

Chapter 6

ASSESSING IMPLICIT COGNITIONS WITH A PAPER-FORMAT IMPLICIT ASSOCIATION TEST

*Kristi M. Lemm¹, Kristin A. Lane², David N. Sattler¹,
Saera R. Khan³, and Brian A. Nosek⁴*

1. Western Washington University, USA

2. Bard College, USA

3. University of San Francisco, USA

4. University of Virginia, USA

ABSTRACT

The Implicit Association Test (IAT; Greenwald, McGhee, and Schwartz, 1998) is a frequently used measure of implicit cognitions that is typically administered on computers. This chapter reports development of an IAT that can be administered on paper. First, it describes a suggested analytic procedure for paper IAT data. Next, two studies measuring implicit racial preferences are reported that suggest that the paper-format IAT elicits similar but somewhat weaker mean effects than the computer-format IAT, and shows test-retest reliability comparable to the computer-format IAT. The paper format IAT may be more sensitive to the type of stimuli used in the task. It performed better with all-verbal stimuli compared with pictures of faces. Use of the paper-format IAT with verbal stimuli may be a useful supplement to computerized data collections, or a viable approach when computer data collection is not feasible.

INTRODUCTION

The assessment of attitudes has a long theoretical and methodological history in social psychology. One recent insight, that people may possess attitudes of which they are

1 Correspondence concerning this chapter should be addressed to Kristi M. Lemm, Department of Psychology, Western Washington University, 516 High Street, MS 9089, Bellingham, WA 98225; Email: Kristi.Lemm@wwu.edu.

consciously unaware, has spurred the development of measures to assess such cognitions without relying on verbal self-reports. Most of these measures use response latency as the dependent variable, and a computer for administration. Less conscious measures of attitudes typically reveal more negativity toward socially stigmatized outgroups than merely asking people how they feel (Nosek et al., 2007). Additionally, implicit and explicit measures explain unique variance in people's behavior (Greenwald, Poehlman, Uhlmann, and Banaji, in press). Sole reliance on computers to administer such tasks restricts researchers' ability to collect a large amount of data in a single setting and potentially to reach non-traditional populations (e.g., conference attendees, students on a camping retreat, beach-goers, people on the train). "Low-tech" means of assessing such attitudes may prove to be a useful supplement to computer-based measures (Vargas, Sekaquaptewa, and von Hippel, 2007).

The Implicit Association Test (IAT; Greenwald et al., 1998) is a widely used tool for assessing implicit attitudes and beliefs (see Lane, Banaji, Nosek, and Greenwald, 2007; Nosek, Greenwald and Banaji, 2007 for reviews). The IAT measures the relative strength of association between concepts in different categories (e.g., flower/insect and pleasant/unpleasant) by measuring the amount of time it takes to categorize stimuli from the four categories with just two response options. In one condition, participants give one response (e.g., a key press on a computer) for both *flowers* and *pleasant* words and an alternate response for both *insects* and *unpleasant* words. In a second condition, participants give one response for both *flowers* and *unpleasant* words and the alternate response for *insects* and *pleasant* words. The difference in average response latency between the two pairings reflects the underlying relative strength of associations. In this example, most participants would be able to categorize the items faster in the first condition than the second, indicating relatively more positive associations toward flowers than insects. Response latency is typically assessed using computers. This provides maximal information about each response but restricts data collection to situations in which computers are readily available.

Although computer-based data collection is optimal for response latency measures, we introduce a paper-format IAT for circumstances in which a computer is not available. We summarize previous uses of this task, provide a detailed description of the procedure for administering a paper-format IAT, and describe an algorithm that is effective for scoring a paper-format version. Additionally, we report two studies that demonstrate that the psychometric properties of a paper-format IAT, while not as consistent as computer versions, are reasonable for research uses of the IAT.

DESCRIPTION AND SOME SUCCESSFUL RESEARCH APPLICATIONS OF THE PAPER-FORMAT IAT

When completing a paper-format IAT, participants see a page with two columns of stimuli and they are instructed to mark a circle to the left or right of each stimulus item to indicate its appropriate category. Two of the categories correspond with a *left* response, and the other two correspond with a *right* response. Each IAT consists of two conditions. For example, Figure 1 depicts one condition that requires participants to categorize names typical among Black people and pleasant words on the left side of the column and names typical among White people and unpleasant words on the right side.

Black Pleasant	White Unpleasant	Black Pleasant	White Unpleasant
<input type="radio"/>	Melanie	<input type="radio"/>	Rachel
<input type="radio"/>	Love	<input type="radio"/>	poison
<input type="radio"/>	Brandon	<input type="radio"/>	Peter
<input type="radio"/>	Evil	<input type="radio"/>	good
<input type="radio"/>	Malik	<input type="radio"/>	Jamal
<input type="radio"/>	Terrific	<input type="radio"/>	evil
<input type="radio"/>	Rachel	<input type="radio"/>	Brandon
<input type="radio"/>	poison	<input type="radio"/>	love
<input type="radio"/>	Sharise	<input type="radio"/>	Lionel
<input type="radio"/>	vomit	<input type="radio"/>	hatred
<input type="radio"/>	Lionel	<input type="radio"/>	Melanie
<input type="radio"/>	joy	<input type="radio"/>	joy
<input type="radio"/>	Lashelle	<input type="radio"/>	Malik
<input type="radio"/>	hatred	<input type="radio"/>	vomit
<input type="radio"/>	Amber	<input type="radio"/>	Sharise
<input type="radio"/>	bad	<input type="radio"/>	happy
<input type="radio"/>	Justin	<input type="radio"/>	Peter
<input type="radio"/>	good	<input type="radio"/>	hatred
<input type="radio"/>	Tanisha	<input type="radio"/>	Lashelle
<input type="radio"/>	happy	<input type="radio"/>	bad

Figure 1. Sample page of paper-format IAT.

In a second condition on a new page, the category pairings switch, and White names and pleasant words are categorized on the left and Black names and unpleasant words on the right side.² Participants receive a fixed time period (e.g., 20 seconds) to categorize as many items as possible starting from the top of the page and working down. The number of correct categorizations in one condition compared with the other is the key measure of relative association strength. This approach is different from computer-based IATs, in which the number of responses is fixed and the amount of time to complete them is the dependent variable. Even so, the procedural logic is the same: close association between the concepts

² The order of the critical pages and the left-right location of the pairings are counterbalanced across participants.

that share a response should make the task easier, leading to better performance (i.e., more correct responses). So, more categorizations in the second compared with the first condition above would be interpreted as an implicit preference for Whites compared with Blacks.

