Understanding and Using the Brief Implicit Association Test: I. Recommended Scoring Procedures

Brian A. Nosek
University of Virginia

Yoav Bar-Anan
Ben Gurion University of the Negev

N. Sriram
University of Virginia

Anthony G. Greenwald
University of Washington

Authors' note: This project was supported by a gift from Project Implicit. Nosek and Greenwald are officers of Project Implicit, Inc., a non-profit organization that includes in its mission "To develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors."

Abstract

Sriram and Greenwald (2009) introduced a Brief version of the Implicit Association Test (BIAT). The present research identified analytical best practices for overall psychometric performance of the BIAT. In 7 studies and multiple replications, we investigated analytic practices with several evaluation criteria: sensitivity to detecting known effects and group differences, internal consistency, relations with implicit measures of the same topic, relations with explicit measures of the same topic and other criterion variables, and resistance to an extraneous influence of average response time. Two data transformation algorithms, *G* and *D*, outperformed other approaches. We conclude with recommended analytic practices for standard use of the BIAT.

Abstract = 104 words

Keywords: Brief Implicit Association Test, Non-parametric measures, Response Latency, Implicit Cognition, Test Validity

Understanding and Using the Brief Implicit Association Test: I. Recommended Scoring Procedures

Even the most brilliant research ideas can flounder if data collection procedures and data analytic strategies applied in the pursuit of these ideas are suboptimal. Research efficiency, the knowledge gained in proportion to resources invested, can be improved by maximizing the quality of procedural and analytical methods. Numerous paradigms in mental chronometry, such as Stroop and lexical-decision tasks, define constructs derived from contrasting response latencies between performance conditions (Meyer, Osman, Irwin, & Yantis, 1988). The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is a chronometric procedure that quantifies strength of conceptual associations by contrasting latencies across conditions (Nosek & Sriram, 2007). Participants categorize stimuli representing four categories (e.g., Democrats, Republicans, good words, bad words) in two different conditions – (a) categorizing Democrats and good words together with one response key, and Republicans and bad words together with another response key; and (b), categorizing Republicans and good words together with one response key, and Democrats and bad words with the other. The difference in average response latency between conditions is taken as an indicator of differential association strengths among the concepts. Since its introduction, the IAT has gained in acceptance and influence, and implicit measures generally have had a wide impact on behavioral research (Gawronski & Payne, 2010; Nosek, Hawkins, & Frazier, 2011).

Even when the procedures in a given mental chronometric paradigm are defined unambiguously, there may be various methods to derive scores quantifying the construct of interest. Different scoring practices may lead to unique findings, and across articles it may be difficult to identify scoring procedures as responsible for producing distinct effects. Also, in the absence of standard analytic procedures, researchers may drift into exploratory search and inflate false positives by converging on that scoring strategy that reveals an effect most consistent with the hypothesis. Therefore, standards regarding the scoring procedures contribute to the integrity of research. Ideally, a standard method maximizes reliability and validity of the resulting scores and findings.

Originally, the IAT score, like many other chronometric constructs, was defined as the mean latency (log latency) difference between conditions. Subsequently, Greenwald, Nosek, and Banaji (2003) evaluated candidate scoring procedures for the IAT on numerous criteria. Compared to the original IAT score, the recommended IAT *D* score improved the sensitivity and power of the measure (e.g., a 38% decrease in needed sample size to detect the average correlation; Greenwald et al., 2003, p. 214). The present article similarly evaluates candidate scoring procedures for the Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009). The goal of this investigation is to determine optimal data analytic strategies for deriving association scores from the BIAT.

The Brief Implicit Association Test

The BIAT was developed to shorten the time required to measure associations, while retaining some of the valuable design properties of the IAT. The BIAT can use as few as two response blocks of 20 trials each, which can be completed in a little over a minute. The design that we evaluate here is a sequence of four response blocks of 20 trials each that is preceded by a 16-trial warm-up block (see Table 1).

The categories and exemplars used in the BIAT and the mapping of category exemplars to response keys are the same as those used in the combined blocks of the IAT (see Nosek, Greenwald, & Banaji, 2007 for a full description of the standard IAT procedure). Both procedures use items from four categories (e.g., Democrats, Republicans, good words, bad words) and, within a block, each item is mapped to one of two responses. Whereas in the IAT each category is explicitly associated with one of the two response options, in the BIAT, participants focus on just two of the four categories. Items from these two *focal* categories are categorized with one response key (e.g., the “i” key), and any other items that appear on the screen (*non-focal*) are categorized with the other response key (e.g., the "e" key; see Figure 1). In the two blocks, the focal attribute is kept fixed (e.g., *Good* in the case of attitude) and the two contrasted concepts (Democrats, Republicans) are focal in separate blocks. These design changes simplify instructions and decrease the need for practice shortening total administration time.

Sriram and Greenwald (2009) established the ability of the BIAT to function effectively as a measure of attitudes, identities, and stereotypes. Also, the BIAT has found application in many studies already (e.g., Greenwald et al., 2009; Petróczi et al., 2010; Rüsch et al., 2010; Sheets, Domke & Greenwald, 2011).

Evaluation Criteria

Sensitivity to known effects – main effects and group differences. All else being equal, better scoring procedures should be more sensitive to the measured construct. Comparing among scoring methods of the same data, eliciting a larger overall effect magnitude was considered a desirable criterion. Two of the three topics for the present study – racial attitudes and self-esteem – were appropriate for this criterion. Both elicit strong effects favoring whites to blacks and self to others respectively (Nosek, Banaji, & Greenwald, 2002; Nosek, Smyth, et al., 2007; Yamaguchi et al., 2007).

The third topic – political attitudes – is polarized with liberals or Democrats preferring Democrats and conservatives or Republicans preferring Republicans, even implicitly (e.g., Lindner & Nosek, 2009; Smith, Ratliff, & Nosek, in press). The better scoring algorithms will be more sensitive to detect that group difference. Likewise, Black and White people have different implicit racial attitudes – each more inclined to show positivity toward their own racial group, though Whites more so – providing another group difference criterion (Nosek, Smyth, et al., 2007).[1]

Internal consistency. Better scoring procedures should maximize the internal consistency of the measurement. Scoring algorithms that elicited higher internal consistencies with the same data were favored over those that elicited lower internal consistencies.

Relations with other implicit measures of the same topic (convergent validity). The studies reported in this article made use of a very large dataset designed as a comparative investigation of many implicit and explicit measures of social cognition (Bar-Anan & Nosek, 2012). This offered an opportunity to examine the correlation of the BIAT with seven other procedures for implicit measurement. Stronger correlations with other implicit measures of the same topic indicated better performance by the BIAT scoring procedures.

Relations with parallel self-report measures and criterion variables. Better scoring procedures will elicit stronger relations with known correlates of a measure than worse scoring procedures. For example, height and weight are distinct, correlated constructs. Assessments that minimize random error in measurement of height and weight will maximize their correlation, getting it closest to the true correlation (assuming that the assessments are not influenced by the same extraneous influence; see also Greenwald et al., 2003). The data collection included self-reported attitudes and other direct measurements of known covariates for each of the topics.

Dual-process and dual-system perspectives in attitudes research presume that implicit and explicit attitudes are distinct constructs – the former measured indirectly with procedures like the BIAT, the latter measured directly with self-report procedures (see Nosek, Hawkins, & Frazier, 2011 for a review). Justification of distinct implicit and explicit constructs requires evidence of divergent validity. Nevertheless, it is well established that these constructs are related (Nosek, 2005). Thus, like height and weight, the best measure of both will maximize their relationship by minimizing random error. Separate

[1] For implicit self-esteem consistent group differences are not observed in the reported literature. For example, while self-reported self-esteem differs between people from Eastern and Western cultures, this difference is not observed implicitly (Yamaguchi et al., 2007).

evidence is required to justify the interpretation of the measures as assessments of distinct constructs (for more in-depth treatment of this topic see Greenwald & Nosek, 2008; Nosek & Smyth, 2007).

Resistance to extraneous influences. Extraneous influences are procedural or other factors that produce variation in measurement that is unrelated to the construct of interest. Two extraneous influences are common for response latency measures: participants' average response time, and the order of the measurement blocks. Participants who respond more slowly on average also tend to show larger effects on many response latency measures, especially when computing difference scores (Sriram, Greenwald, & Nosek, 2010). Also, the order of measurement blocks is a well-known influence on response latency measures like the IAT (Greenwald et al., 1998; Nosek, Greenwald, & Banaji, 2005). It is more desirable to have a scoring procedure that is less sensitive to these influences. Ultimately, in the present studies, the order effect did not serve as a criterion variable because the procedural design effectively eliminated that common extraneous influence (see supplementary information), so we examined only the average speed of responding.

Candidate Data Transformations

Five candidate data transformations were compared: mean difference of average latencies, mean difference of average reciprocals, mean difference of log transformed latencies, *D*, and *G*.

Difference scores of mean untransformed or transformed latencies. The most straightforward method for comparing average response latency between contrasted conditions is to average the response latencies in each condition and subtract one from the other. Many research applications apply a data transformation log or reciprocal (inverse) to the raw latencies prior to averaging. Whether transformed or untransformed, these approaches are vulnerable to intra- and inter-individual biases on difference scores (Sriram et al., 2010). As such, we expected that these would not be among the best performing algorithms.

D algorithm. Greenwald and colleagues (2003) introduced the *D*-algorithm as a substantial improvement for scoring the IAT. *D* is the difference between the average response latencies between contrasted conditions divided by the standard deviation of response latencies across the conditions (distinct from the pooled within-conditions standard deviation). Functionally, it is an individual effect size assessment that is similar to Cohen's *d* except, with the same number of trials per condition, *D* has a theoretical minimum of -2 and maximum of +2 when blocks of the same size are compared (Nosek & Sriram, 2007). With the IAT, *D* reduces the impact of extraneous influences like average response latency (Cai, Sriram, Greenwald, & McFarland, 2004; Mierke & Klauer, 2003), and increases sensitivity to detecting relations with known covariates (Greenwald et al., 2003).

G algorithm. In investigating the properties of *D*, Sriram and colleagues (Sriram, Greenwald, & Nosek, 2010; Sriram, Nosek, & Greenwald, 2012) discovered that it is approximately equivalent to a dominance function (Handcock & Morris, 1999). *G,* or the gaussian rank latency difference, is a non-parametric dominance measure that may offer even better psychometric properties as compared to *D*. G is computed by first deriving the fractional ranks (percentiles) of the subject response latencies in the two response conditions as a combined set (i.e., all the trials, regardless of response conditions). These fractional percentiles are distributed uniformly. Outliers in the raw latency distribution have reduced influence as every observation contributes equally to the mean rank. *G* is the difference between the means of the gaussian rank latencies in the two conditions. Like mean percentile differences, *G* is a scale invariant, non-parametric measure. Specific instructions for calculating *G* appear in Appendix E.

Other Data Treatment Considerations

The data transformation is only one of a variety of possible data treatment decisions for analysis. We evaluated four additional criteria in the order presented below.

Warm-up trials. Sriram and Greenwald (2009) defined the four initial trials of each response block that only presented target concepts as practice trials to learn the performance rules for that block. They deleted these trials prior to calculation of BIAT scores. For short response blocks, this is a relatively large amount of data loss. In Study 1, we evaluate whether the warm-up trials provide added value for construct measurement.

Errors. When a participant presses the wrong key in response to a stimulus item, the task presents a red X and waits for the correct response to be made. In the IAT, Greenwald and colleagues (1998; see also Greenwald et al., 2003) found that error responses contain useful information for measuring the intended construct. For blocks with more difficult response configurations, participants are likely to go slower *and* make more errors than blocks with easier response configurations. As such, incorporating errant responses that are delayed by the need to correct them may have positive benefits for measuring association strengths with the BIAT. In Study 5, we compared the effects of deleting error trials or retaining them "as is" (recording the response latency from the beginning of the trial to the correct response regardless of whether an error was made).

Very fast and very slow response trials. Extreme responses – either very slow or very fast – can indicate inattention to the task performance rules. It is not possible, for example, for humans to process and respond to stimulus items with the BIAT rules faster than about 200ms. Likewise, taking more than 5000ms to make a response is unlikely to occur when the participant is attending to the task. In Study 3, we tested cut-offs for very slow and very fast response latencies, and compared treatments of deleting versus recoding the outliers to the cut-off boundary response latency.

Exclusion criteria for overall task performance. Separate from computing an individual score, researchers typically consider a variety of criteria for excluding all of the data from a given task or participant. For example, if a participant is sufficiently disinterested or unable to adhere to the task instructions his or her performance data may be sufficiently invalid so that its inclusion in analyses impairs criteria of efficiency and validity. For example, some participants may press the keys quickly at random to get through the task as rapidly as possible, paying no attention to accuracy. Identifying and removing such non-cooperative participants can improve the validity of a data set. At the same time, unless there is compelling reason to do so, it is good practice to retain as many participants as possible. Study 4 tested different exclusion rules for error rates, number of very fast response latencies, and number of very slow response latencies.

Data Source

Participants were volunteers from the Project Implicit participant pool (https://implicit.harvard.edu/; see Nosek, 2005 for more information). Participants register an identity and are randomly assigned to a study from the pool each time they visit the site. The participant sample is very diverse, but not representative of any identifiable population.

The present data came from a very large data collection termed "Attitudes 3.0" (Bar-Anan & Nosek, 2012) collected from November 6, 2007 to May 30, 2008. In Attitudes 3.0, each study session was comprised of a random selection of measures from a large pool of possible measures. The total session time was designed to be approximately 15 minutes. Each session administered a small subsample of the available measures. The measures assessed evaluations related to three topics: political attitudes, racial attitudes, and self-esteem (see Bar-Anan & Nosek, 2012 for a full description of the procedure, measures, and constraints on the random selection process). Participants could complete the study multiple times. Each time, the participant would receive a new random selection of measures. In total, there were almost 40,000 sessions. For the present studies, we selected the sessions in which participants completed at least one of the three Brief IATs – political attitudes, racial attitudes, and self-esteem. For each topic, there were more than 2,000 completed BIATs.

