D.-Y., & Greenwald, A. G. (1998, May). *Voluntary controllability of implicit cognition: Can implicit attitudes be faked?* Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Lappin, J. D., & Disch, K. (1972). Latency operating characteristic: I. Effects of stimulus probability on choice reaction time. *Journal of Experimental Psychology, 92,* 419–427.

Maison, D., Greenwald, A. G., & Bruin, R. (2001). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin, 2,* 61–79.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123,* 34–80.

Nosek, B. A. (2003). [Implicit and explicit attitudes toward sexual orientation]. Unpublished data.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics, 6,* 101–115.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83,* 44–59.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2003). *Understanding and using the Implicit Association Test: II.* Unpublished manuscript, University of Virginia.

Ratcliff, R. (1993). Methods of dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin and Review, 7,* 1–25.

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

Rudman, L. A., Feinberg, J., & Fairchild, K. (2002). Minority members' implicit attitudes: Automatic ingroup bias as a function of group status. *Social Cognition, 20,* 294–320.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology* (Rev. ed.). New York: Holt.

Yellott, J. I. (1971). Correction for fast guessing and the speed–accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology, 8,* 159–199.

(*Appendix follows*)

Appendix

Questions Used to Obtain Optional Self-Report Measures Prior to Implicit Association Test (IAT) Measures

Likert Items

One 5-point Likert item was used in conjunction with each IAT, illustrated here for the Age IAT:

> Which statement best describes you?
>
> I strongly prefer *young people* to *old people.*
>
> I moderately prefer *young people* to *old people.*
>
> I like *young people* and *old people* equally.
>
> I moderately prefer *old people* to *young people.*
>
> I strongly prefer *old people* to *young people.*

For the Race IAT, the italicized concept words were replaced with *European Americans* and *African Americans.* For the Election 2000 IAT the concepts were *George W. Bush* and *Al Gore.*

For the Gender–Science IAT, the Likert item was as follows:

> Which statement best describes you?
>
> I strongly associate *liberal arts* with *females* and *science* with *males.*
>
> I moderately associate *liberal arts* with *females* and *science* with *males.*
>
> I associate *males* and *females* with *science* and *liberal arts* equally.
>
> I moderately associate *science* with *females* and *liberal arts* with *males.*
>
> I strongly associate *science* with *females* and *liberal arts* with *males.*

Thermometer Items

Two 11-point items were used in conjunction with each IAT, illustrated here for the Age IAT:

> Please rate how warm or cold you feel toward the following groups (0 = coldest feelings, 5 = neutral, 10 = warmest feelings).
>
> Old people
>
> Young people

A drop-down list with numbers 0–10 was provided to the right of each of the two concepts. The thermometer score was computed as the numerical difference between the two responses. For the race and Election 2000 IATs, the concept labels were replaced in the same fashion as for the Likert items.

For the Gender–Science IAT, the thermometer measure was as follows:

> Please rate how much you associate the following domains with males or females.
>
> Science
>
> Liberal arts

The drop-down list to the right of each of the two concepts provided five options: *strongly male, somewhat male, neither male or female, somewhat female,* and *strongly female.* Scoring these five options, respectively, as 1–5, the thermometer score was computed as the numerical difference between the two responses.

Received November 7, 2002
Revision received March 9, 2003
Accepted March 25, 2003 ■