The patterns of data obtained from paper-format IATs generally parallel those of conceptually similar computerized IATs. For example, studies using paper-format IATs have demonstrated strong pro-White (relative to Black) implicit preference (Lowery, Hardin, and Sinclair, 2001; Sinclair, Lowery, Hardin, and Colangelo, 2005) and strong pro-thin (relative to obese) implicit preference (Teachman and Brownell, 2001; Teachman, Gapinski, Brownell, Rawlins, and Jeyaram, 2003), just as is observed with computer administration (Nosek et al., 2007). Attitudes assessed with the paper-format IAT appear to be sensitive to the same kinds of situational cues as those assessed with the computerized IAT (Blair, 2002). For example, White (but not Asian American) participants showed weaker pro-White attitudes when the test was administered by a Black experimenter than by a White experimenter (Lowery et al., 2001); anti-fat attitudes were stronger after exposure to news stories emphasizing behavioral (versus genetic) causes of obesity but weaker after exposure to an empathy-inducing story (the latter effect was only observed among overweight participants; Teachman et al., 2003). Finally, paper-format IATs have been used to measure associations of self with good (self-esteem), self with one's university (identity), and one's university with good (attitude) providing evidence for cognitive consistency among these constructs (Lane, Mitchell, and Banaji, 2005), as predicted by recent conceptions of consistency among implicit cognitions, and observed with computer-based measures (Greenwald, Banaji, Rudman, Farnham, Nosek, and Mellott, 2002). These studies suggest that paper-format forms of the IAT may be a useful instrument for implicit social cognition research.

Use of the paper-format IAT is predicated on the assumption that the computerized and paper-format versions of the IAT share the same underlying logic, and thus should measure the same construct. To date, this assumption has not been tested directly. The current research is the first to compare the results and psychometric properties of the paper-format IAT and its computer-based counterpart. In addition, because the dependent variable for the paper-format IAT is the number of responses completed in a fixed time, rather than the amount of time to complete a fixed number of responses, the recommended scoring procedures for the computer-based IAT (Greenwald, Nosek, and Banaji, 2003) do not apply. Through simulation and empirical tests, we investigated the validity of alternative analytic approaches for paper-format IAT data.

STUDY 1

The goal of Study 1 was to determine an appropriate method for analyzing paper-format IAT data. In a paper-format IAT, the dependent measure is based on the number of items completed in a fixed amount of time. Intuition may suggest using the difference between the number of items completed on the two blocks of the paper-format IAT as an index of the "IAT effect." However, individual differences in response speed create an unwanted artifact in the use of difference scores in response latency comparisons in general (Sriram, Greenwald, and Nosek, 2007) and on the IAT in particular (Greenwald et al., 2003; Hummert,

Garstka, O'Brien, Greenwald, and Mellott, 2002).³ This effect can be reduced by using the D algorithm that divides the average difference in response latency by the standard deviation of all the response trials (Cai, Sriram, Greenwald, and McFarland, 2004; Greenwald et al., 2003; Mierke and Klauer, 2003). As the paper format IAT does not provide response latencies to individual trials, this approach is not a viable strategy for the paper-format IAT. However, a number of mathematical transformations may reduce artifactual influences of individual differences in average response latency.

Consider two participants who each complete a paper-format IAT measuring implicit preference for flowers, relative to insects. Participant 1 completes 20 items in the flower+good block and 10 items in the flower+bad block, and Participant 2 completes 30 items and 20 items in those blocks, respectively. Based on the simple difference in the number of items completed (difference = 10), these two participants exhibit the same implicit preference for flowers over insects. However, Participant 1 completes twice as many items in the flower+good pairing than in the flower+bad pairing, whereas Participant 2 completes only 50% more items in the flower+good pairing than in the flower+bad pairing. As a ratio, Participant 1 has a much larger effect than Participant 2. This example illustrates how individual differences in participant speed may mask or exaggerate the magnitude of implicit preferences as gauged by a simple difference score.

Study 1 explores a variety of different procedures for analyzing paper-format IAT data. To evaluate the efficacy of these procedures, we first created a simulated data set of paper-format IAT data with individual differences in average response time. Next, we created a second data set designed to be correlated with the IAT score from the first and included random error. As the datasets were manufactured to be positively correlated, the validity of scoring procedures can be observed by identifying the approach that maximizes the correlation between the data sets.

METHOD

Simulated Datasets

We first generated a simulated dataset (D_1) of 1000 data observations. Each observation consisted of two values (A_1 and B_1) that were drawn from a range similar to those that have been seen in actual paper-format IAT data sets (Lane et al., 2005; Lemm, 2000). For ease of presentation, we assumed that these data represent an attitudinal IAT in which blocks A_1 and B_1 denote the pairings A+good and B+good, respectively. Values of A_1 greater than B_1 would reflect implicit preference for A over B. A_1 and B_1 were randomly selected integers between 10 and 40. Thus, this simulated data set assumes that items completed on any given block are uniformly distributed between 10 and 40. (Simulated data sets with normally distributed data revealed a similar pattern of results.)

3 The effect of overall response speed differs for the two versions of the task: on the computerized version, participants with slower responses overall show larger IAT effects for a measure based on difference scores; on the paper-pencil version, they would be expected to complete fewer items overall and show a smaller IAT effect based on a difference score calculation.

Nine secondary data sets (D_2) were then generated as a function of D_1 with the following formulas: (1) $A_2 = r' * A_1 + (1-r') * X$ and (2) $B_2 = r' * B_1 + (1-r') * Z$, where r' varied between 0.1 and 0.9, and X and Z were randomly selected integers between 10 and 40. That is, each new data point was a joint function of the original data set and a random value. Greater values of r' therefore reflect stronger relationships between the two data sets. In the extreme case, for example, where $r' = 1.0$, D_1 would be isomorphic to D_2 .

Finally, “extreme” data sets (D_3) were generated that were a function of D_1 but that consisted solely of extreme scores. The same formulas were used to generate values of data points of A_3 and B_3 , with the additional requirement that for each observation, the larger value of the two blocks was increased by two standard deviations. These simulated data sets allowed us to explore each proposed analytic procedure’s ability to detect correspondence between the data sets even when one of the data sets consists of extreme scores.

POTENTIAL SCORING PROCEDURES

An infinite number of approaches could be used in analyzing paper-format IAT data. In the current study, we examined the properties of seven scoring procedures. The procedures varied in their emphasis on the additive difference ($A - B$) versus the ratio difference (A/B) between the two IAT blocks. Additionally, common approaches for normalizing data (square root or inverse calculations) were included in some of the scoring approaches to test their ability to minimize the extraneous effects of extreme scores. In all cases, higher numbers reflect stronger implicit preference for A over B.

Difference score. The difference score considers only the difference in number of items completed across the two blocks, and does not adjust for participant speed. The difference score is calculated as $A - B$.

Simple ratio. The simple ratio is calculated as $A/B - 1$, and is sensitive to the relative speed of response on each block. The ratio approach provides an index of a subject’s performance on block A, relative to his or her performance on block B. One is subtracted from each ratio value in order to define the zero point as no difference in the number of correct responses between the two response conditions.

Max/min ratio. The max/min ratio is calculated as $X/Y - 1$, where X is the greater of A or B, and Y is the smaller of A or B. (If $A = B$, the formula produces a result of 0 regardless of which original value is selected as X and which is selected as Y .) As with the simple ratio approach, this approach adjusts for overall speed by assessing performance on one block of the IAT relative to performance on the other block of the IAT. However, unlike the simple ratio, this approach minimizes the effects of larger difference scores from faster subjects, relative to slower ones. Resulting values should be multiplied by -1 if B is greater than A in order to preserve the directionality of the IAT score.

Latency conversion. This approach calculates the difference in the average latency per item between the two blocks. The total number of items completed is first transformed into an average response latency per item (per second) by taking the reciprocal of the number of items completed on each block. Next, the difference between these two values is calculated. The formula for this approach is $1000 * ((1/B) - (1/A))$.