Overview of Studies

We conducted 7 studies and multiple replications to evaluate data treatment alternatives for the BIAT. Here, we provide a full report of the studies using the politics data and briefly summarize replication studies with race and self-esteem as target concepts (additional details are available at http://briannosek.com/).

Studies 1 and 2 examined psychometric properties of five data transformations, which are described below. Study 1 demonstrated that retention of four warm-up trials for each response block at least slightly damaged psychometric criteria, on the basis of which these warm-up trials were removed from the data for all remaining studies. Study 2 compared the five transformations in terms of possible contamination by subjects' overall speed of responding, finding that two were clearly superior to the other three. As a consequence, the remaining studies focused on these two superior algorithms. Studies 3–6 focused on the two surviving algorithms. Study 3 considered alternatives for dealing with extreme latency trials – very fast or very slow, finding that both algorithms were only mildly affected by latency tail treatments, but also that both could be slightly improved by reducing the impact of fast or slow outliers. Study 4 considered criteria for considering protocols acceptable and found that including those with more than 10% of trials being very fast (<300 ms) disrupted psychometric properties of the BIAT enough to warrant excluding them. High error rates (>30%) also reduced sensitivity but not substantially. On the basis of Study 4, results for all studies are reported excluding respondents who had more than 10% of trials faster than 300 ms. Study 5 established that retaining error trials was superior to removing them. Study 6 found that the "good-focal" response blocks were considerably more reliable and valid than "bad-focal" response blocks, confirming the result previously reported by Sriram and Greenwald (2009). As a consequence of that finding, results from Studies 1–5 are reported here only for BIATs that used good-focal response blocks. Study 7 showed that data from the first and second of two internal replications of the BIAT procedure were comparable in psychometric properties. The General Discussion summarizes the recommended analytic practices based on this investigation.

Method for Political Attitude Studies

Participants

2,358 study sessions of the Attitudes 3.0 dataset included the politics Brief IAT with at least 4 completed blocks for either good-focal or bad-focal tasks. Average age of the participants was 29.4 (SD = 12.1), 65% were female, and 73.1% were White, 6.5% Black, 3.1% east Asian, 2.2% south Asian, 6.7% multiracial, and the remainder other or unknown.

Measures

BIAT. In the Brief IAT, two categories (e.g., *Democrats* and *good words*) are "focal". Stimulus items appear one at a time in the middle of the screen and participants must categorize the stimulus items as either belonging to one of the focal categories (press the 'i' key) or not (press the 'e' key). If the participant makes an error, a red "X" appears below the stimulus and the trial continues until the correct key is pressed. In this study, the stimulus items that appeared but did not belong to the focal categories were always the contrasting stimuli for the other tasks (e.g., *Republicans* and *bad words* when *Democrats* and *good words* were the focal categories).

To evaluate both good and bad-focal conditions, the Brief IAT sequence included nine blocks of trials (Table 1). In each block, the first four trials were selected from the target categories (e.g., *Democrats, Republicans*). The remaining trials for each block alternated between target categories and attributes (good, bad words). The first block was a practice round of 16 total trials with *mammals* and *good words* as the focal categories and *birds* and *bad words* as non-focal categories. The other eight blocks had the four category-only warm-up trials, and then 16 category-attribute alternating trials. The 2nd through 5th blocks had the same focal attribute (e.g., *good words*) and alternated the focal category

(*Democrats, Republicans*) such that one appeared in blocks 2 and 4, and the other appeared in blocks 3 and 5. The 6th through 9th blocks had the other attribute focal (*bad words*) and likewise alternated the focal category between blocks. The order of attributes and categories as focal was randomized between subjects resulting in four between-subjects conditions (*good* or *bad* first; *Democrats* or *Republicans* first) for each topic (politics, race, self-esteem).

Sriram and Greenwald (2009) observed that the Brief IAT showed stronger construct validity for response blocks when *good* was focal compared to when *bad* was focal. This experimental design enabled comparison of good and bad blocks for replication of this effect with a variety of evaluation criteria. Study 6 strongly confirmed Sriram and Greenwald's observation that good-focal blocks outperformed bad-focal blocks.

Other implicit measures. In addition to Brief IATs measuring the three topics of interest, participants were randomly assigned to complete one or more of seven other implicit measures about the same topics – the IAT, Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), single-target IAT (ST-IAT; Karpinski & Steinman, 2006), Sorting Paired Features (Bar-Anan, Nosek, & Vianello, 2009), Evaluative Priming (EPT; Fazio et al., 1995), Affect Misattribution Procedure (AMP; Payne et al., 2005), and – a direct measure with time pressure – speeded self-report (SPD; Ranganath, Smith, & Nosek, 2008). A full report of the procedural details of each implicit measure appears in Bar-Anan and Nosek (2012).

Self-reported attitudes and individual difference measures. Each participant received a random selection of self-report measures including (a) 7-point relative preference for Democrats compared to Republicans, White people compared to Black people, and Self compared to Other; (b) 11-point warmth ratings for each of those target concepts in dependently; (c) liking rating of 5 exemplars of Democrats, Republicans, black people or white people, averaged within topic for analysis (Range = 0 to 8); (d) 14-item Right-Wing Authoritarianism scale (Altemeyer, 1996; range = 1 to 6); (e) Modern Racism Scale (McConahay, 1986; range = 1 to 6); (f) Rosenberg Self-Esteem (Rosenberg, 1965; Range = 1 to 6); (g) self-attributes questionnaire (Pelham & Swann, 1989; Range = 1 to 7); (h) reported 2004 U.S. presidential vote (Kerry or Bush) and 2008 U.S. president voting intention (Democratic or Republican); (i) reported frequency of friendly contact with black people (Range = 1 to 6); and (j) reported recency of positive and negative feedback (Range = 1 to 6; see Bar-Anan & Nosek, 2012 for comprehensive detail on the measures). All variables were coded so that positive correlations would indicate a relationship with the BIAT in the predicted direction.[2]

Demographics. During registration, participants reported a variety of demographic characteristics. Two of those were relevant for the present study: race (categorical identification including Black/African American and White/Caucasian), and political ideology (7-point scale from strongly conservative to strongly liberal).

Procedure

Participants registered for the research participant pool at Project Implicit and completed a demographics questionnaire. On each subsequent visit to the site, participants were randomly assigned to studies from those presently available in the pool. Participants randomly assigned to this study were given a random selection of implicit and self-report measures that required a total time of about 15 minutes to complete.

[2] In the original study design, recency of positive feedback was predicted to have a positive relation to implicit self-esteem – more recent explicitly reported positive feedback predicting higher implicit self-esteem. The empirical result was a weak relationship in the opposite direction. For the present analyses, we followed the empirical result for evaluation of candidate algorithms.

Data Preparation

Study sessions with at least one completed BIAT were retained. Response trials greater than 10,000 milliseconds indicate inattention to the task and were removed (456 of the total of 379,800 trials were removed, with removals disproportionately from the block-warm-up trials, which were 20% of trials and 48% of removals).

Study 1: Including vs. Removing Warm-up Trials

Sriram and Greenwald (2009) removed the first four trials of each BIAT response block presuming that the shortened overall format of the procedure would make those trials particularly unreliable and vulnerable to irrelevant influences. However, with each block being just 20 total trials, removal of the first 4 trials is a substantial 20% reduction of the available data. Study 1 tested whether the initial trials could contribute to the measure's validity by comparing the performance of the BIAT with and without the first four trials. Like most other analyses presented in this article, it used data from which trials with latencies greater than 10,000 ms had been filtered and excluded subjects whose non-cooperation with instructions was indicated by their having more than 10% of latencies faster than 300 ms.

Results and Discussion

Analyses are summarized in text only for good-focal blocks. Similar findings were observed for analyses of data from bad-focal blocks (Table 2). The findings showed that the warmup trials provided no useful data. Across the five candidate data transformations, removing the first four trials left sensitivity in the BIAT to differences across political ideology unchanged (average *r*s = .492 and .490 for warmup trials retained and discarded respectively). Also, the internal consistency of the BIAT was slightly higher without the first four trials (average αs = .548 and .559). BIAT correlations with other implicit measures were unaffected by removing the first four trials (average *r*s = .545 and .549). Finally, BIAT correlations with parallel self-report measures and criterion variables were not different with and without the first four trials (average *r*s = .567 and .565). Similar (non)findings were obtained in with racial attitude measures (Appendix A), and with self-esteem measures (Appendix B).

Performance on the several evaluation criteria varied substantially across the five candidate transformations. For example, the correlation with political ideology with warmup trials removed ranged from .213 (reciprocal) to .563 (*D* and *G*). Internal consistency ranged from .340 (reciprocal) to .779 (*G*). Average correlations with other implicit measures ranged from .464 (reciprocal) to .608 (*D*). And average correlations with parallel self-report measures ranged from .213 (reciprocal) to .681 (*G*). In general, both *D* and *G* were similar and superior in these psychometric criteria to the other three measures. Logarithm was consistently close to these and reciprocal generally last. The poor performance of the reciprocal measure was almost certainly due to the weight it accords to fast responses.[3]

The greatest effect of removing the four warmup trials on any of the psychometric criteria was a slight increase in internal consistency, indicating that the initial four trials of each response block did not

[3] To understand this problem of reciprocals of latency measures, consider that the average of reciprocals of 200 and 500 ms (assume 1000 ms in numerator) is 3.5, which is the reciprocal of 286 ms, 64 ms faster than the latency average, clearly giving greater weight to faster responses. In contrast, the average of reciprocals of 500 and 800 ms (also a 300-ms difference) is 1.625, which is the reciprocal of 615 ms, only 35 faster than the average latency. Various strategies for reducing the contribution of latencies faster than 400 ms substantially improved performance of the reciprocal measure. For example, the reciprocal's correlation with political ideology sharply improved from .213 to .669 when trials with latencies faster than 400 ms were dropped from the data set.

contribute positively to reliability and validity. Removing them is therefore a sensible analytic practice. Data analyses for subsequent studies reported here therefore also removed the four warmup trials.

Study 2: Sensitivity to Average Speed of Responding

Evaluations of the sensitivity of the five potential data transformation procedures to respondent differences in average latency of responding used the method of constructing *latency operating characteristics* similar to those reported in Figures 1 and 2 of Greenwald et al. (2003). For this purpose each subject's average latency in milliseconds was used, excluding from that computation the latencies for the four warmup trials and latencies slower than 10s. Very similar findings are obtained when overall average reciprocal or overall average log latency are used as the indicator of average speed of responding.

Results and Discussion

On the basis of Study 1, we removed data for the first four trials of each response block, in addition to latencies slower than 10,000 ms and excluding subjects with more than 10% of responses faster than 300 ms.

Sensitivity to sample central tendency. The population sampled for this research was known to be politically liberal. On a scale ranging from –3 (strongly conservative) to 3 (strongly liberal), the sample mean was 0.93 (N = 2,232, SD = 1.64; for difference from 0, t_{2231} = 25.76, p = 10^{-137}). It was therefore expected that means for the political BIAT should be numerically positive, reflecting the ideologically liberal preference in the sample. Figure 2 presents a latency operating characteristic for mean values of each transformation, simultaneously displaying differences among the transformations in magnitudes of effect sizes for the mean (higher is better) and in stability of the mean across variations in subjects' overall speed of responding. To enable comparison among the transformations, each decile's mean for each transformation was converted to a Cohen's *d* by dividing it by the transformation's *SD* for the full sample. The figure plots the mean of each transformation for each of 10 latency deciles (overall N = 2,023, Ns per decile range from 202 to 203). The desired characteristic of these latency operating characteristics is their *stability* across the ten deciles.[4]

Figure 2 shows that all five transformations were sensitive to the politically liberal character of the sample. Nevertheless, they varied considerably both in sensitivity and in stability across the 10 deciles. The most obvious deviation from stability in Figure 2 is for the untransformed latency difference measure, which was clearly larger in value for slow than fast deciles. This was true to a lesser extent for the log transformation. The *D* and *G* transformations both showed the opposite trend, being smaller for the slowest subjects. To assess stability statistically, the five transformations were entered as criteria in separate multiple regression analyses that used linear, quadratic, cubic, and quartic trends of the average latency measure for each subject as predictors. Stability is revealed by a *small* size of the Multiple *R* in this analysis. Ordered from greatest to least stability, the five transformations were *Reciprocal* (R = .034, p = .69), *G* (R = .047, p = .35), *Log* (R = .052, p = .24), *D* (R = .076, p = .02), and *Latency* (R = .154, p =10^{-9}).

To show the influence of fast responding on the five transformations, Figure 3 shows the same latency operating characteristics as Figure 2, but for measures in which, additionally, latencies faster than 400 ms were deleted before computing the measure. The patterns are partly the same. The

[4] Conceivably, true values of attitudes are correlated with a third variable that is associated with response latency, which might mean that stability of the LOC should not be expected. Age is such a possible third variable. However, in this sample age was only very weakly and nonsignificantly correlated with the single-item conservatism–liberalism measure (r = .021, p = .36, 2-tailed, N = 1,947), rendering it implausible as a source of non-stability of latency operating characteristics.

latency and logarithm transformations still show greater values for slower subjects, and the *D* and G measures still show smaller values for slower subjects. The most dramatic difference is for the reciprocal transformation, which has values nearly double those in Figure 2. Results for the polynomial regressions were very similar to those for the data in Figure 2.

Sensitivity to correlation with political orientation. Latency operating characteristics parallel to those in Figures 2 and 3 were also computed using as criterion measure the correlation between each BIAT transformation and the 1-item measure of conservative–liberal political orientation. The results led to conclusions similar to those for the mean values of the transformations shown in Figures 2 and 3. For analyses based on all trials other than the four warmup trials and those slower than 10s, the reciprocal measure performed least well, the untransformed latency measure somewhat better, and the other three measures best, with both *G* and *D* slightly superior to the logarithm transformation. When latencies faster than 400 ms were deleted prior to computing the measures, the measures were much more similar, but still the untransformed latency measure showed weaker correlations across the range of latencies.