Simple product. This approach makes use of both the overall difference score and the ratio score by multiplying the difference score by the max/min ratio. It is calculated as $(A - B) * (X/Y)$, where X is the greater of A or B, and Y is the smaller of A or B.

Product: squared ratio. This approach also incorporates the overall difference and ratio scores. However, the relative performance on the two blocks is squared, and thus weighted more heavily than the difference score. It is calculated as $(A - B)*(X/Y)^2$, where X is the greater of A or B, and Y is the smaller of A or B.

Product: square root of difference. This approach includes both difference score and ratio information. In this case, the square root of the difference between the number of items completed between the two blocks is multiplied by the ratio of items completed. It is calculated as $(X/Y)*\text{Square Root of } (X - Y)$, where X is the greater of A or B, and Y is the smaller of A or B. If B is greater than A, the resulting values are multiplied by -1 to retain the directionality of the IAT effect.

RESULTS

Correlations among Simulated Data Sets

IAT effects were calculated for the simulated data using each of the proposed transformations. Each scoring procedure was applied to the data sets D_1 (original data set), D_2 (non-extreme data set that was a function of D_1), and D_3 (extreme data set that was a function of D_1) at varying levels of correspondence (r') between D_1 and the secondary data sets D_2 and D_3 . As D_2 and D_3 are direct functions of D_1 , there is a true relationship among them. Superior scoring procedures ought to reproduce this functional relationship, and thus maximize the correlation between the data sets D_1 and D_2 and between the data sets D_1 and D_3 .

The different scoring procedures did not affect correlations among IAT scores at low levels of correspondence between D_1 and D_2 ($r' = 0.1, 0.2$). However, at higher degrees of correspondence between the data sets, transformation selection did influence the observed correlation among the data. As illustrated in Figure 2, the *simple ratio* transformation was far inferior to the other transformations, and the *product: squared ratio* and *max/min ratio* transformations lagged behind the others in maximizing the observed correlations between D_1 and D_2 . As these procedures were inferior for non-extreme data, they were not considered in the analysis of the extreme data set.

We next examined the ability of the four remaining scoring procedures to maximize correspondence between D_1 and D_3 , which was functionally related to D_1 but consisted of extreme scores. As can be seen in Figure 3, at eight out of nine levels of r' , the *product: square root of difference* approach outperformed the other scoring procedures. For the two highest levels of defined correspondence ($r' = 0.8, 0.9$) between D_1 and D_3 , the *product: square root of difference* approach explained significantly more variance than the next-best transformation, $z_s > 2.00$.

DISCUSSION

The ability of seven proposed scoring procedures to maximize the correlation between functionally related data sets was explored. Of the seven potential scoring procedures, the *product: square root of difference* approach evidenced the best performance, particularly when one of the data sets consisted of extreme scores. Consequently, we used this scoring approach in subsequent studies with data collected from human participants.

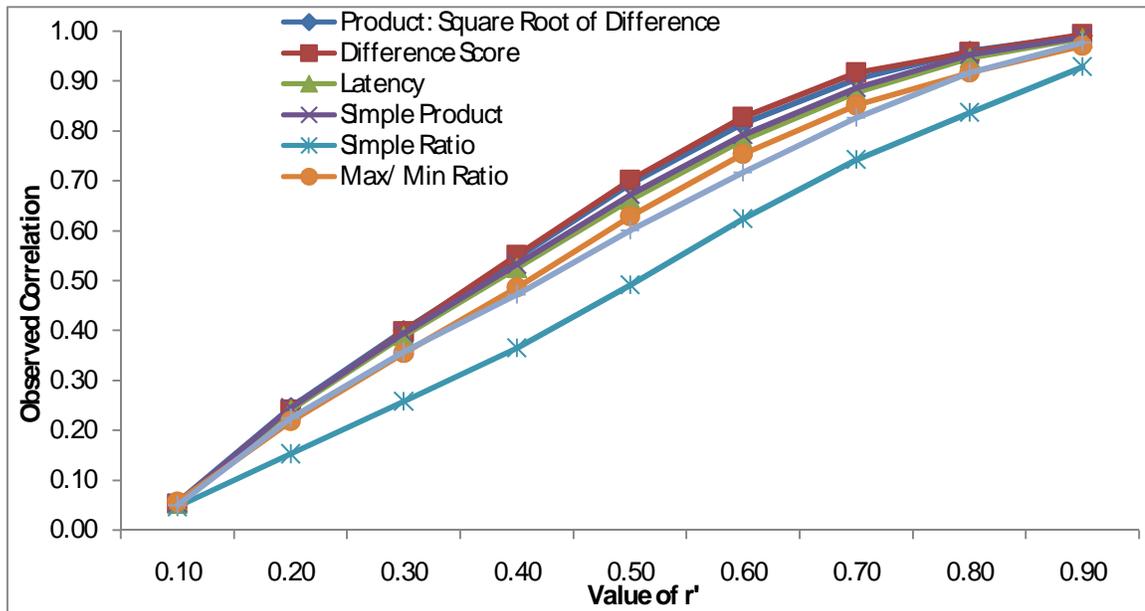


Figure 2. Comparison of proposed procedures for analyzing paper-format IAT data. Figure shows the correlation between two functionally related data sets - D_1 and D_2 - at different levels of correspondence between the two (r'). Higher levels of r' indicate stronger initial relationships between D_1 and D_2 .

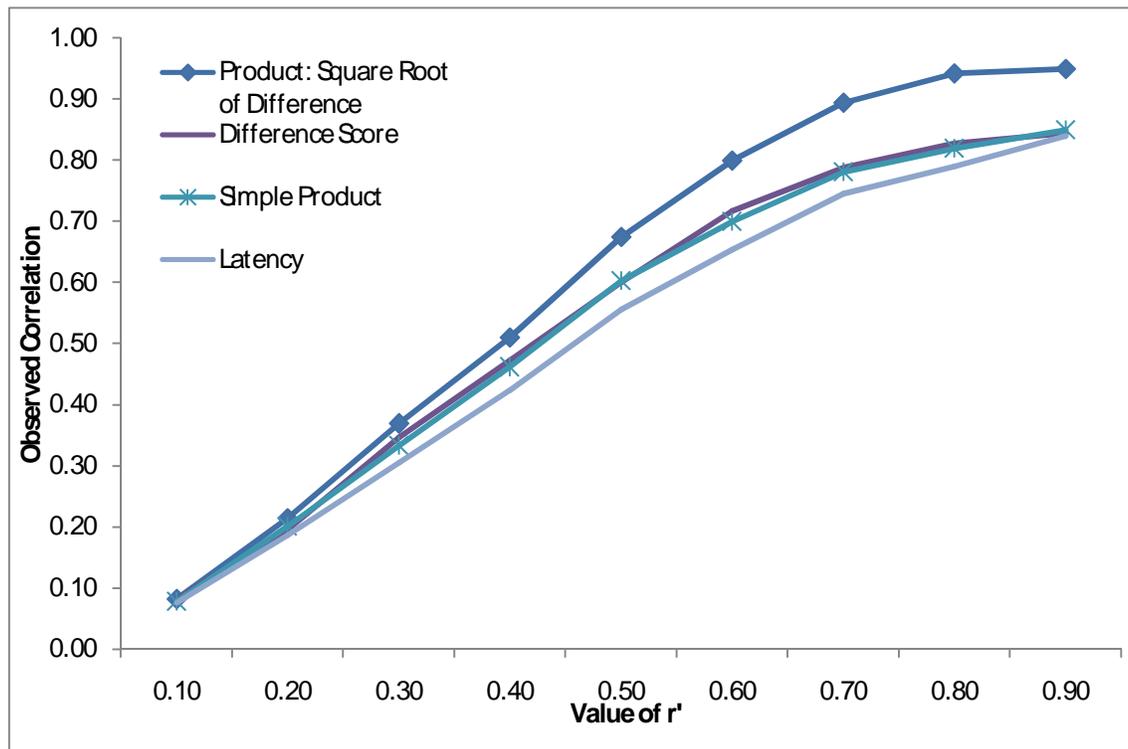


Figure 3. Comparison of proposed procedures for analyzing paper-format IAT data for extreme scores. Figure shows the correlation between two functionally related data sets - D_1 and D_3 - at different levels of correspondence between the two (r'). Higher levels of r' indicate stronger initial relationships between D_1 and D_3 . This data set consisted of extreme scores, with the larger value of the two data points increased by two standard deviations.

STUDIES 2A AND 2B

Paper-format IATs have been used in previous research with the tacit assumption that the psychometric properties of the measure are similar to computer-based IATs. The purpose of Studies 2a and 2b was to assess systematically whether results from a paper-format IAT would mimic those of a computer-based IAT when using the same stimulus items. Participants completed paper- and computer-based race IATs (Black-White/pleasant-unpleasant) with names or pictures for the race dimension and words for the pleasant-unpleasant dimension. Test-retest reliability was assessed after a one-week interval. If the paper-format IAT is to be a viable supplement to the computerized version, then they should show similar patterns of racial attitudes, be correlated with one another, and have comparable levels of test-retest reliability.

STUDY 2A

Method

Participants

Fifty-three female and 20 male undergraduates (aged 18-40, $M = 20.8$, $SD = 4.1$; 75% White) completed Session 1 in exchange for partial course credit; 19 men and 50 women returned for Session 2. Data from two women who reported not being fluent in English were omitted from further analysis. Due to a printing error on the measure, paper IAT responses from three men and 19 women were unusable, and due to computer failure, computer IAT data were unusable for two men and 10 women, resulting in unequal N s for different analyses.

Materials

Black and White names and pleasant and unpleasant word stimuli were selected from Greenwald et al. (1998). Race stimuli for the picture IATs were positive expression color photographs of African American and European American faces (four men and four women of each race).⁴ The individuals pictured all had similar hair color and were matched on attractiveness by a sample of college students.

PROCEDURE

Experimenters were White undergraduate students. Participants were seated in individual laboratory rooms, with a maximum of three participants (in separate rooms) per session. Room doors were left open and the experimenter stood in the hall where participants could clearly hear him or her deliver verbal instructions for the paper-format IAT. Participants first completed a flower/insect training IAT followed by the critical race attitude paper IAT measures in counterbalanced order. The experimenter then instructed participants to begin working on the computer and closed the individual lab room doors. Participants completed a computer format flower/insect training IAT followed by the race attitude computer IATs in counterbalanced order. Following the first session, participants returned approximately one week later and completed the same measures again. At the end of Session 1 only, participants completed a demographic form that assessed their gender, race, and English fluency.

Computer-based IATs. Each of the two race IAT measures consisted of two 60-trial blocks, one in which White was paired with pleasant and Black with unpleasant, and one in which Black was paired with pleasant and White with unpleasant. In a deviation from the standard IAT procedure (Nosek et al., 2006), there were no single-category practice blocks. This has the benefit of making the format of the computer IAT congruent with the format of the paper IAT, and the potential risk of undermining the computer IAT's psychometric properties. The standard version is examined in Study 2. For analysis purposes, the first 20 trials of each block were considered practice trials and the remaining 40 trials were considered critical.

⁴ Photos were senior year high-school photos used with permission from a 2000 high-school yearbook.

Paper-based IATs. Each paper IAT consisted of two critical blocks. Each block was presented on a single page, with trials alternating between the race and valence dimensions (see Figure 1). The experimenter stood in a hallway outside the experimental rooms where he or she could be heard clearly but could not see the participants' responses. Reading from a script, the experimenter orally instructed participants to categorize each stimulus item by making a check mark in the appropriate circle to the left or right of the item, beginning in the upper left and working down, then over to the second column. They were told to avoid making mistakes, but if they did make a mistake to keep going. Participants were given 20 seconds per page to make as many categorizations as possible. Prior to each page, the experimenter instructed participants to look at the category labels at the top of the page for three seconds. The experimenter then said, "Ready, set, go" and started a timer set for 20 seconds. The procedure was repeated for each block.

RESULTS

Preparation of Data for Analysis

Computer IAT data. In accordance with the strategy described by Greenwald et al. (2003), differences between the two IAT blocks were divided by the standard deviation of all trials, separately for the practice and critical trials. Their average yielded a *D* score. Among the 18 men and 43 women in the final data set, the overall error rate was 5.8%, with slightly more errors recorded in Session 2 (6.1%), than Session 1 (5.4%), $F(1, 58) = 3.76, p = .06, d = 0.25$.⁵ As expected, collapsing across both sessions, participants made significantly more errors in the *White+unpleasant* pairing condition (7.0%) than the *White+pleasant* pairing condition (4.8%), $F(1, 63) = 24.00, p < .001, d = 0.62$.

Paper IAT data. Although overall error rates were low, participants had a significantly higher error rate in Session 1 (3.4%) than Session 2 (0.7%), $F(1, 44) = 4.07, p < .05, d = 0.30$. Participants with more than 20% errors in any session (one White male and one Hispanic female) were omitted from further analysis. In addition, we set a criterion that only blocks with at least eight correct responses would be retained for analysis, which resulted in the deletion of one IAT for three participants. Among the remaining participants, collapsing across both sessions, participants made significantly more errors in the *White+unpleasant* pairing condition (1.5%) than in the *White+pleasant* pairing condition (0.7%), $F(1, 42) = 6.62, p < .05, d = 0.39$. IAT attitude scores were computed on correct responses for each participant using the *product: square root of difference* algorithm described in Study 1.

DEMONSTRATION OF IMPLICIT ATTITUDE

Computer IATs with name and picture stimuli showed the same overall pattern of responses and were significantly correlated, $r(58) = .54, p < .001$ for Session 1 and Session 2, $r(52) = .45, p < .001$. As predicted, responses were significantly faster under the

⁵ Effect sizes for all two-group, repeated-measures comparisons are reported as Cohen's *d*, which is computed as $d = t/\sqrt{df}$. Cohen (1988) provides guidelines of $d = .2$ as a small effect, $d = .5$ as a medium effect, and $d = .8$ as a large effect for behavioral science research.

White+pleasant response pairing than the *Black+pleasant* pairing, indicating an implicit preference for White people relative to Black people. As shown in Table 1, the effect size for this difference was moderately strong, with Cohen's *d* ranging from 0.51 to almost 0.77.