Other criteria. Summary results across the other evaluation criteria with latencies faster than 400 ms deleted are summarized in the right-side panel of Table 2 for both good-focal and bad-focal response blocks (see Appendix A for results with race and Appendix B for results with self-esteem). Across criteria, *G* and *D* performed consistently strongly, with *Reciprocal* also performing well in many cases, and *Log* and particularly *Latency* performing less well, albeit only slightly in some cases for *Log*. Overall performance of the transformations coincided with the strength of relationship with *G* (Table 3) – *D* was most strongly related, followed closely by *Reciprocal*, then *Log*, and then *Latency*.

As a summary of Study 2, *G* and *D* were nearly indistinguishable in performing best among the five transformations, with more comparisons slightly favoring *G* than slightly favoring *D*. Without truncation of fast responses *Reciprocal* showed the least sensitivity to expected effects, and *Latency* showed greatest susceptibility to artifact associated with speed of responding and did not perform as well on the other criteria. *Log* was satisfactory in many respects, but was nevertheless consistently outperformed by both *G* and *D*.

Study 3: Treatment of Extreme Latencies

Because the preceding studies had made clear that G and *D* were the most effective measures, starting with Study 3 we changed focus to comparing those two measures and trying to find their best forms. Although the reciprocal transformation was consistently third best in analyses when responses faster than 400 ms were removed, it was not considered further because its properties were quite poor without those removals (see, e.g., Figure 2).

In speeded response tasks, very rapid and very slow responses are often treated as due to subjects deviating from instructed behavior. Study 3 examined alternative methods for reducing the influence of these outlying observations on psychometric properties of the *G* and *D* measures. Each latency-tail treatment was identified by boundary values for fast and slow responding. For each candidate boundary value, we examined effects either of removing trials outside that boundary or of recoding outlying trials to the boundary values, or both. The boundaries and strategies that were examined were: no removal, deleting below 200 or 400 ms boundaries, recoding below 400 ms to that boundary value, deleting above 2000, 3000 or 4000 ms, and recoding above 2000, 3000, or 4000 ms to those boundary values. Because *G* is a nonparametric measure that is computed from rankings of latencies, extreme score treatments may be less consequential than they are for other measures that preserve distributional information. Although the *D* measure preserves distribution information, it also reduces the impact of outlying observations by using the subject's variability as a denominator for latency differences between treatments. Outlying observations, in effect, reduce their own impact by contributing to the magnitude of the denominator.

Results and Discussion

Analyses were conducted after first deleting the first four trials of each response block and excluding participants who had more than 10% of latencies faster than 300ms. Findings are reported for good-focal BIATs (see Table 4). We also report results for replications with race (Appendix C) and self-esteem (Appendix D).

G performed comparably with no treatment of fast trials and with recoding or removing trials below the 400 and above the 2000 millisecond boundaries. For *G*, the comparable performance of using no treatment at all and recoding at the 400 and 2000 boundaries leads us to prefer the no treatment option because it offers the conceptual advantage of minimizing data manipulation. For *D*, recoding with 400 and 2000 boundaries produced the best performance. Overall, the results suggest that both *D* and *G* are relatively insensitive to treatments of extreme latencies.

Study 4: Error trial treatment

When participants make a categorization error in the BIAT they must correct it before moving on to the next trial. The trial latency is the time from stimulus onset until the correct response is made. Studies 1–3 retained all trials whether or not an error occurred. Alternative analytic strategies are to remove or recode error trials before calculating BIAT scores. On the basis of evidence obtained with the IAT (Greenwald et al., 2003), we expected that error trials would provide useful data and that it would likely therefore be best to retain them in computing the measure.[5]

Results and Discussion

We removed the first four trials of each response block and trials faster than 400ms, excluded participants having more than 10% of trials with response latencies faster than 300ms, and we summarize results for the good-focal blocks (Table 5).

We compared BIAT scores with and without error trials removed for five evaluation criteria for *G* and *D*. Results were very similar across scoring transformations, so for simplicity, the results for *G* only are summarized here. The political attitude BIAT was more sensitive to differences between liberals and conservatives when error trials were retained (r = .565) than when they were removed (r = .537). Also, internal consistency was higher with error trials retained (α = .783) than when removed (α = .731). The BIAT correlated more strongly with each implicit measure with error trials retained (average r = .601) than with error trials removed (average r = .572). Further, the BIAT correlated more strongly with 7 of 8 self-reported attitudes and other criterion variables with error trials retained (average r = .593) than with error trials removed (average r = .570). Removing the error trials resulted in a slightly weaker relationship with the average response latency extraneous influence (r = -.032) than retaining the error trials (r = -.055), but the differences were already near zero. These findings support retaining error trials as useful contributors to the BIAT measure.

Study 5: Analytic Strategy – Respondent Exclusion Criteria

When participants neglect task instructions they may produce data that is relatively useless for measurement. Two available indicators of failure to perform the BIAT as instructed were responding more rapidly than is plausible for intentional, accurate responding and making frequent errors. These are correlated indicators, because subjects who respond too rapidly will also have increased error rates. Study 4 showed that error trials can provide useful data. The study that produced the currently

[5] There are research applications using the IAT in which respondents are not required to correct errors. We do not consider that procedural format in this manuscript. If such procedures are used, Greenwald et al. (2003) should be consulted for appropriate scoring practices.

preferred IAT scoring algorithm (Greenwald et al., 2003) found that subjects who had more than 10% responses faster than 300 ms ("fast" responses) provided generally useless data and were best dropped from analyses. We compared three exclusion criteria based on response speed: no exclusions and exclusions based on exceeding either 10% or 20% of responses faster than 300ms. We also examined three exclusion criteria based on error rates: no exclusion and exclusions based on exceeding either 30% or 40% error rates.

Results and Discussion

Results were similar for the *D* and *G* transformations and are reported here only for *G*. Results were computed using data sets from which the four warmup responses and latencies slower than 10s were initially removed. Results are reported for BIAT measures based on good-focal blocks. The results were similar for measures computed from bad-focal blocks (Table 6).

We examined the exclusion criteria sequentially – first comparing the fast trial exclusion rules, and then comparing the error rate exclusion rules. As was previously found for the IAT, excluding subjects with more than 10% fast responses (3.8% of subjects) produced psychometric benefits superior to either no exclusion or to the 20% criterion which excluded fewer (3.0% of subjects). The 10%-fast-response exclusion criterion produced best psychometric properties for detecting known group differences, for internal consistency, for correlations with parallel self-report attitude and other criterion variables, for correlations with parallel implicit measures, and for freedom from contamination by variations in average latency of responding.

Starting from the base of excluding subjects with more than 10% of "fast" responses, the 40% error criterion eliminated another 1.3% of the sample (5.0% excluded in total), while the 30% error criterion eliminated 4.7% of the sample (8.3% excluded in total). The 30% exclusion criterion afforded greater sensitivity to known group differences and slightly stronger internal consistency compared to no error-based exclusion, but did not improve over no extra error-based exclusion for relations with implicit measures, relations with self-reported attitudes and criterion variables, and freedom from contamination by variations in average latency of responding.

We replicated these results with racial attitudes and self-esteem (see supplementary materials). The results suggest that excluding subjects with more than 10% fast responses has benefits for overall psychometric performance. Because the 10%-fast-responding criterion effectively excludes most subjects who have higher error rates, additional exclusion of remaining subjects with 30% or more errors has only a small additional beneficial effect. Nevertheless, it is possible that the 30%-error-rate exclusion criterion will prove useful in small samples.

Study 6: Comparing Good-focal and Bad-focal Blocks

Sriram and Greenwald (2009) observed that response blocks in which *Good words* was a focal category and *Bad words* was nonfocal showed more reliable effects and stronger correlations with criterion variables than did those in which *Bad words* was focal and *Good words* was nonfocal. This is a curious phenomenon because the two variations are structurally identical. In both cases, there are two response blocks: In the political BIAT (a) in one block *Democrats* and *Good words* are categorized with one key and *Republican* and *Bad words* are categorized with the other key, and (b)in the other block *Republicans* and *Good words* are categorized with one key and *Democrats* and *Bad words* with the other key. The only difference between the good-focal and bad-focal conditions is in the category labels that appear on screen and to which participants are instructed to attend. In the "good-focal" condition, the category labels appear as "Democrats and Good" and "Republicans and Good" naming the two categories required for one of the responses keys in the respective response blocks described above. Respondents are instructed to categorize "anything else" with the second key. In the "bad-focal" condition, the category labels appear as "Republicans and Bad" and "Democrats and Bad" for the same response blocks. Sriram and Greenwald's finding that good-focal and bad-focal conditions elicit different degrees of validity was intriguing and important to clarify. Study 6 sought to examine this phenomenon with a variety of evaluation criteria. The results demonstrate the procedural advantage of using "good" as the focal attribute for attitude BIATs.

Results and Discussion

We applied the same data preparation practices as in Study 2, and likewise compared focal conditions on all five candidate data transformation approaches. We compared good-focal and bad-focal conditions on (a) sensitivity to known group differences, (b) internal consistency, (c) relations with other implicit measures, and (d) relations with parallel self-report measures and criterion variables.

We compared the correlation of the political attitudes BIAT with self-reported political orientation between the two focal conditions. Across the five data transformation procedures, political orientation was more strongly correlated with the good-focal BIAT (average *r* = .534; range among scoring transformations .457 to .565) than with the bad-focal BIAT (average *r* = .312; range .282 to .393; see Table 2). In other words, political ideology accounted for almost 200% more shared variance in the good-focal BIAT (28.2%) than in the bad-focal BIAT (9.7%) despite them being structurally identical. Likewise, the good-focal BIAT (average α = .749; range .690 to .783) showed much stronger internal consistency than did the bad-focal BIAT (average α = .547; range .389 to .603). Further, the good-focal BIAT (average *r* = .567; range .474 to .601) correlated more strongly with seven other implicit measures of political attitudes than did the bad-focal BIAT (average *r* = .387; range .291 to .419). Finally, the good-focal BIAT (average *r* = .561; range .488 to .593) correlated more strongly with eight self-reported criterion variables such as past voting and voting intention than did the bad-focal BIAT (average *r* = .369; range .291 to .398). These differences indicate sizable internal consistency and validity advantages for the good-focal over the bad-focal conditions.

We replicated the comparison of good and bad-focal blocks with racial attitude measures, and with self-esteem measures (see Appendices A and B). The results consistently replicated for racial attitudes, and offer the same conclusion but somewhat less definitively for self-esteem. In particular, the self-esteem BIAT showed weak relations with other implicit measures and with the criterion variables for both good and bad-focal blocks (see also Bosson, Swann, & Pennebaker, 2000). On the other criteria, good-focal retained a clear advantage.[6] These results suggest that attitude BIATs will be much more effective by using *good* as the focal category instead of *bad*. In additional laboratory data, there is a similar advantage for using *self* as the focal category instead of *other* for identity-related IATs (Sriram & Greenwald, 2009). Identification of the mechanism underlying these differences may assist in selecting focal and background categories for other applications. A qualification of this conclusion is the possibility that *bad* and *other* focal blocks reveal distinct validity, even though their psychometric performance is weaker overall. As such, there may be many research applications in which collecting data for both focal conditions is theoretically relevant and advisable.

Study 7 — First 40 vs. Second 40 Trials

In the development of the IAT scoring algorithm, data from the first blocks of each combined task produced a measure slightly superior to that from the second blocks of each task (Greenwald et al. (2003, p. 202). In the present research, each of the good-focal and bad-focal BIATs was conducted in

[6] Another potential contributor is evidence that "self" functions better as a focal category than does "other" – a factor that was manipulated within focal condition in this context. Future research might focus on self-esteem measurement in particular given its peculiarities. Here, we retain an emphasis on the aggregate results across topics.

four blocks, producing one measure for the first two blocks and another for the second two. These two sub-measures provided the basis for previously described internal consistency tests.

For Study 7's comparisons of the two sub-measures we used the best performing versions of the *G* measure and *D* measures — ones computed from data sets for which 4 warmup trials of each block and latencies greater than 10s had been removed, and for which latencies faster than 400 ms and slower than 2000 ms had been recoded to those boundary values. Also, data for subjects having more than 10% of responses faster than 300 ms were excluded. These *G* and *D* measures were compared in their sensitivity to the liberal character of the subject population, and their average correlations with the self-reported and implicit political attitude measures, and also their (weaker) average correlations with three self-report race attitude measures and with seven available implicit race attitude measures.

Results and Discussion

Table 7 compares properties of *D* and *G* measures based on first 40 trials versus second 40 trials of the political BIAT measure. Each set of 40 trials consisted of two 20-trial blocks, one with Democrat and good focal, the other with Republican and good focal. Results for *D* and *G* measures computed without any latency tail treatments are included in Table 7 for comparison.

The most striking feature of the results in Table 7 were that average correlations with political measures were quite substantial for both *D* and *G*, regardless of whether the measures used tail treatments or not and whether they were based on the first 40 or last 40 trials. Although the averaged correlations with self-report and implicit race attitude measures were lower, all of the individual correlations with explicit race attitude measures were statistically significant, and most of those with implicit measures were likewise statistically significant. Additionally, the internal consistency data showed Cronbach's alphas close to .80.

Those observations, however, do not bear on the main two reasons for interest in these data, which were (a) to determine whether there was a difference between the two sets of 40 trials in their sensitivity to expected effects and (b) to compare performance of the *D* and *G* measures.

In fact, the data provide no clear basis either for preferring *G* to *D* or for preferring measures based on the first 40 or the second 40 trials of each BIAT. And, both sets of trials contribute to measurement validity with the first 40 performing best on a few criteria and the second 40 on other criteria. In examining these data in conjunction with replications using the race and self-esteem BIATs, there are indications that *G* measures are quite robust in showing little difference between versions of the measure with and without treatment of tails of latency distributions. Also, the *G* measures consistently had slightly higher internal consistency than the *D* measures. *D* measures consistently showed small benefits of tail treatments. The *D* measures were very slightly superior to the *G* measures when tail treatments were applied. It may require data sets with considerably more observations than even the large data set of the present research to establish the generalizability of these observations.