Table 1. Study 2a: Computer and Paper Format IAT Scores

Computer IAT

	IAT D score (<i>SD</i>)	<i>t</i> for difference		Cohen's <i>d</i>
		from zero	<i>p</i>	
Name Stimuli				
Session 1	.27 (.39)	5.12	.000004	0.68
Session 2	.24 (.33)	5.45	.000002	0.77
Picture Stimuli				
Session 1	.21 (.42)	3.86	.0003	0.51
Session 2	.23 (.32)	4.85	.00002	0.69

Paper IAT

	IAT score (<i>SD</i>)	<i>t</i> for difference		Cohen's <i>d</i>
		from zero	<i>p</i>	
Name Stimuli				
Session 1	1.62 (2.52)	4.07	.0002	0.64
Session 2	.68 (2.31)	1.87	.07	0.29
Picture Stimuli				
Session 1	-.28 (3.07)	-0.57	.57	0.09
Session 2	-.19 (2.14)	-0.54	.59	0.09

Note. Computerized IAT scores (D) are calculated according to the recommendations of Greenwald et al. (2003). Paper format IAT scores are computed using the *Product: square root of difference* algorithm. Cohen's *d* for one-sample t-tests were computed as the absolute value of the sample mean divided by the sample standard deviation.

The relationship between IATs with name and picture stimuli was weaker for paper-format IATs: $r(43) = .28$, $p = .07$ for Session 1 and $r(42) = .45$, $p < .01$ for Session 2. Importantly, the pattern of results for the name and picture versions was markedly different. On the paper IAT with name stimuli, participants demonstrated a pro-White attitude, but the effect size was smaller than that observed with the computer IAT and was only significant in Session 1. Unexpectedly, the predicted pro-White attitude was not observed on the paper IAT with picture stimuli. Participants showed no difference in their latency to respond when White was paired with pleasant versus when Black was paired with pleasant. This suggests that the paper format IAT may not be as powerful a measure for detecting effects as the computer format, and the paper format may be more sensitive to the type of stimuli used in the task.

Test-Retest Reliability

Reliability between sessions separated by one week was similar to that observed in previous, computer-based IAT research (e.g., Nosek et al., 2006). For the computer IAT, attitude scores for Session 1 and Session 2 were correlated, $r(49) = .53$, $p < .001$ for names

and $r(49) = .46, p < .001$ for pictures. Test-retest reliability for the paper IATs was in line with that of the computerized task: $r(41) = .62, p < .001$ for names and $r(40) = .49, p < .001$ for pictures. These correlations, although small by standards for explicit measures, are consistent with reliability coefficients from administrations of the computerized IAT: In a summary of 23 data collections that have included IAT measurements on more than one occasion, mean and median test-retest reliabilities were .50 (Lane et al., 2007).

Relationship between Paper and Computer IAT

For the session in which the IAT effect was strongest (Session 1), the correlation between the paper and computer IATs with name stimuli was also moderately strong, $r(30) = .43, p < .05$. However, in Session 2, in which the IAT effect with names was weaker, the correlation between the computer and paper versions was also weak, $r(26) = .19, ns$. For the IATs with picture stimuli, the paper-computer correlations were $r(30) = .27, ns$ in Session 1 and $r(26) = .48, p < .05$ in Session 2.

Comparing Scoring Procedures

We examined the efficacy of each of the proposed paper IAT scoring strategies. Results from actual data corroborated the analyses of simulated data sets, offering support for use of the *product: square root of difference* algorithm for analyzing paper IAT data. We assumed that a better scoring approach would maximize the relationship between computerized and paper-pencil measures of the same construct. For IATs with name stimuli, computer-paper correlations using the *product: square root of difference* algorithm were approximately equal or higher in both sessions than those using any of the six alternate algorithms ($r = .43$ versus .31 to .40 for other algorithms in Session 1 and $r = .19$ versus .11 to .20 for Session 2). For IATs with picture stimuli, computer-paper correlations in Session 1 were equal or higher using the *product: square root of difference* algorithm compared with all other algorithms ($r = .27$ versus .15 to .27 in Session 1). In Session 2, the *product: square root of difference* algorithm produced computer-paper correlations that were stronger than three of the other algorithms, $r = .48$ versus .43 to .47; correlations were higher using the *simple difference* ($r = .50$), *simple product* ($r = .52$) and *product: squared ratio* ($r = .57$) algorithms. However, of the candidate scoring procedures, no alternative was consistently superior to the *product: square root of difference*.

Effect of Potential Artifacts on Paper IAT Scores

Order of IAT block presentation. The order in which participants complete the two IAT blocks (i.e., whether participants complete flower+good or flower+bad first) is one artifact of the procedure that has consistently been shown to influence the overall IAT effect with computer-based measures (Mierke and Klauer, 2003; Nosek, Greenwald, and Banaji, 2005). Contrary to most research with computer IATs, the typical IAT order effect was not observed with paper IAT measures in the present study. Order of presentation of the IAT block was not

a significant factor for the paper IAT with name stimuli in Session 1, $F(1, 38) = 0.44$, *ns*, $d = 0.11$, or Session 2, $F(1, 39) = 0.48$, *ns*, $d = 0.11$. There was a significant order effect for the paper IAT with picture stimuli in Session 1, $F(1, 38) = 15.78$, $p < .001$, $d = 0.64$. Surprisingly, this strong order effect was in the opposite direction of what is commonly observed, with participants who completed the Black+pleasant pairing first demonstrating a strong pro-White attitude ($M = 1.37$, $SD = 2.97$) and participants who completed the White+pleasant pairing first demonstrating a strong pro-Black attitude ($M = -1.92$, $SD = 2.22$). This unexpected reverse order effect did not appear in Session 2, where there was no significant difference between the two order conditions, $F(1, 39) = 0.00$, *ns*.⁶ It is possible that categorization of the faces was difficult initially, but became easier with repeated administrations. This may account for the lack of consistent effects with the picture version.

Overall response speed. Original scoring procedures of the computerized version of the IAT were susceptible to differences in overall response speed: slower participants tended to exhibit larger overall IAT effects. This artifact of average response time was one of the motivations for developing newer scoring procedures for the computerized IAT (Greenwald et al., 2003). We examined whether overall differences in response speed would also influence the magnitude of IAT effects on the paper-format version. For each IAT in the study, the average number of items completed across the two blocks of each paper-format IAT was calculated to provide an index of general speed of response. This value was then correlated with participants' calculated IAT scores. For the IATs with name stimuli, the total number of items completed was uncorrelated with the strength of the IAT effect in Session 1 ($r = -.03$, *ns*) and Session 2 ($r = .08$, *ns*). Similar results were observed for IATs with picture stimuli, with low and nonsignificant correlations in Session 1 ($r = .08$, *ns*) as well as Session 2 ($r = .16$, *ns*). The magnitude of the correlation between total number of items completed and IAT score was higher in most instances when alternate algorithms were used to compute IAT scores (*rs* ranging from .01 to .31). No single alternate algorithm consistently outperformed the *product: square root of difference* algorithm in minimizing the strength of this relationship.

SUMMARY

Results of Study 2a provide support for the use of a paper IAT with all verbal stimuli; however, the paper IAT with picture stimuli for the race dimension was not effective. One concern is that the pictures that were used for the race stimuli, which were color photos of attractive young people with a positive expression, may not have been maximally effective for assessing racial attitudes, at least on the paper IAT. Although we cannot be certain that this is the case, it is possible that participants did not even need to process the pictures as human faces in order to categorize them as Black or White people – these categorizations could have been made entirely on the basis of ink color.

⁶ These artifacts were also tested for the computer IAT. Typical order effects were observed on the computer IAT, with significantly larger IAT effects observed when the White+pleasant block was completed first compared with when the Black+pleasant block was completed first, across both IAT measures and both sessions (all $ps < .05$). Correlations between average response latency and magnitude of IAT effect were non-significant. Similar results were observed in Study 2b.

In Study 2b, we replaced the color photos with black and white photos of people with neutral expressions, using the same stimulus items used in previous IAT research (e.g., Cunningham, Preacher, and Banaji, 2001; Nosek, Banaji, and Greenwald, 2002). As these pictures use similar ink colors for both categories (light and dark shades of gray), participants must process the faces more thoroughly to identify features that distinguish White and Black faces. In addition, practice blocks were added to the computer IATs to parallel the more typical design for computer IATs (Greenwald et al., 1998).

STUDY 2B

Method

Participants

Twenty-eight male and 57 female undergraduates (aged 18-26, $M = 19.9$, $SD = 1.9$; 92% White) completed Session 1 in exchange for partial course credit; 26 men and 54 women returned for Session 2. Data from two men who reported not being fluent in English were omitted from further analysis.

Materials

IAT measures used the same word and name stimuli as Study 2a. Race stimuli for the picture IATs were neutral expression, closely cropped black and white photographs of adult African American and European American faces (three men and three women of each race).⁷

PROCEDURE

Experiment 2b followed the same overall procedure as Experiment 2a. Experimenters for the study were White undergraduate students. Participants were placed in individual lab rooms, with a maximum of three participants in a session. Experimenters delivered verbal instructions for the paper-format IAT from the hall outside the lab rooms. Participants completed the paper-format IAT measures prior to the computer IAT measures. Both the paper- and computer-format IAT sets began with a flower/insect training IAT, followed by the critical race attitude IATs in counterbalanced order. Participants returned approximately one week after Session 1 and completed the same measures again. Demographic information (gender, race, and English fluency) was assessed following Session 1 only.

Computer-based IATs. Each of the two critical IAT measures consisted of five blocks; only data from blocks 3 and 5 are used in the analysis. The first two blocks were 20-trial single-category practice blocks of Black-White and pleasant-unpleasant judgments. The third block was a 42-trial block with judgments from both categories using the key mappings previously practiced (e.g., *Black+pleasant* and *White+unpleasant*). A 20-trial block of pleasant and unpleasant judgments followed, with key mappings reversed relative to the

⁷ The race photos used are identical to those employed in the race attitude task on the IAT website, available at <https://implicit.harvard.edu/>.

preceding blocks. In the final 42-trial block, participants categorized stimuli from the two dimensions using response key mappings reversed from the previous critical block (e.g., *Black+unpleasant* and *White+pleasant*). In blocks 3 and 5, for analysis purposes, the first 14 trials are considered practice trials and the subsequent 28 trials are considered critical. Incorrect responses were indicated with a red X and participants were required to correct their response before advancing. Within the critical blocks, trials alternated between the pleasant-unpleasant and Black-White dimensions.

Paper-based IATs. The paper IATs with name stimuli were identical to those used in Study 2a. The paper IATs with picture stimuli were modified to include the black and white photos described above.

RESULTS

Preparation of Data for Analysis

Computer IAT data. Data for the computer IATs were prepared following the same conventions used in Study 2a. The overall error rate was 5.4%. Replicating Study 2a, participants made significantly more errors in Session 2 (6.0%), than Session 1 (5.0%), $F(1,78) = 8.30, p < .01, d = 0.33$. Also replicating Study 2a, participants made significantly more errors in the *White+unpleasant* pairing condition (7.0%) than the *White+pleasant* pairing condition (4.3%), $F(1, 84) = 62.12, p < .001, d = 0.86$.

Paper IAT data. Data for the paper IAT were prepared in the same way as Study 2a. Replicating Study 2a, participants had a significantly higher error rate in Session 1 (6.0 %) than Session 2 (1.7%), $F(1, 78) = 14.71, p < .001, d = 0.43$, partly due to several participants with unusually high error rates in Session 1 not returning for Session 2. Data from sessions in which participants made more than 20% errors were omitted from further analysis (resulting in deletion, from both sessions, of data provided by one White woman and Session 1 data only for three White men and four White women). Critical blocks in which fewer than 8 correct responses were made (8.1% of all critical blocks) were omitted from further analysis. Replicating Study 2a, participants remaining in the sample made significantly more errors across both sessions in the *White+unpleasant* pairing condition (3.2%) than in the *White+pleasant* pairing condition (2.3%), $F(1, 70) = 5.55, p < .05, d = 0.28$.

Demonstration of Implicit Attitude

Results from Study 2a were replicated for the computer IATs with name and picture stimuli, with participants indicating strong pro-White attitudes. As shown in Table 2, these effects were very strong, with Cohen's d values near 1.0 for the picture IAT and in excess of 1.0 for the name IAT. The name and picture versions of the IAT were moderately correlated, $r(80) = .43, p < .001$ in Session 1 and $r(80) = .26, p < .05$ in Session 2.

Pro-White implicit attitudes were also demonstrated on the paper IATs with both name and picture stimuli. As in Study 2a, the effect size was smaller with the paper IAT than with the computer IAT. However, in Study 2b, participants demonstrated significant pro-White implicit attitudes, with Cohen's d ranging from 0.26 to 0.62. Correspondence between name and picture IATs was similar to the computerized task: $r(68) = .38, p < .01$ in Session 1 and $r(74) = .36, p < .01$ in Session 2.

Table 2. Study 2b: Computer and Paper Format IAT Scores

<i>Computer IAT</i>				
	IAT D score (<i>SD</i>)	<i>t</i> for difference from zero	<i>p</i>	Cohen's <i>d</i>
Name Stimuli				
Session 1	.46 (.38)	11.07	6.09 x 10 ⁻¹⁸	1.22
Session 2	.32 (.31)	9.00	1.17 x 10 ⁻¹³	1.02
Picture Stimuli				
Session 1	.36 (.36)	9.14	3.84 x 10 ⁻¹⁴	1.00
Session 2	.29 (.32)	7.93	1.39 x 10 ⁻¹¹	0.90
<i>Paper IAT</i>				
	IAT score (<i>SD</i>)	<i>t</i> for difference from zero	<i>p</i>	Cohen's <i>d</i>
Name Stimuli				
Session 1	1.55 (2.48)	5.18	.000002	0.62
Session 2	1.38 (2.24)	5.23	.000002	0.62
Picture Stimuli				
Session 1	.79 (2.48)	2.63	.01	0.32
Session 2	.50 (1.92)	2.26	.03	0.26

Note. Computerized IAT scores (D) were calculated according to the recommendations of Greenwald et al. (2003). Paper format IAT scores were computed using the *Product: square root of difference* algorithm. Cohen's *d* for one-sample t-tests were computed as the absolute value of the sample mean divided by the sample standard deviation.