General Discussion

The present studies identified analytic practices that maximized (a) sensitivity to known effects and group differences, (b) internal consistency, (c) relations with other implicit measures of the same topic, (d) relations with self-report measures of the same topic and other criterion variables, and (e) resistance to the extraneous influence of average response time for the Brief Implicit Association Test. The studies and replications showed that (a) the four warm-up trials at the beginning of each response block do not contribute to the measures' validity (Study 1), (b) *D* and *G* data transformations perform better than variations that use differences between average response times (Study 2), (c) trials in which an error is made provide useful information and should be retained in analysis (Study 3), (d) task performances with a high frequency of unreasonably fast responses and high error rates (to a lesser degree) may be removed to improve overall sensitivity and measure performance (Study 4), (e) treatment of extreme latencies has relatively small effects, particularly with *G*, but can improve *D* slightly by either recoding or removing very fast and very slow trials (Study 5), (f) "good-focal" response blocks possess much stronger psychometric properties than "bad-focal" response blocks (Study 6), and (g) the first and second halves of the BIAT contributed approximately equally to the measures' validity (Study 7). These findings converge to recommended BIAT analytic practices that are presented in Table 8. Future research may identify additional improvements among the variations.

The best performing *G* and *D* algorithms performed similarly well in the present data, though *G*'s performance was less contingent on manipulation of the originating data (i.e., trimming outliers). We present both a *G* and *D* algorithm for investigators that wish to compare performance of them on other datasets. The present studies had the advantage of evaluating the algorithms with known findings, large samples, and multiple replications allowing for inference based on algorithm performance. In typical research applications using the BIAT, especially when the hypothesized outcome is not already known to exist, selection of scoring algorithm should occur prior to data analysis, not following analysis based on which one elicited the best performance in that particular dataset. Applying standard analytic practices will facilitate comparison of effects across research applications.

Implications and future directions

Good primacy. The performance advantage of "good-focal" over "bad-focal" blocks is stunning considering that they are behaviorally identical. Participants are using the same keys, with the same response assignments, and categorizing the same stimuli. The only difference are the instructions for what information to attend to, and the labels that appear on screen. This illustrates the significant impact instructions and the mental context can have on implicit measurement, and it offers an intriguing puzzle that neither Sriram and Greenwald (2009) nor we have the evidence to solve. Unkelbach and colleagues proposed that positive information is more similar to other positive information and is more strongly associated with other positive information, in comparison to the similarity and association that negative information share with other negative information (Unkelbach et. al, 2008). It is conceivable that this could contribute to good primacy in the BIAT.

Potential applicability of G and D to other procedures. *D*, and now *G*, were developed for analysis of response latency data in the contrasting conditions of the IAT and BIAT. However, they have the potential for much broader application. Sriram, Nosek, and Greenwald (2012) propose scale invariance and validity maximization as defining properties of admissible latency contrasts, and perhaps for other measures that have mean-variance correlations (i.e., differences in means between conditions are associated with differences in variances between conditions). Across a range of paradigms including the IAT, evaluative priming, task-switching, visual search, and the Stroop task, the scale invariant, relative *G* measure outperformed the mean latency difference and other absolute measures[7]. *G* is not spuriously related to general processing speed, has greater power in detecting valid effects, and exhibits superior internal consistency. *G* may have the potential to advance the analysis of latency contrasts generally.

Conclusion

Research efficiency – the amount of knowledge gained compared to the resources expended – is improved by maximizing the validity of measurement methods. In the present article, we identified analytic practices that improve the validity of the BIAT. Applying these practices, and adapting further

[7] Relative measures depend on a contrast between 2 conditions. The rank score for trials within a block necessarily depends on the observations in a contrasting block because the rank is computed across trials in both blocks. In contrast, log latency is an absolute measure because the logarithm of a latency within a block does not depend on other observations.

improvements when they are identified, will accelerate the discovery of the relevance of implicit cognition for human behavior.

References

Altemeyer, B. (1996). *The Authoritarian Specter*. Cambridge, MA: Harvard University Press.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, *56*, 329-343.

Bar-Anan, Y. & Nosek, B. A. (2012). A comparative investigation of seven indirect measures of social cognition. Unpublished manuscript.

Bosson, J.K., Swann, W.B., & Pennebaker, J.W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631-643.

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

Fazio, R.H., Jackson, J.R., Dunton, B.C., & Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013-1027.

Gawronski, B., & Payne, B.K. (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. Guilford Press.

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

Greenwald, A. G., & Nosek, B. A. (2008). Attitudinal dissociation: What does it mean? In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the New Implicit Measures* (pp. 65-82). Hillsdale, NJ: Erlbaum.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85(2)*, 197-216.

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

Handcock, M. S., & Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*. Springer-Verlag: New York.

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

Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology, 30,* 67-92.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, Discrimination, and Racism* (pp. 91-125). San Diego, CA: Academic Press.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern Mental Chronometry. *Biological Psychology, 26*, 3-67.

Mierke, J., & Klauer, K.C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*, 1180-1192.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*, 565-584.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19(6)*, 625-666.

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

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31(2)*, 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265-292). New York: Psychology Press.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15*, 152-159.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology*, *54*, 14-29.

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

Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales, and Christie (2006). *Journal of Experimental Social Psychology, 43*, 393-398.

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

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology, 44*, 386-396.

Pelham, B. W., & Swann, W. B. (1989). From self-conceptions to self-worth: On the sources and structure of global self-esteem. *Journal of Personality and Social Psychology, 57*, 672-680.

Petróczi, A., Aidman, E. V., Hussain, I., Deshmukh, N., Nepusz, T., Uvacsek, M., Tóth, M., Barker, J., & Naughton, D. P. (2010) Virtue or pretense? Looking behind self-declared innocence in doping. *PLoS ONE*, *5*, e10457.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rüsch, N., Todd, A. R., Bodenhausen, G. V., Olschewski, M., & Corrigan, P. W. (2010). Automatically activated shame reactions and perceived legitimacy of discrimination: A longitudinal study among people with mental illness. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*, 60–63.

Sheets, P., Domke, D., & Greenwald, A. G. (2011). God and country: The partisan psychology of the presidency, religion, and nation. *Political Psychology*, *32*, 459–484.

Smith, C. T., Ratliff, K. A., & Nosek, B. A. (in press). Rapid assimilation: Automatically integrating new information with existing beliefs. *Social Cognition.*

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

Sriram, N., Greenwald, A. G., & Nosek, B. A. (2010). Correlational biases in mean response latency differences. *Statistical Methodology, 7*, 277-291.

Sriram, N., Nosek, B. A., & Greenwald, A. G. (2011). Scale invariant contrasts of response latency distributions. Unpublished manuscript.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*, 36-49.

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science, 18*, 498–500.

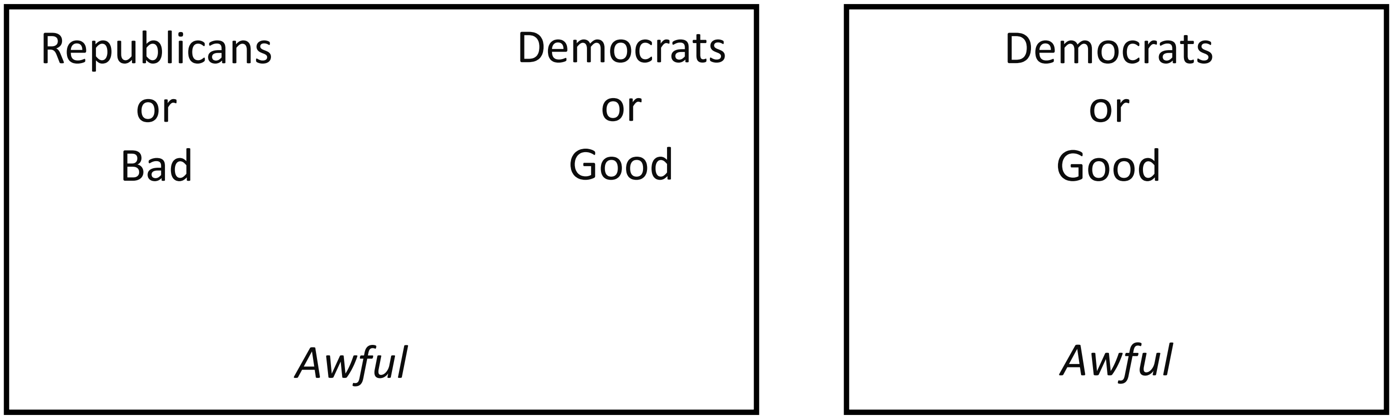

Figure 1. Schematics of the same single response trial of one block of the IAT on the left, and the BIAT on the right.

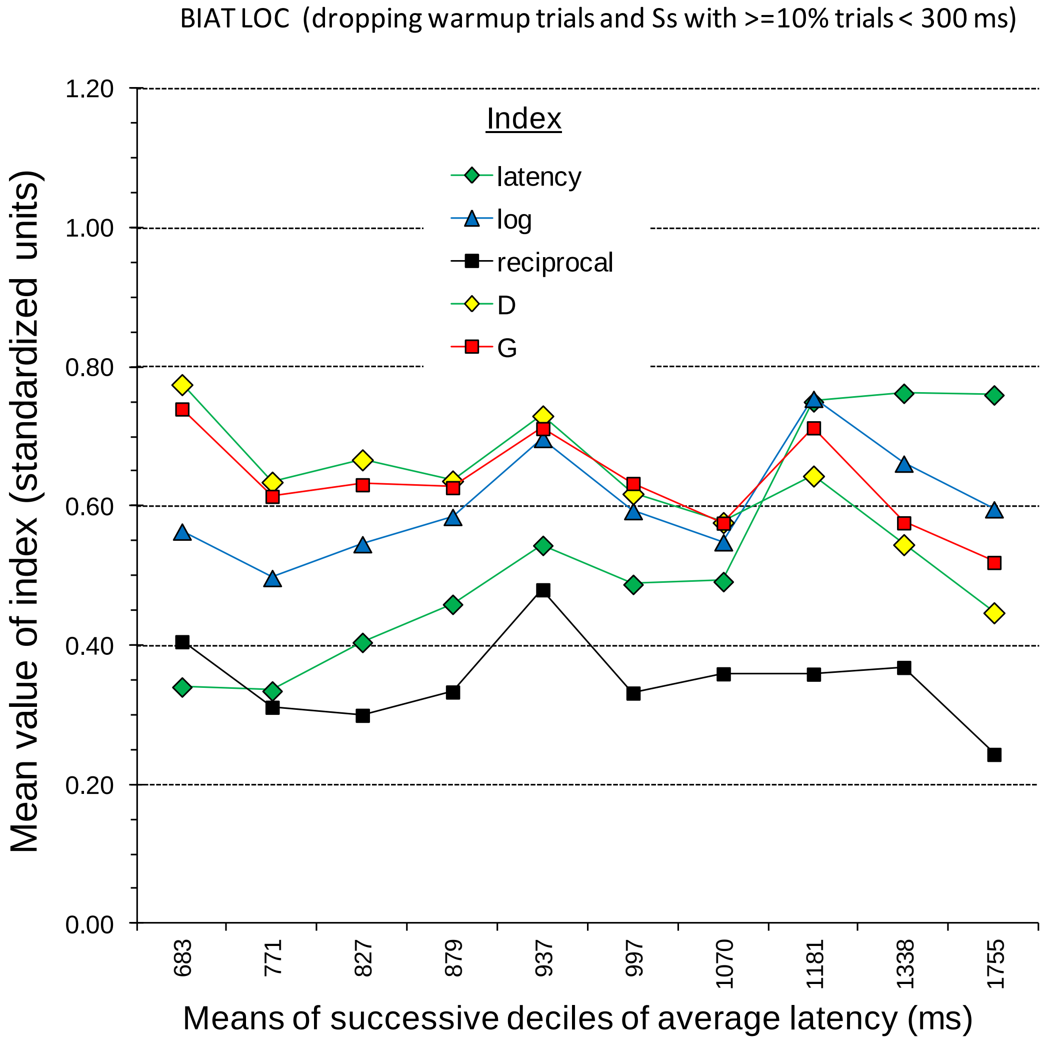

Figure 2. [illegible] BIAT scoring algorithms across deciles of the sample's distribution of average speed of responding. For this plot, the algorithms were computed after deleting 4 warmup trials from each response block and also deleting latencies greater than 10,000 ms. There were 202 or 203 respondents in each decile. Most noticeable in the graph is the inferior performance (smaller effect sizes) for the reciprocal measure, and strongest performance for the *D* and *G* measures. Also noticeable is that the D and G measures were smallest for the slowest subjects, whereas the log and latency measures were largest for the slowest subjects.

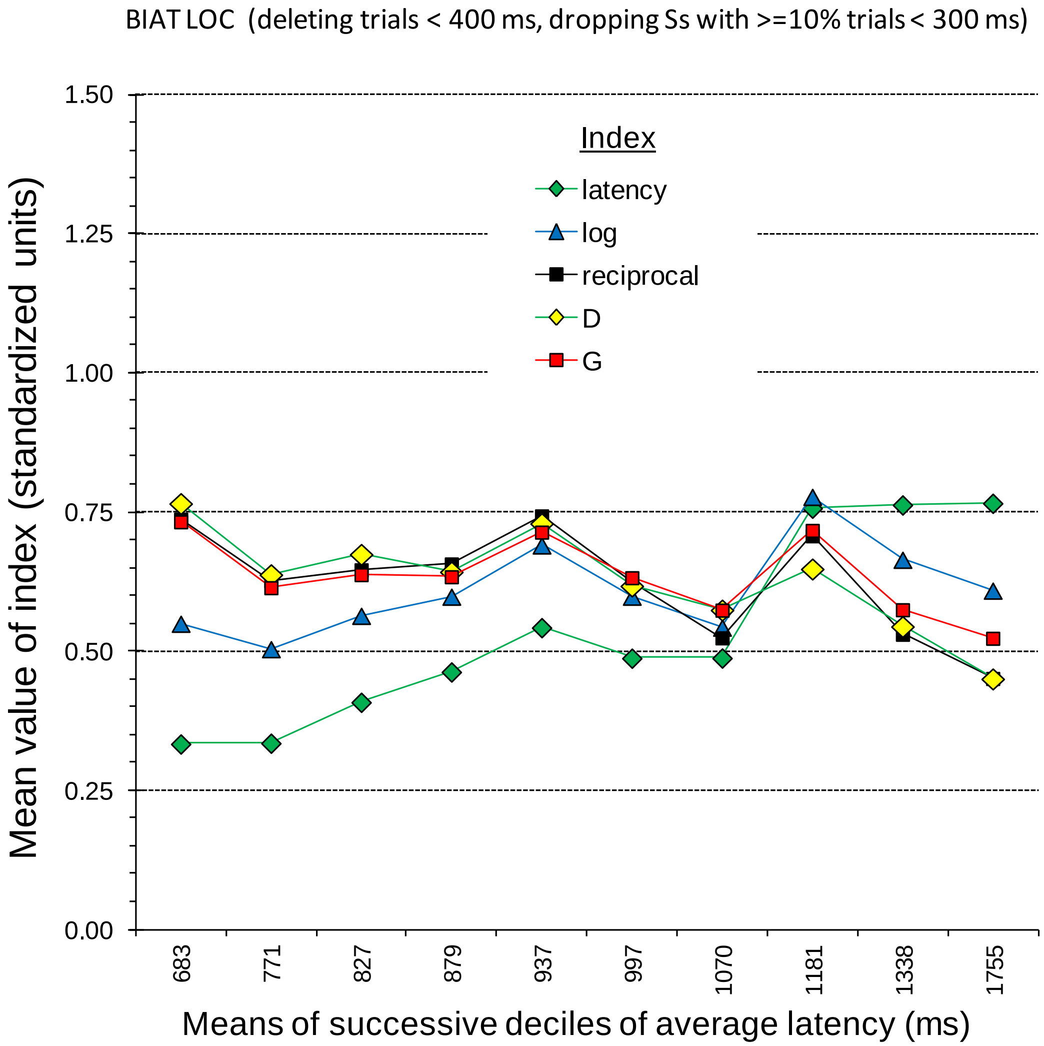

Figure 3. [illegible] BIAT scoring algorithms across deciles of the sample's distribution of average speed of responding. Pretreatment of the data involved removing 4 warmup trials per block, latencies slower than 10s, and latencies faster than 400 ms. There were 202 or 203 respondents in each decile. The most noticeable effect visible in the graph are improvement in performance of the reciprocal measure relative to its poor showing in Figure 2, and the contrast between the relative stability across speed variations for four of the measures and the increasing magnitude of the (untransformed) latency-difference measure as responding went from fast (left of graph) to slow.

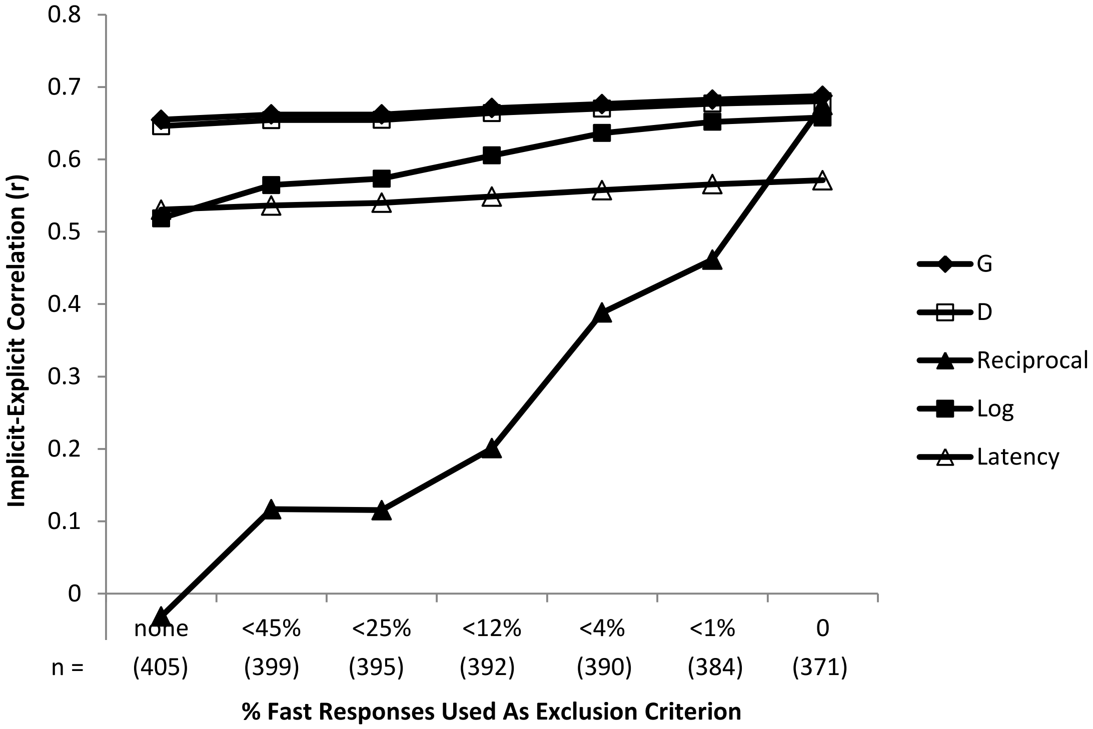

Figure 4. Effects of seven criteria for excluding respondents as a function of their proportion of fast responses (latency < 300 ms) on correlations with self-reported preference between Democrats and Republicans for five BIAT data transformations (Study 4). Higher correlations indicate better performance. The furthest left datapoint indicates no exclusion of participants; the furthest right datapoint indicates exclusion of all participants that had even a single fast response. Sample size (n) on the x-axis indicates the number of participants retained with that exclusion criterion.

Table 1. BIAT procedure

Block	Trials	Trial structure	Example focal	Example non-focal
1	16	4 attribute only + 12 trials alternating category and attribute	Good words (attribute) and mammals (category)	Bad words (attribute) and birds (category)
2	20	4 attribute only + 16 trials alternating category and attribute	Good words (attribute) and Democrats (category)	Bad words (attribute) and Republicans (category)
3	20	4 attribute only + 16 trials alternating category and attribute	Good words (attribute) and Republicans (category)	Bad words (attribute) and Democrats (category)
4	20	4 attribute only + 16 trials alternating category and attribute	Good words (attribute) and Democrats (category)	Bad words (attribute) and Republicans (category)
5	20	4 attribute only + 16 trials alternating category and attribute	Good words (attribute) and Republicans (category)	Bad words (attribute) and Democrats (category)

Notes: Procedure displays for political attitude measure. Between subjects randomize order of blocks 2 and 4 with blocks 3 and 5. A trial begins with the onset of the stimulus and ends once a correct categorization is made. Clarity between the two dimensions (Democrats/Republicans and Good/Bad) was enhanced by presenting the labels and stimulus items from each dimension in a different color or stimulus format (e.g., Democrats/Republicans as images; Good/Bad as words).

Table 2. Comparison of retaining or removing 1st four trials of each block (Study 1), candidate data transformations (Study 2), and bad versus good focal blocks (Study 6) across evaluation criteria for political attitudes. Correlations averaged after Fisher's z-transformation and then

		Retain all trials					Remove 1st four trials of each block					Average across algorithms	
	N	G	D	Reciprocal Diff	Log Diff	Latency Diff	G	D	Reciprocal Diff	Log Diff	Latency Diff	All Trials	Remove 1st
KNOWN GROUP DIFFERENCES													
Political Ideology (Bad Focal)	2026	0.346	0.317	0.353	0.316	0.226	0.389	0.378	0.393	0.365	0.282	0.312	0.362
Political Ideology (Good Focal)	2026	0.564	0.548	0.557	0.533	0.453	0.565	0.557	0.556	0.535	0.457	0.532	0.535
INTERNAL CONSISTENCY													
Alpha (Bad Focal)	2012	0.588	0.506	0.612	0.550	0.399	0.603	0.564	0.623	0.557	0.389	0.535	0.552
Alpha (Good Focal)	2024	0.781	0.750	0.779	0.762	0.672	0.783	0.773	0.778	0.763	0.690	0.751	0.759
RELATIONS WITH OTHER IMPLICIT MEASURES													
BAD FOCAL													
IAT	290	0.434	0.392	0.454	0.392	0.262	0.483	0.472	0.488	0.439	0.315	0.389	0.441
GNAT	251	0.523	0.484	0.521	0.484	0.354	0.571	0.556	0.558	0.533	0.419	0.475	0.529
ST-IAT	244	0.388	0.357	0.391	0.350	0.244	0.401	0.400	0.404	0.371	0.281	0.347	0.372
SPF	271	0.447	0.420	0.463	0.413	0.314	0.474	0.471	0.487	0.447	0.354	0.413	0.448
EPT	262	0.320	0.304	0.342	0.282	0.179	0.353	0.350	0.371	0.318	0.229	0.286	0.325
AMP	361	0.242	0.227	0.248	0.195	0.118	0.273	0.272	0.276	0.228	0.151	0.206	0.241
SPD	417	0.356	0.321	0.361	0.326	0.233	0.380	0.375	0.381	0.359	0.289	0.320	0.357
GOOD FOCAL													
IAT	290	0.629	0.610	0.619	0.586	0.477	0.643	0.645	0.629	0.602	0.505	0.587	0.607
GNAT	251	0.668	0.648	0.664	0.639	0.553	0.674	0.676	0.668	0.652	0.578	0.636	0.651
ST-IAT	244	0.596	0.549	0.595	0.558	0.423	0.603	0.607	0.591	0.561	0.419	0.547	0.560
SPF	271	0.611	0.627	0.612	0.588	0.487	0.606	0.619	0.604	0.584	0.496	0.587	0.583
EPT	262	0.553	0.535	0.549	0.506	0.422	0.560	0.561	0.555	0.516	0.427	0.515	0.526
AMP	361	0.509	0.481	0.505	0.475	0.393	0.499	0.484	0.497	0.463	0.365	0.474	0.463
SPD	417	0.627	0.588	0.609	0.587	0.511	0.622	0.617	0.601	0.588	0.527	0.586	0.592
Bad focal average		0.390	0.360	0.400	0.352	0.245	0.424	0.418	0.428	0.389	0.293	0.351	0.391
Good focal average		0.601	0.579	0.595	0.565	0.468	0.604	0.604	0.594	0.569	0.477	0.564	0.571
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES													
BAD FOCAL													
Dem-Rep Preference	396	0.393	0.358	0.390	0.362	0.264	0.439	0.430	0.427	0.410	0.316	0.354	0.405
Warmth for Democrats	401	0.321	0.277	0.320	0.315	0.256	0.358	0.341	0.343	0.345	0.289	0.298	0.335
Warmth for Republicans	398	0.364	0.339	0.363	0.324	0.223	0.399	0.401	0.394	0.375	0.293	0.323	0.373
Right-Wing Authoritarianism	438	0.442	0.411	0.455	0.411	0.281	0.468	0.451	0.480	0.456	0.364	0.402	0.445
Avg liking of 5 Democrats	335	0.306	0.238	0.325	0.253	0.137	0.355	0.327	0.369	0.318	0.219	0.253	0.319
Avg liking of 5 Republicans	335	0.335	0.300	0.363	0.323	0.233	0.379	0.360	0.397	0.374	0.305	0.311	0.363
Intended Vote in 2008 (D or R cand.)	375	0.296	0.265	0.291	0.261	0.177	0.334	0.321	0.324	0.303	0.221	0.258	0.301
Vote in 2004 (Bush or Kerry)	216	0.404	0.364	0.400	0.359	0.247	0.453	0.450	0.444	0.423	0.323	0.356	0.420
GOOD FOCAL													
Dem-Rep Preference	396	0.682	0.659	0.669	0.642	0.544	0.680	0.672	0.667	0.646	0.556	0.642	0.646
Warmth for Democrats	401	0.471	0.441	0.463	0.451	0.390	0.465	0.452	0.456	0.449	0.392	0.444	0.443
Warmth for Republicans	398	0.580	0.554	0.574	0.550	0.464	0.583	0.572	0.575	0.558	0.480	0.546	0.555
Right-Wing Authoritarianism	438	0.547	0.528	0.530	0.488	0.397	0.547	0.538	0.528	0.495	0.408	0.500	0.505
Avg liking of 5 Democrats	335	0.579	0.556	0.565	0.558	0.497	0.578	0.567	0.561	0.554	0.491	0.552	0.551
Avg liking of 5 Republicans	335	0.640	0.596	0.622	0.593	0.491	0.642	0.617	0.628	0.598	0.491	0.591	0.598
Intended Vote in 2008 (D or R cand.)	375	0.557	0.547	0.548	0.547	0.490	0.566	0.562	0.550	0.548	0.491	0.538	0.544
Vote in 2004 (Bush or Kerry)	216	0.672	0.663	0.654	0.651	0.594	0.684	0.692	0.659	0.657	0.597	0.648	0.659
Bad focal average		0.359	0.320	0.364	0.327	0.228	0.399	0.386	0.398	0.377	0.292	0.320	0.371
Good focal average		0.595	0.572	0.582	0.563	0.486	0.597	0.589	0.582	0.567	0.491	0.561	0.566
RELATIONS WITH EXTRANEOUS INFLUENCE													
BAD FOCAL												average of absolute values	
Relation with average reciprocal	2055	0.049	0.053	0.151	-0.090	-0.336	0.074	0.089	0.172	-0.065	-0.324	0.138	0.147
Relation with average log	2055	-0.064	-0.064	-0.166	0.115	0.421	-0.106	-0.121	-0.201	0.076	0.401	0.170	0.184
Relation with average latency	2055	-0.058	-0.062	-0.144	0.132	0.455	-0.102	-0.115	-0.180	0.090	0.437	0.176	0.189
GOOD FOCAL												average of absolute values	
Relation with average reciprocal	2063	0.032	0.080	0.110	-0.111	-0.318	0.013	0.067	0.105	-0.114	-0.331	0.132	0.128
Relation with average log	2063	-0.065	-0.116	-0.132	0.124	0.384	-0.052	-0.100	-0.137	0.121	0.396	0.167	0.165
Relation with average latency	2063	-0.064	-0.115	-0.109	0.137	0.406	-0.055	-0.095	-0.118	0.129	0.415	0.170	0.166

Table 3. Correlations among candidate data transformations for politics after removing the first four trials, responses >10,000 ms, and responses <400 ms. Good focal blocks are to the left of the diagonal; Bad focal blocks are to the right of the diagonal.