Test-Retest Reliability

The computer IAT showed significant test-retest reliability, $r(78) = .45$, $p < .001$, for the name IAT, and $r(78) = .41$, $p < .001$ for the picture IAT. Reliability was similar for the paper IAT with name stimuli, $r(64) = .49$, $p < .001$. The paper IAT with picture stimuli, which showed the smallest overall IAT effect, was not significantly correlated between Sessions 1 and 2, $r(64) = .21$, $p = .10$.

Relationship between Paper and Computer IAT

Name-based paper and computer IATs were significantly correlated: $r(70) = .36$, $p = .01$ for Session 1 and $r(73) = .40$, $p = .001$ for Session 2. In contrast, with picture stimuli, paper-computer correlations were considerably weaker: $r(68) = .17$, $p = .17$ for Session 1 and $r(76) = .23$, $p < .05$ for Session 2.

Comparing Scoring Procedures

As in the prior study, results supported use of the *product: square root of difference* algorithm, which maximized the relationship between computerized and paper-based versions of the task. For IATs with name stimuli, correlations between computer and paper IATs were higher when the *product: square root of difference* algorithm was used compared with any of the six alternate algorithms ($r = .36$ versus $.28$ to $.35$ for other algorithms in Session 1 and $r = .40$ versus $.34$ to $.39$ for Session 2). For IATs with picture stimuli, all correlations were stronger with the *product: square root of difference* algorithm compared with other

algorithms in Session 1 ($r = .17$ versus $.12$ to $.14$ for other algorithms). In Session 2, the *product: square root of difference* algorithm produced an equal or stronger paper-computer correlation ($r = .23$) than four of the alternate algorithms (r s from $.20$ to $.23$). The only exceptions were modestly stronger correlations in Session 2 using the *max-min ratio* conversion ($r = .26$) and the *latency conversion* algorithm ($r = .30$). As in Study 2a, no alternative was consistently superior to the *product: square root of difference*.

The Paper-Format and Computer Versions of the IAT Assess a Single Construct

Due to the relatively high degree of noise in response latency data, zero-order correlations with response-latency measures such as the IAT may be attenuated due to the low reliability of the measures. Cunningham et al. (2001) demonstrated that Structural Equation Modeling (SEM) can be used to separate variance due to latent attitude constructs from variance due to error inherent in response latency measurements, resulting in substantially improved correlations among measures. Analyzing IATs with name and picture stimuli separately, we fit latent variable models in which the paper IAT measures for Session 1 and 2 loaded onto a single latent factor that was correlated with a second latent factor assessed by computer IAT measures for Session 1 and 2. We then compared the two-factor models to models in which all IAT measures loaded onto a single latent factor. Model fit was assessed following guidelines proposed by Hu and Bentler (1999), who suggest that adequate model fit is indicated by standardized Root Mean Square Residual (sRMR) values less than $.08$, Non-Normed Fit Index (NNFI) and Comparative Fit Index (CFI) greater than $.95$, and Root Mean Square Error of Approximation (RMSEA) less than $.06$.

Name IATs. A model fitting paper and computer IAT scores onto separate, correlated latent factors was a moderately good fit to the data, $\chi^2(1) = 1.39$, $p = .24$, sRMR = $.026$, NNFI = $.943$, CFI = $.991$, RMSEA = $.08$. The correlation between the latent factors was $r = .78$, indicating a very strong relationship between the paper and computer versions of the IAT.

To determine whether scores on the paper IAT and computer IAT represent a single underlying factor, a second model was fit to the data in which all four measures (paper and computer IAT scores in Session 1 and 2) loaded onto a single latent factor. A nested chi-square test demonstrated that the fit of this model was not significantly different from the two-factor model described above, $\Delta\chi^2(1) = 1.95$, ns , suggesting that the simpler and more parsimonious representation that the paper and computer measures assess the same construct can be retained.

Picture IATs. Results for IATs with picture stimuli were similar to those with name stimuli. A model fitting paper and computer IAT scores onto separate, correlated latent factors was a very good fit to the data, $\chi^2(1) = .345$, $p = .56$, sRMR = $.019$, NNFI = 1.000 , CFI = 1.000 , RMSEA = $.000$. The correlation between the latent factors was $r = .51$, indicating that the paper and computer versions of the IAT were related. A second model with all four measures loading onto a single latent factor did not provide significantly different fit from the two-factor model described above, $\Delta\chi^2(1) = 1.53$, ns , consistent with the claim that the paper and computer IATs are assessing the same underlying construct.

Effect of Potential Artifacts on Paper IAT Scores

Order of IAT block presentation. Block order effects for the paper IAT were similar to those observed in Study 2a: implicit attitudes as measured by the IAT did not seem to be affected by the order of presentation of IAT blocks for name stimuli in Session 1, $F(1, 71) = 1.63$, *ns*, $d = 0.15$, or Session 2, $F(1, 73) = 0.73$, *ns*, $d = 0.10$. A significant order effect was observed for the paper IAT with picture stimuli in Session 1, $F(1, 69) = 10.60$, $p < .01$, $d = 0.39$. Surprisingly, this order effect was in the same reverse direction as that observed in Study 2a, with participants who completed the *Black+pleasant* pairing first demonstrating a larger IAT effect ($M = 1.63$, $SD = 2.58$) than participants who completed the *White+pleasant* pairing first ($M = -.18$, $SD = 2.01$). This unexpected reverse order effect did not appear in Session 2, where there was no significant difference between the two order conditions, $F(1, 76) = 1.42$, *ns*, $d = 0.14$.

Overall response speed. Replicating Study 2a, the average number of items completed across the two blocks of each paper-format IAT was uncorrelated with participants' calculated IAT scores. For the IATs with picture stimuli, the total number of items completed was uncorrelated with the strength of the IAT effect in Session 1 ($r = .02$) and Session 2 ($r = -.02$). IATs with name stimuli showed a similar pattern of weak correlations in both Session 1 ($r = .15$) and Session 2 ($r = .17$). Also replicating Study 2a, these correlations were generally higher when alternate algorithms were used to compute IAT scores (r s ranging from .001 to .22), and no single alternate algorithm produced consistently lower correlations compared with the *product: square root of difference* algorithm.

To supplement the present finding that overall response speed did not produce a meaningful artifact on the paper-format IAT, we also examined the effect of overall response speed in another available data set (Lane et al., 2005). These results supported the conclusion that there was no systematic relationship between participants' general response speed and their IAT scores. In this study, participants completed six paper-format IATs. Three measures of overall response speed were calculated: the average number of items completed in the two blocks of each IAT, the average number of items completed on all 12 blocks of the IATs in the study, and the number of items completed on the two blocks of a single, non-social (flower relative to insect) IAT.⁸ As seen in Table 3, IAT effects, as calculated by the suggested scoring procedure, were unrelated to the average number of items completed in the two blocks of the IAT (average $r = .03$) and the average number of items completed on all IATs in the study (average $r = .02$). Additionally, five social IATs (measures of university and school dormitory attitude and identity, and self-esteem) were unrelated to the number of items completed on the individual blocks of an unrelated, non-social (flower) IAT (average $r = .01$).

Although these findings suggest that overall response speed does not produce a large artifact in paper-format IAT effects, it is worth noting that all of the participants were drawn from student populations and were likely familiar and comfortable with completing rapid or timed tasks. It is possible that an overall effect of speed may emerge among different populations that have notably slower overall response speeds (e.g., Hummert et al., 2002).