	G	D	Reciprocal Diff	Log Diff	Latency Diff
G	-	0.971	0.965	0.953	0.817
D	0.976	-	0.944	0.948	0.835
Reciprocal Diff	0.971	0.956	-	0.962	0.797
Log Diff	0.955	0.945	0.971	-	0.925
Latency Diff	0.846	0.843	0.849	0.946	-

Table 4. Comparison of fast and slow latency treatments across evaluation criteria for politics good focal response blocks (Study 3)

	Fast Latency Treatment								Slow Latency Treatment											
	Deleting			Deleting			Recoding		Deleting			Deleting			Recoding			Recoding		
	G400	G200	G none	D400	D200	D none	G400	D400	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000
KNOWN GROUP DIFFERENCE (Political ID)	0.565	0.563	0.562	0.557	0.556	0.556	0.563	0.557	0.565	0.567	0.564	0.569	0.567	0.560	0.564	0.563	0.563	0.567	0.560	0.559
INTERNAL CONSISTENCY (alpha)	0.783	0.781	0.780	0.772	0.770	0.768	0.780	0.770	0.771	0.775	0.779	0.770	0.765	0.769	0.781	0.780	0.781	0.779	0.772	0.771
RELATIONS WITH OTHER IMPLICIT MEASURES																				
IAT	0.643	0.641	0.640	0.645	0.643	0.641	0.640	0.643	0.645	0.645	0.647	0.654	0.653	0.657	0.641	0.641	0.640	0.648	0.645	0.643
GNAT	0.674	0.674	0.674	0.676	0.676	0.675	0.674	0.677	0.665	0.668	0.669	0.667	0.667	0.667	0.674	0.673	0.673	0.677	0.674	0.675
ST-IAT	0.603	0.602	0.602	0.607	0.607	0.606	0.603	0.607	0.612	0.610	0.604	0.622	0.612	0.604	0.606	0.603	0.603	0.612	0.605	0.605
SPF	0.606	0.594	0.592	0.619	0.611	0.610	0.593	0.611	0.610	0.607	0.607	0.622	0.616	0.620	0.595	0.593	0.593	0.611	0.610	0.609
EPT	0.560	0.558	0.558	0.561	0.559	0.559	0.558	0.560	0.566	0.562	0.556	0.561	0.554	0.553	0.561	0.559	0.558	0.556	0.558	0.559
AMP	0.499	0.495	0.495	0.484	0.482	0.482	0.495	0.482	0.516	0.506	0.506	0.515	0.494	0.496	0.499	0.497	0.496	0.491	0.483	0.482
SPD	0.622	0.625	0.625	0.617	0.621	0.620	0.626	0.621	0.615	0.621	0.620	0.610	0.614	0.613	0.625	0.625	0.625	0.624	0.621	0.621
Average	0.604	0.601	0.601	0.604	0.603	0.602	0.601	0.603	0.606	0.605	0.604	0.609	0.604	0.604	0.603	0.601	0.601	0.606	0.602	0.602
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES																				
Dem-Rep Preference	0.680	0.675	0.671	0.672	0.668	0.664	0.671	0.666	0.687	0.685	0.682	0.694	0.689	0.680	0.674	0.673	0.672	0.679	0.673	0.670
Warmth for Democrats	0.465	0.465	0.461	0.452	0.451	0.446	0.461	0.448	0.472	0.467	0.467	0.475	0.460	0.457	0.461	0.461	0.461	0.459	0.452	0.450
Warmth for Republicans	0.583	0.582	0.579	0.572	0.571	0.567	0.579	0.570	0.587	0.589	0.584	0.590	0.588	0.579	0.580	0.580	0.580	0.580	0.573	0.572
Right-Wing Authoritarianism	0.547	0.545	0.544	0.538	0.537	0.537	0.545	0.537	0.541	0.544	0.547	0.531	0.533	0.537	0.546	0.546	0.545	0.535	0.537	0.537
Avg liking of 5 Democrats	0.578	0.582	0.581	0.567	0.570	0.568	0.581	0.568	0.566	0.574	0.577	0.566	0.568	0.570	0.580	0.581	0.581	0.577	0.571	0.569
Avg liking of 5 Republicans	0.642	0.643	0.641	0.617	0.617	0.615	0.641	0.615	0.645	0.653	0.645	0.644	0.646	0.628	0.647	0.643	0.641	0.639	0.623	0.618
Intended Vote in 2008 (D or R cand.)	0.566	0.560	0.561	0.562	0.558	0.558	0.561	0.559	0.570	0.565	0.566	0.582	0.568	0.566	0.560	0.561	0.561	0.570	0.563	0.561
Vote in 2004 (Bush or Kerry)	0.684	0.680	0.676	0.692	0.690	0.685	0.677	0.687	0.680	0.683	0.684	0.690	0.698	0.698	0.677	0.677	0.677	0.692	0.691	0.690
Average	0.597	0.596	0.593	0.589	0.587	0.584	0.594	0.586	0.598	0.600	0.598	0.601	0.599	0.594	0.595	0.594	0.594	0.596	0.590	0.588
RELATIONS WITH EXTRANEOUS INFLUENCE																				
Relation with average reciprocal	0.013	0.013	0.019	0.067	0.065	0.070	0.017	0.069	0.029	0.021	0.018	0.026	0.040	0.056	0.035	0.021	0.018	0.015	0.044	0.056
Relation with average log	-0.052	-0.053	-0.053	-0.111	-0.111	-0.111	-0.052	-0.111	-0.064	-0.063	-0.056	-0.055	-0.080	-0.094	-0.076	-0.057	-0.053	-0.047	-0.079	-0.094
Relation with average latency	-0.055	-0.055	-0.054	-0.111	-0.111	-0.110	-0.054	-0.111	-0.062	-0.064	-0.057	-0.051	-0.078	-0.092	-0.081	-0.059	-0.055	-0.046	-0.077	-0.092

Table 5. Comparing effects of removing versus retaining error trials for good focal blocks on evaluation criteria (Study 4).

	Politics				Race				Self-Esteem			
	G		D		G		D		G		D	
	Remove	Retain	Remove	Retain	Remove	Retain	Remove	Retain	Remove	Retain	Remove	Retain
MAGNITUDE OF MAIN EFFECT	.	.	.	.	0.344	0.423	0.357	0.450	1.036	1.118	1.026	1.097
KNOWN GROUP DIFFERENCES	0.537	0.565	0.528	0.557	0.180	0.196	0.178	0.192	.	.	.	.
INTERNAL CONSISTENCY	0.731	0.783	0.721	0.773	0.554	0.586	0.545	0.569	0.438	0.462	0.402	0.432
RELATIONS WITH OTHER IMPLICIT MEASURES	0.574	0.604	0.577	0.604	0.319	0.334	0.315	0.327	0.062	0.063	0.055	0.050
RELATIONS WITH CRITERION VARIABLES	0.573	0.597	0.569	0.589	0.183	0.196	0.189	0.198	0.061	0.064	0.051	0.059
		-		-				-		-		-
RELATIONS WITH EXTRANEOUS INFLUENCE	-0.032	0.055	-0.091	0.095	0.050	0.027	0.002	0.025	0.030	0.004	-0.044	0.070

Notes: Magnitude of main effect is Cohen's d of average BIAT score, others are correlation coefficients. Correlation coeffecients underwent Fisher's z transformation before averaging.

Table 6. Effects of applying task exclusion criteria on evaluation criteria for politics (Study 5)

	Exclusion for % fast trials (no error exclusion)			Exclusion for % errors (in addition to >10% fast exclusion)	
	none	20%	10%	40%	30%
N	2145	2080	2063	2037	1966
% of tasks excluded		3.0%	3.8%	1.3%	4.7%
cumulative % of tasks excluded				5.0%	8.3%
KNOWN GROUP DIFFERENCES					
Political Ideology (Bad Focal)	0.372	0.384	0.388	0.389	0.400
Political Ideology (Good Focal)	0.548	0.558	0.562	0.567	0.569
INTERNAL CONSISTENCY					
Alpha (Bad Focal)	0.585	0.602	0.604	0.604	0.612
Alpha (Good Focal)	0.763	0.775	0.780	0.782	0.783
RELATIONS WITH OTHER IMPLICIT MEASURES					
BAD FOCAL					
IAT	0.471	0.485	0.485	0.490	0.486
GNAT	0.569	0.570	0.573	0.572	0.572
ST-IAT	0.412	0.406	0.405	0.400	0.390
SPF	0.460	0.466	0.472	0.472	0.467
EPT	0.366	0.362	0.358	0.357	0.360
AMP	0.272	0.271	0.274	0.287	0.290
SPD	0.388	0.385	0.383	0.381	0.381
GOOD FOCAL					
IAT	0.638	0.640	0.640	0.649	0.653
GNAT	0.665	0.673	0.674	0.674	0.676
ST-IAT	0.609	0.603	0.602	0.604	0.595
SPF	0.592	0.592	0.592	0.592	0.592
EPT	0.559	0.560	0.558	0.558	0.557
AMP	0.497	0.495	0.495	0.504	0.504
SPD	0.626	0.626	0.625	0.627	0.625
Bad focal average	0.424	0.425	0.426	0.427	0.425
Good focal average	0.600	0.601	0.601	0.604	0.603
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES					
BAD FOCAL					
Dem-Rep Preference	0.431	0.437	0.437	0.434	0.428
Warmth for Democrats	0.335	0.348	0.350	0.348	0.356
Warmth for Republicans	0.394	0.401	0.401	0.399	0.389
Right-Wing Authoritarianism	0.468	0.466	0.467	0.464	0.464
Avg liking of 5 Democrats	0.332	0.344	0.344	0.351	0.352
Avg liking of 5 Republicans	0.368	0.382	0.382	0.377	0.372
Intended Vote in 2008 (D or R cand.)	0.323	0.334	0.334	0.339	0.333
Vote in 2004 (Bush or Kerry)	0.456	0.458	0.452	0.452	0.452
GOOD FOCAL					
Dem-Rep Preference	0.655	0.671	0.671	0.670	0.677
Warmth for Democrats	0.445	0.465	0.461	0.458	0.473
Warmth for Republicans	0.560	0.570	0.579	0.576	0.572
Right-Wing Authoritarianism	0.541	0.545	0.544	0.548	0.548
Avg liking of 5 Democrats	0.571	0.581	0.581	0.579	0.581
Avg liking of 5 Republicans	0.634	0.641	0.641	0.637	0.639
Intended Vote in 2008 (D or R cand.)	0.538	0.556	0.561	0.557	0.560
Vote in 2004 (Bush or Kerry)	0.618	0.660	0.676	0.673	0.669
Bad focal average	0.390	0.397	0.397	0.397	0.394
Good focal average	0.573	0.590	0.593	0.591	0.594
RELATIONS WITH EXTRANEOUS INFLUENCE					
BAD FOCAL					
Relation with average reciprocal	-0.072	0.047	0.073	0.070	0.075
Relation with average log	-0.010	-0.106	-0.107	-0.100	-0.102
Relation with average latency	-0.069	-0.103	-0.104	-0.096	-0.100
GOOD FOCAL					
Relation with average reciprocal	-0.092	0.001	0.019	0.018	0.022
Relation with average log	0.047	-0.049	-0.053	-0.048	-0.050
Relation with average latency	-0.015	-0.054	-0.054	-0.047	-0.049

Table 7. Analyses of *D* and *G* measures Based on First 40 Trials vs. Second 40 Trials

	Recoding	Trial subsets	BIAT	$r_{\text{explicit political}}$	$r_{\text{explicit race}}$	$r_{\text{implicit political}}$	$r_{\text{implicit race}}$	Cronbach α
D	none	1st 40	0.583	.543	.248	.543	***.152***	.768
		2nd 40	***0.558***	.581	***.241***	***.562***	***.170***	
	<400 = 400; >2000 = 2000	1st 40	***0.591***	***.559***	.249	.549	***.154***	.779
		2nd 40	0.567	.590	***.245***	***.563***	***.173***	
G	none	1st 40	***0.587***	***.556***	***.252***	***.550***	.149	***.780***
		2nd 40	0.569	***.589***	.235	.551	.163	
	<400 = 400; >2000 = 2000	1st 40	0.586	.558	.249	***.552***	.150	***.781***
		2nd 40	0.567	.590	.236	.553	.164	

Note. Recoding treatments are described in text. *D* and *G* are the two best performing BIAT scoring algorithms as described in the text. Underlines indicate the trial subset (1st or 2nd) with larger value for each combination of measure type and recoding trearment. Bold italics indicate the measure type (*D* or *G*) with larger value for each combination of trial subset and recoding treatment (3 ties left unmarked). The "BIAT" column gives Cohen's d effect size measures for difference of mean BIAT scores from zero. $r_{\text{explicit.political}}$ is the averaged correlation of the political BIAT with seven self-report measures of political beliefs (range of Ns: 229–2,057); $r_{\text{explicit.race}}$ is the averaged correlation of the political BIAT with three self-report measures of racial attitudes (range of Ns: 446–463); $r_{\text{implicit.political}}$ is the BIAT's average correlation with 7 other implicit political measures (range of Ns: 255–435); $r_{\text{implicit.race}}$ is the BIAT's average correlation with 7 implicit measures of race attitudes (range of Ns: 256–425); Cronbach's α is a measure of internal consistency based on each pair of 40-trial measures (N=2,136).

Table 8. Recommended scoring practices for BIAT using procedure described in Table 1.

	Steps for scoring with G		Steps for scoring with D	
	Recommended	Nearly equivalent alternatives	Recommended	Nearly equivalent alternatives
1	Remove trials >10000 milliseconds		Remove trials >10000 milliseconds	
2	Remove 1st four trials of each response block		Remove 1st four trials of each response block	
3	Retain error trials		Retain error trials	
4	No extreme latency treatment	Recode <400ms to 400ms and >2000ms to 2000ms	Recode <400ms to 400ms and >2000ms to 2000ms	Remove <400ms and >2000ms
5	Compute G separately for blocks 2, 3 and 4, 5 and then average		Compute D separately for blocks 2, 3 and 4, 5 and then average	
6	Remove tasks with >10% fast responses	Remove tasks with >10% fast responses and >30% errors	Remove tasks with >10% fast responses	Remove tasks with >10% fast responses and >30% errors

Appendix A. Comparison of bad and good focal blocks, retaining or removing 1st four trials of each block, and candidate data transformations on evaluation criteria for racial attitudes. Magnitude of main effect is Cohen's d of average BIAT score, others are correlation coefficients. Correlations averaged after Fisher's z-transformation and then converted back to a correlation.

		Retain all trials					Remove 1st four trials of each block					Average across algorithms	
	N	G	D	Reciprocal Diff	Log Diff	Latency Diff	G	D	Reciprocal Diff	Log Diff	Latency Diff	All Trials	Remove 1st four
MAGNITUDE OF MAIN EFFECT													
Preference for White (Bad Focal)	2232	0.589	0.572	0.589	0.562	0.425	0.666	0.692	0.666	0.65	0.522	0.593	0.639
Preference for White (Good Focal)	2230	0.391	0.409	0.417	0.408	0.338	0.423	0.450	0.442	0.438	0.376	0.393	0.426
KNOWN GROUP DIFFERENCES													
Participant Race - Black or White (Bad Focal)	1783	0.079	0.085	0.078	0.072	0.048	0.093	0.099	0.090	0.086	0.060	0.072	0.086
Participant Race - Black or White (Good Focal)	1783	0.169	0.173	0.178	0.179	0.155	0.178	0.184	0.178	0.187	0.183	0.171	0.182
Political Ideology (Bad Focal)	2141	0.065	0.066	0.066	0.060	0.045	0.081	0.079	0.079	0.070	0.047	0.060	0.071
Political Ideology (Good Focal)	2141	0.213	0.194	0.217	0.215	0.187	0.213	0.199	0.210	0.209	0.183	0.205	0.203
INTERNAL CONSISTENCY													
Alpha (Bad Focal)	2012	0.410	0.365	0.413	0.406	0.359	0.409	0.373	0.418	0.401	0.345	0.391	0.390
Alpha (Good Focal)	2024	0.606	0.562	0.609	0.592	0.518	0.586	0.569	0.592	0.573	0.492	0.578	0.563
RELATIONS WITH OTHER IMPLICIT MEASURES													
BAD FOCAL													
IAT	293	0.348	0.329	0.365	0.323	0.209	0.393	0.399	0.411	0.374	0.264	0.316	0.369
GNAT	316	0.231	0.228	0.240	0.210	0.136	0.248	0.249	0.252	0.220	0.136	0.209	0.221
ST-IAT	296	0.233	0.232	0.266	0.259	0.198	0.250	0.253	0.275	0.270	0.215	0.238	0.253
SPF	304	0.215	0.214	0.194	0.155	0.097	0.218	0.193	0.197	0.147	0.066	0.175	0.165
EPT	286	0.113	0.056	0.100	0.092	0.074	0.102	0.055	0.095	0.078	0.051	0.087	0.076
AMP	399	0.095	0.076	0.106	0.111	0.102	0.102	0.096	0.103	0.101	0.087	0.098	0.098
SPD	425	0.110	0.070	0.127	0.098	0.054	0.137	0.118	0.152	0.122	0.075	0.092	0.121
GOOD FOCAL													
IAT	293	0.421	0.399	0.436	0.421	0.347	0.435	0.404	0.458	0.447	0.388	0.405	0.427
GNAT	316	0.372	0.386	0.380	0.378	0.352	0.362	0.358	0.365	0.365	0.342	0.374	0.358
ST-IAT	296	0.379	0.328	0.406	0.374	0.287	0.392	0.375	0.413	0.385	0.311	0.356	0.376
SPF	304	0.278	0.286	0.283	0.242	0.158	0.302	0.312	0.305	0.268	0.196	0.250	0.277
EPT	286	0.318	0.303	0.312	0.290	0.335	0.316	0.308	0.318	0.299	0.241	0.312	0.297
AMP	399	0.203	0.217	0.221	0.218	0.179	0.206	0.227	0.224	0.227	0.203	0.208	0.217
SPD	425	0.311	0.293	0.326	0.314	0.264	0.312	0.297	0.327	0.316	0.276	0.302	0.306
Bad focal average		0.194	0.174	0.202	0.180	0.125	0.209	0.197	0.215	0.190	0.129	0.175	0.188
Good focal average		0.328	0.317	0.340	0.321	0.276	0.334	0.327	0.346	0.331	0.281	0.316	0.324
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES													
BAD FOCAL													
Black-White Preference	487	0.132	0.149	0.129	0.135	0.126	0.135	0.140	0.139	0.152	0.149	0.134	0.143
Warmth for Blacks	503	0.044	0.052	0.058	0.054	0.039	0.176	0.192	0.081	0.087	0.080	0.049	0.124
Warmth for Whites	503	0.069	0.068	0.059	0.053	0.040	0.043	0.034	0.043	0.034	0.023	0.058	0.035
Avg liking of 5 Black people	343	0.043	0.017	0.071	0.051	0.011	0.087	0.072	0.107	0.090	0.043	0.039	0.080
Avg liking of 5 White people	342	0.099	0.085	0.114	0.130	0.126	0.118	0.110	0.124	0.143	0.146	0.111	0.128
Modern Racism Scale	484	0.133	0.135	0.119	0.102	0.080	0.156	0.134	0.140	0.114	0.075	0.114	0.124
Contact with Black people	499	0.069	0.104	0.076	0.090	0.103	0.078	0.091	0.092	0.106	0.113	0.088	0.096
Right-Wing Authoritarianism	453	0.132	0.095	0.141	0.115	0.067	0.151	0.120	0.154	0.137	0.098	0.110	0.132
GOOD FOCAL													
Black-White Preference	487	0.267	0.282	0.278	0.282	0.243	0.267	0.278	0.274	0.285	0.272	0.270	0.275
Warmth for Blacks	503	0.121	0.133	0.129	0.136	0.138	0.111	0.115	0.114	0.118	0.115	0.131	0.115
Warmth for Whites	503	0.158	0.156	0.152	0.147	0.115	0.159	0.163	0.155	0.152	0.133	0.146	0.152
Avg liking of 5 Black people	343	0.125	0.131	0.156	0.161	0.152	0.136	0.133	0.158	0.162	0.153	0.145	0.148
Avg liking of 5 White people	342	0.188	0.167	0.200	0.199	0.184	0.217	0.206	0.212	0.217	0.216	0.188	0.214
Modern Racism Scale	484	0.326	0.314	0.363	0.331	0.257	0.324	0.334	0.357	0.336	0.281	0.319	0.327
Contact with Black people	499	0.100	0.094	0.091	0.087	0.070	0.107	0.111	0.097	0.099	0.094	0.088	0.102
Right-Wing Authoritarianism	453	0.236	0.253	0.255	0.282	0.282	0.234	0.236	0.247	0.271	0.277	0.262	0.253
Bad focal average		0.090	0.088	0.096	0.091	0.074	0.118	0.112	0.110	0.108	0.091	0.088	0.108
Good focal average		0.191	0.192	0.205	0.205	0.181	0.196	0.198	0.203	0.206	0.194	0.195	0.199
RELATIONS WITH EXTRANEOUS INFLUENCE													
BAD FOCAL												average of absolute values	
Relation with average reciprocal	2232	-0.015	-0.011	0.090	-0.111	-0.306	-0.006	0.023	0.121	-0.112	-0.345	0.108	0.124
Relation with average log	2232	0.061	0.051	-0.058	0.215	0.478	0.040	0.008	-0.081	0.184	0.459	0.180	0.161
Relation with average latency	2232	0.069	0.055	-0.028	0.241	0.519	0.048	0.015	-0.051	0.211	0.506	0.192	0.175
GOOD FOCAL												average of absolute values	
Relation with average reciprocal	2230	-0.054	-0.003	0.025	-0.146	-0.290	-0.030	0.019	0.055	-0.147	-0.342	0.105	0.121
Relation with average log	2230	0.059	-0.001	-0.013	0.210	0.412	0.035	-0.018	-0.036	0.185	0.411	0.144	0.142
Relation with average latency	2230	0.055	-0.006	0.005	0.225	0.441	0.027	-0.025	-0.022	0.195	0.435	0.152	0.146

Appendix B. Comparison of bad and good focal blocks, retaining or removing 1st four trials of each block, and candidate data transformations on evaluation criteria for self-esteem. Magnitude of main effect is Cohen's d of average BIAT score, others are correlation coefficients. Correlations averaged after Fisher's z-transformation and then converted back to a correlation.

		Retain all trials					Remove 1st four trials of each block					Average across algorithms	
	N	G	D	Reciprocal Diff	Log Diff	Latency Diff	G	D	Reciprocal Diff	Log Diff	Latency Diff	All Trials	Remove 1st four
MAGNITUDE OF MAIN EFFECT													
Self-Esteem (Bad Focal)	2152	0.284	0.277	0.289	0.264	0.182	0.415	0.439	0.425	0.408	0.310	0.259	0.399
Self-esteem (Good Focal)	2127	1.092	0.988	1.010	0.987	0.783	1.118	1.097	1.045	1.033	0.861	0.972	1.031
INTERNAL CONSISTENCY													
Alpha (Bad Focal)	2004	0.458	0.384	0.436	0.420	0.355	0.450	0.394	0.423	0.408	0.369	0.411	0.409
Alpha (Good Focal)	2014	0.466	0.431	0.468	0.453	0.355	0.462	0.432	0.462	0.449	0.370	0.435	0.436
RELATIONS WITH OTHER IMPLICIT MEASURES													
BAD FOCAL													
IAT	291	0.117	0.095	0.105	0.105	0.092	0.142	0.132	0.132	0.131	0.110	0.103	0.129
GNAT	295	0.063	0.084	0.069	0.083	0.103	0.078	0.077	0.073	0.081	0.092	0.080	0.080
ST-IAT	300	0.083	0.082	0.064	0.063	0.069	0.108	0.089	0.086	0.054	0.070	0.072	0.081
SPF	286	0.056	0.025	0.055	0.034	-0.055	0.073	0.074	0.074	0.057	0.019	0.023	0.059
EPT	275	0.007	0.053	0.010	0.037	0.065	0.022	0.040	0.018	0.038	0.059	0.034	0.035
AMP	340	0.082	0.061	0.083	0.073	0.067	0.088	0.057	0.075	0.065	0.057	0.073	0.068
SPD	401	-0.031	-0.034	-0.033	-0.014	0.009	-0.039	-0.042	-0.037	-0.013	0.023	-0.021	-0.022
GOOD FOCAL													
IAT	291	0.086	0.069	0.120	0.077	0.013	0.103	0.089	0.132	0.092	0.028	0.073	0.089
GNAT	295	0.086	0.030	0.083	0.047	-0.003	0.114	0.093	0.090	0.059	0.016	0.049	0.074
ST-IAT	300	0.018	-0.046	-0.011	-0.033	-0.065	0.041	0.002	0.011	0.004	-0.015	-0.027	0.009
SPF	286	0.036	0.002	0.037	-0.009	-0.054	0.023	0.001	0.040	0.004	-0.039	0.002	0.006
EPT	275	0.123	0.113	0.155	0.158	0.127	0.140	0.149	0.167	0.174	0.167	0.135	0.159
AMP	340	-0.001	0.022	0.003	0.004	0.011	0.016	0.019	0.016	0.02	0.016	0.008	0.017
SPD	401	-0.017	-0.029	-0.037	-0.032	-0.021	0.004	-0.001	-0.014	-0.006	0.001	-0.027	-0.003
Bad focal average		0.054	0.052	0.051	0.054	0.050	0.068	0.061	0.060	0.059	0.061	0.052	0.062
Good focal average		0.047	0.023	0.050	0.030	0.001	0.063	0.050	0.063	0.050	0.025	0.030	0.050
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES													
BAD FOCAL													
Self-Other Preference	477	0.133	0.121	0.138	0.127	0.101	0.126	0.140	0.134	0.121	0.102	0.124	0.125
Warmth for Self	479	0.057	0.012	0.051	0.031	-0.002	0.042	0.034	0.042	0.024	0.002	0.030	0.029
Warmth for Others	480	0.038	0.064	0.050	0.045	0.043	0.048	0.056	0.068	0.058	0.047	0.048	0.055
Self-Attributes Questionnaire	427	-0.015	-0.020	0.001	-0.024	-0.048	-0.008	-0.026	0.010	-0.011	-0.037	-0.021	-0.014
Rosenberg Self-Esteem	472	0.161	0.133	0.172	0.174	0.156	0.172	0.166	0.176	0.182	0.173	0.159	0.174
Recency of Positive Feedback	479	0.026	0.034	0.015	0.030	0.042	0.019	0.028	0.015	0.027	0.035	0.029	0.025
Recency of Negative Feedback	479	0.027	0.032	0.024	0.003	-0.012	0.021	0.016	0.019	-0.002	-0.029	0.015	0.005
GOOD FOCAL													
Self-Other Preference	477	0.084	0.044	0.085	0.052	0.004	0.085	0.076	0.090	0.059	0.018	0.054	0.066
Warmth for Self	479	0.039	0.042	0.037	0.035	0.016	0.053	0.066	0.052	0.051	0.035	0.034	0.051
Warmth for Others	480	0.108	0.069	0.094	0.086	0.069	0.112	0.077	0.084	0.076	0.061	0.085	0.082
Self-Attributes Questionnaire	427	0.077	0.055	0.085	0.090	0.083	0.074	0.044	0.085	0.081	0.057	0.078	0.068
Rosenberg Self-Esteem	472	0.069	0.026	0.038	0.036	0.006	0.083	0.058	0.051	0.052	0.026	0.035	0.054
Recency of Positive Feedback	479	0.101	0.078	0.097	0.093	0.070	0.098	0.093	0.089	0.083	0.065	0.088	0.086
Recency of Negative Feedback	479	0.006	0.080	0.039	0.050	0.057	0.006	0.055	0.040	0.046	0.047	0.046	0.039
Bad focal average		0.054	0.047	0.057	0.049	0.035	0.053	0.052	0.058	0.050	0.037	0.048	0.050
Good focal average		0.061	0.049	0.059	0.055	0.038	0.064	0.059	0.061	0.056	0.039	0.053	0.056
RELATIONS WITH EXTRANEOUS INFLUENCE													
BAD FOCAL												average of absolute values	
Relation with average reciprocal	2048	-0.026	-0.030	0.099	-0.140	-0.338	-0.015	-0.009	0.119	-0.115	-0.320	0.129	0.118
Relation with average log	2048	0.031	0.036	-0.095	0.190	0.445	0.002	-0.011	-0.133	0.151	0.425	0.165	0.149
Relation with average latency	2048	0.033	0.038	-0.069	0.213	0.487	0.000	-0.010	-0.109	0.174	0.472	0.176	0.160
GOOD FOCAL												average of absolute values	
Relation with average reciprocal	2066	-0.019	0.053	0.090	-0.124	-0.318	-0.032	0.029	0.071	-0.131	-0.316	0.123	0.118
Relation with average log	2066	0.010	-0.067	-0.089	0.161	0.398	0.015	-0.052	-0.081	0.165	0.402	0.149	0.147
Relation with average latency	2066	-0.006	-0.083	-0.079	0.165	0.420	-0.004	-0.070	-0.073	0.167	0.420	0.155	0.152

Appendix C. Comparison of fast and slow latency treatments across evaluation criteria for race. Magnitude of main effect is Cohen's d of average BIAT score, others are correlation coefficients. Correlations averaged after Fisher's z-transformation and then converted back to a correlation.

	Fast Latency Treatment								Slow Latency Treatment											
	Deleting			Deleting			Recoding		Deleting			Deleting			Recoding			Recoding		
	G400	G200	G none	D400	D200	D none	G400	D400	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000
MAGNITUDE OF MAIN EFFECT	0.423	0.415	0.415	0.450	0.444	0.444	0.415	0.445	0.419	0.420	0.422	0.460	0.451	0.453	0.415	0.415	0.415	0.454	0.447	0.446
KNOWN GROUP DIFFERENCE (Race)	0.178	0.177	0.177	0.184	0.184	0.183	0.177	0.183	0.167	0.175	0.178	0.173	0.183	0.186	0.175	0.177	0.177	0.183	0.187	0.185
KNOWN GROUP DIFFERENCE (Political ID)	0.213	0.220	0.221	0.199	0.204	0.206	0.221	0.205	0.209	0.214	0.212	0.207	0.208	0.201	0.221	0.220	0.221	0.213	0.208	0.206
INTERNAL CONSISTENCY (alpha)	0.586	0.594	0.594	0.569	0.575	0.576	0.595	0.575	0.572	0.582	0.584	0.566	0.564	0.568	0.597	0.596	0.595	0.578	0.578	0.577
RELATIONS WITH OTHER IMPLICIT MEASURES																				
IAT	0.435	0.451	0.448	0.404	0.415	0.412	0.453	0.415	0.438	0.440	0.438	0.441	0.430	0.421	0.454	0.453	0.453	0.436	0.423	0.418
GNAT	0.362	0.362	0.361	0.358	0.360	0.360	0.361	0.360	0.349	0.360	0.361	0.350	0.360	0.360	0.359	0.360	0.361	0.357	0.362	0.361
ST-IAT	0.392	0.395	0.395	0.375	0.379	0.379	0.394	0.379	0.404	0.398	0.395	0.401	0.390	0.383	0.396	0.393	0.393	0.389	0.383	0.379
SPF	0.302	0.308	0.308	0.312	0.318	0.318	0.308	0.318	0.298	0.302	0.300	0.301	0.306	0.311	0.309	0.308	0.307	0.311	0.316	0.318
EPT	0.316	0.330	0.330	0.308	0.324	0.323	0.329	0.322	0.316	0.320	0.315	0.313	0.315	0.307	0.330	0.329	0.329	0.321	0.321	0.321
AMP	0.206	0.205	0.205	0.227	0.227	0.227	0.206	0.227	0.199	0.206	0.206	0.219	0.229	0.227	0.205	0.206	0.206	0.226	0.228	0.228
SPD	0.312	0.307	0.306	0.297	0.297	0.296	0.306	0.297	0.311	0.317	0.315	0.310	0.311	0.304	0.309	0.307	0.306	0.309	0.300	0.298
Average	0.334	0.339	0.338	0.327	0.333	0.332	0.339	0.332	0.333	0.337	0.335	0.335	0.336	0.332	0.339	0.339	0.338	0.337	0.335	0.333
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES																				
Black-White Preference	0.267	0.269	0.268	0.278	0.278	0.277	0.269	0.278	0.258	0.264	0.264	0.269	0.274	0.275	0.269	0.269	0.268	0.281	0.280	0.278
Warmth for Blacks	0.111	0.109	0.109	0.115	0.116	0.115	0.110	0.116	0.103	0.105	0.108	0.103	0.104	0.108	0.107	0.110	0.110	0.110	0.113	0.116
Warmth for Whites	0.159	0.163	0.162	0.163	0.164	0.164	0.162	0.163	0.162	0.157	0.158	0.157	0.162	0.164	0.163	0.162	0.161	0.166	0.164	0.163
Avg liking of 5 Black people	0.136	0.146	0.146	0.133	0.143	0.143	0.147	0.142	0.140	0.136	0.136	0.143	0.135	0.132	0.146	0.147	0.147	0.148	0.142	0.142
Avg liking of 5 White people	0.217	0.222	0.222	0.206	0.211	0.211	0.222	0.210	0.225	0.207	0.207	0.222	0.190	0.185	0.218	0.221	0.222	0.208	0.205	0.209
Modern Racism Scale	0.324	0.321	0.321	0.334	0.335	0.335	0.324	0.336	0.321	0.323	0.325	0.336	0.330	0.336	0.326	0.325	0.324	0.338	0.337	0.337
Contact with Black people	0.107	0.100	0.099	0.111	0.107	0.106	0.101	0.108	0.103	0.102	0.103	0.104	0.099	0.102	0.100	0.102	0.101	0.103	0.106	0.107
Right-Wing Authoritarianism	0.234	0.248	0.251	0.237	0.248	0.250	0.252	0.250	0.220	0.235	0.232	0.236	0.243	0.237	0.251	0.252	0.252	0.260	0.255	0.253
Average	0.196	0.198	0.198	0.198	0.202	0.201	0.200	0.202	0.193	0.192	0.193	0.198	0.193	0.194	0.199	0.200	0.199	0.203	0.202	0.202
RELATIONS WITH EXTRANEOUS INFLUENCE																				
Relation with average reciprocal	-0.030	-0.031	-0.031	0.019	0.012	0.012	-0.035	0.013	-0.017	-0.026	-0.023	-0.006	-0.004	0.009	-0.023	-0.032	-0.034	-0.023	-0.008	0.004
Relation with average log	0.035	0.038	0.038	-0.018	-0.011	-0.010	0.041	-0.011	0.023	0.031	0.027	0.015	0.011	-0.006	0.026	0.037	0.04	0.032	0.016	-0.0001
Relation with average latency	0.027	0.030	0.030	-0.025	-0.019	-0.019	0.032	-0.020	0.018	0.026	0.019	0.014	0.010	-0.010	0.015	0.029	0.031	0.027	0.012	-0.006

Appendix D. Comparison of fast and slow latency treatments across evaluation criteria for self-esteem. Magnitude of main effect is Cohen's d of average BIAT score, others are correlation coefficients. Correlations averaged after Fisher's z-transformation and then converted back to a correlation.

	Fast Latency Treatment								Slow Latency Treatment											
	Deleting			Deleting			Recoding		Deleting			Deleting			Recoding			Recoding		
	G400	G200	G none	D400	D200	D none	G400	D400	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000	G400 + G2000	G400 + G3000	G400 + G4000	D400 + D2000	D400 + D3000	D400 + D4000
MAGNITUDE OF MAIN EFFECT	1.187	1.187	1.184	1.164	1.163	1.160	1.184	1.162	1.163	1.189	1.185	1.178	1.186	1.167	1.187	1.185	1.185	1.190	1.173	1.166
INTERNAL CONSISTENCY (alpha)	0.462	0.459	0.458	0.432	0.429	0.428	0.458	0.429	0.467	0.428	0.436	0.444	0.433	0.431	0.459	0.459	0.458	0.474	0.456	0.461
RELATIONS WITH OTHER IMPLICIT MEASURES																				
IAT	0.103	0.109	0.108	0.089	0.094	0.094	0.107	0.093	0.142	0.124	0.111	0.155	0.121	0.102	0.110	0.109	0.108	0.110	0.098	0.096
GNAT	0.114	0.113	0.113	0.093	0.094	0.094	0.112	0.093	0.125	0.121	0.117	0.114	0.099	0.096	0.114	0.113	0.112	0.092	0.093	0.093
ST-IAT	0.041	0.043	0.053	0.002	0.001	0.011	0.052	0.009	0.063	0.047	0.046	0.045	0.016	0.011	0.054	0.052	0.051	0.021	0.012	0.009
SPF	0.023	0.025	0.025	0.001	0.001	0.001	0.026	0.002	0.049	0.035	0.024	0.035	0.019	-0.003	0.029	0.026	0.026	0.010	0.004	0.003
EPT	0.140	0.137	0.136	0.149	0.144	0.143	0.137	0.146	0.137	0.136	0.138	0.159	0.145	0.147	0.136	0.137	0.137	0.151	0.148	0.147
AMP	0.016	0.026	0.026	0.019	0.027	0.027	0.025	0.026	0.016	0.015	0.016	0.017	0.018	0.015	0.024	0.025	0.025	0.027	0.028	0.029
SPD	0.004	0.007	0.006	-0.001	0.001	-0.0003	0.005	-0.0003	0.012	0.007	0.007	0.016	0.003	0.002	0.005	0.004	0.005	0.003	-0.003	-0.003
Average	0.063	0.066	0.067	0.050	0.052	0.053	0.066	0.053	0.078	0.069	0.066	0.078	0.060	0.053	0.068	0.067	0.066	0.059	0.054	0.054
RELATIONS WITH SELF-REPORT MEASURES AND CRITERION VARIABLES																				
Self-Other Preference	0.085	0.089	0.090	0.076	0.080	0.081	0.089	0.080	0.079	0.086	0.084	0.062	0.064	0.067	0.091	0.089	0.088	0.068	0.072	0.074
Warmth for Self	0.053	0.058	0.059	0.066	0.069	0.070	0.057	0.069	0.047	0.052	0.055	0.061	0.058	0.063	0.056	0.057	0.057	0.062	0.064	0.065
Warmth for Others	0.112	0.109	0.109	0.077	0.079	0.079	0.108	0.079	0.113	0.114	0.114	0.081	0.073	0.074	0.108	0.109	0.108	0.077	0.078	0.079
Self-Attributes Questionnaire	0.074	0.080	0.079	0.044	0.048	0.047	0.080	0.047	0.089	0.085	0.082	0.085	0.073	0.063	0.082	0.080	0.080	0.068	0.056	0.049
Rosenberg Self-Esteem	0.083	0.088	0.084	0.058	0.061	0.057	0.085	0.058	0.093	0.094	0.089	0.085	0.086	0.076	0.088	0.086	0.085	0.075	0.067	0.062
Recency of Positive Feedback	0.098	0.093	0.093	0.093	0.091	0.091	0.092	0.091	0.112	0.099	0.094	0.116	0.092	0.084	0.090	0.090	0.092	0.093	0.089	0.091
Recency of Negative Feedback	0.006	0.002	0.001	0.055	0.051	0.050	0.0003	0.051	-0.013	0.009	0.007	0.025	0.053	0.054	-0.001	-0.001	-0.0003	0.044	0.049	0.051
Average	0.064	0.065	0.064	0.059	0.060	0.059	0.064	0.059	0.065	0.067	0.066	0.064	0.062	0.060	0.064	0.064	0.064	0.061	0.059	0.059
RELATIONS WITH EXTRANEOUS INFLUENCE																				
Relation with average reciprocal	-0.032	-0.025	-0.028	0.029	0.034	0.030	-0.031	0.030	-0.044	-0.031	-0.032	-0.043	-0.005	0.009	-0.022	-0.030	-0.031	-0.028	0.002	0.016
Relation with average log	0.015	0.008	0.008	-0.052	-0.057	-0.057	0.011	-0.056	0.023	0.011	0.016	0.024	-0.013	-0.024	-0.002	0.010	0.011	0.015	-0.018	-0.036
Relation with average latency	-0.004	-0.009	-0.009	-0.07	-0.074	-0.074	-0.007	-0.073	0.007	-0.006	-0.001	0.012	-0.024	-0.036	-0.022	-0.008	-0.007	0.003	-0.030	-0.05

Appendix E. A step-by-step guide for calculating the recommended G and D scores from Table 8		
	n_1 latencies from condition 1 are contrasted with n_2 latencies in condition 2, $n_1 + n_2 = N$	
	Steps for calculating G	Steps for calculating D
1	Assign fractional ranks to N latencies. The longest latency will be assigned a value of 1.0 and the shortest will be assigned a value of 1/N; In case of ties, ranks are averaged across tied values.	Compute the standard deviation of the N latencies, SD.
2	Subtract 1/2N from each fractional rank. Assuming untied values, the largest latency will now have a value of 1-1/2N or (2N-1)/2N and the smallest adjusted fractional rank latency will be 1/2N. The 1/2N downward adjustment applies generally, even when tied values exist.	M_1 is the mean of the latencies in condition 1. M_2 is the mean of the latencies in condition 2.
3	For each of the N observations, compute the standard normal deviate (mean=0 and standard deviation=1) corresponding to the adjusted fractional rank latency.	$D = \frac{M_2 - M_1}{SD}$
4	G_1 is the mean of the normal deviates in condition 1. G_2 is the mean of the normal deviates in condition 2.	
5	$G = G_2 - G_1$	