⁸ These measures were highly correlated with one another, indicating that participants who were faster overall did tend to complete more items across all 12 blocks of the six IATs. For example, the average correlation between the mean number of items completed on the 12 IAT blocks and the mean number completed on each IAT was $r = .80$.

Table 3. Effect of Overall Response Latency on Paper-Format IAT Effects: Correlations Between Four Measures of General Response Speed and Paper-Format IAT Effects (Reanalysis of Lane et al. [2005])

	Average on Relevant IAT	Average on All IATs	Flower+ Good	Flower+Bad
University Attitude	-0.08	-0.02	0.02	0.00
University Identity	0.00	0.05	-0.07	0.06
Self Esteem	-0.03	0.06	0.05	0.09
Dorm Attitude	-0.05	-0.12*	-0.05	0.08
Dorm Identity	-0.01	-0.01	-0.05	0.00
Flower Attitude ⁽¹⁾	0.34**	0.16*	⁽¹⁾	⁽¹⁾
<i>Average</i>	0.03	0.02	-0.02	0.05

Note. Participants completed six paper-format IATs, listed in the left-most column. IAT effects were calculated using the *product: square root of difference* scoring approach. Average on relevant IAT = Average number of items completed on that particular IAT. Average on All IATs = Average number of items completed on all 12 blocks of the six IATs. Flower+Good = Total number of items completed on the *flower+good* pairing of the flower/insect paper-format IAT. Flower+Bad = Total number of items completed on the *flower+bad* pairing of the flower/insect paper-format IAT. *Ns* ranged from 247 to 289. * $p < .05$. ** $p < .0001$.

⁽¹⁾ The flower IAT was designed to familiarize participants with the task in a large group. As such, it was the only task not presented in a counterbalanced order: all participants completed the *flower+good* block prior to the *flower+bad* block of the IAT. Flower attitude effects were a direct function of number of items completed on each IAT block and should be highly correlated with the IAT effects.

SUMMARY

Study 2b largely replicated results of Study 2a, supporting use of a paper-format IAT with all verbal stimuli. This task was positively related to a computerized measure of the same task, and showed comparable patterns of attitudes and test-retest reliability to computerized IATs. Although the paper IAT with black and white, neutral expression pictures showed more similar main effects to the computerized task than the version used in Study 2a, test-retest reliability and correspondence with the computerized IAT were still disappointingly small.

The weak zero-order correlations between paper and computer IATs may be due, in part, to the substantial error variance associated with those scores. Error variance is a particular concern with any response latency measure, and the paper IAT is no exception. SEM analyses, which separate measurement error from latent construct variance, provided support that the computer and paper IATs are strongly correlated and they may be assessing the same underlying construct. At the same time, the data suggest that the paper-format IAT is not quite as effective at eliciting strong mean effects as the computer-format IAT. In addition,

both verbal and picture versions demonstrated that the paper IAT is largely unaffected by common IAT artifacts such as task order and overall response speed.

CONCLUSION

Study 1 compared several potential scoring procedures and demonstrated that the *product: square root of difference* approach, calculated as $(X/Y) \times \text{Square Root of } (X - Y)$, where X is the greater of A or B, and Y is the smaller of A or B, maximized the relationship between functionally related data sets. Data from human participants in Studies 2a and 2b also showed that this scoring approach was consistently superior at maximizing the relationship between computerized and paper-format versions of the same task.

For the IAT with names, correlations between the computer and paper versions were equal to or higher using the *product: square root of difference* method compared with any of the six other proposed algorithms. For the IAT with pictures, computer-paper correlations were higher with the *product: square root of difference* algorithm in most instances, and it was the only algorithm that consistently outperformed the remaining algorithms.

Studies 2a and 2b demonstrated that a paper-format IAT using names for the race categories exhibits similar psychometric properties to a traditional computer IAT with the same stimuli. Test-retest reliabilities for the paper format IAT with names were comparable to computer IAT reliabilities and the paper IAT with names was significantly correlated with the computer IAT in three of four replications. However, results with two versions of a paper-format IAT using pictures for the race dimension were equivocal. This version of the task did not show the expected strong pro-White implicit attitude (weak in Study 2b and absent in Study 2a), exhibited poor test-retest reliability and only weak correspondence with computerized versions of the same task.

Although the paper IATs with picture stimuli showed weak or no effects, computer IATs with identical pictures showed robust pro-White/anti-Black bias. Thus, it appears that pictures are effective stimulus items for a computerized IAT but not for a paper-format IAT, at least in this particular context. Although speculative, one explanation for this may be the simultaneous visibility of all stimuli on a paper-format IAT. On the computer IAT, one stimulus item is shown at a time and a new item appears only after a participant has recorded a response. On the paper IAT, all of the stimuli are shown on the page simultaneously. Although participants are instructed to categorize each item individually, the presence of other stimuli in the visual field during categorization is unavoidable. This may not be a concern with word stimuli because participants most likely attend to each item individually since each word needs to be read and processed before it can be categorized. However, with pictures, participants are able to process multiple pictures in the visual field simultaneously; thus judgment of the target picture may be influenced by nearby pictures that are not the main focus of judgment (Mausfeld, 2003). This may have been a particular problem in Study 2a because the pictures were printed in color, thus emphasizing visual differences between Black and White faces.

The current studies support the use of paper-format IATs with verbal stimuli, analyzed using the *product: square root of difference* algorithm. Our results replicated those of Lowery et al. (2001), who demonstrated anti-Black implicit attitudes using a paper IAT with name

stimuli. They also parallel those of Teachman et al. (2003), who demonstrated anti-fat bias using a paper IAT with word stimuli, and Lane et al. (2005), who used paper IATs to measure college identity, attitude, and self-esteem using verbal stimuli for all dimensions.

Paper-format IATs allow researchers to assess implicit attitudes under circumstances where traditional computer IAT data collection may be impractical or even impossible. We have administered it at beaches, where sand, sun and water pose risks to computers, and in large groups. Under circumstances such as these, where computerized data collection is not feasible, paper format IATs may provide a satisfactory supplement to traditional IATs. However, because the accuracy of timing is substantially less sensitive with paper IATs, and because the effects tend to be smaller than with computer data collection, paper format IATs are unlikely to serve as a substitute for computer IATs in most situations. Paper IATs also do not allow for trial by trial timing, thus it is not possible to compute split-half or alpha reliability. In addition, the computer IAT has been used extensively; numerous studies have explored its psychometric properties; and much more is known about its potential artifacts. Thus, researchers are advised to use computer IAT data collection for laboratory-based research.

Although computer-based data collection is preferable, there are many circumstances in which it is not practical or even possible. When there is evidence that findings from a particular paper-format IAT converge with those of a computer-based counterpart, paper format IATs should prove useful.

AUTHORS' NOTE

This research was supported by R-01 MH68447 from the National Institutes of Health and REC-0634041 from the National Science Foundation to the last author. The authors thank Jonathan Dahl and Rebekah Mitchell for assistance with data collection.

REFERENCES

- Blair, I.V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6, 242-261.
- Cai, H., Sriram, N., Greenwald, A.G., and 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.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Erlbaum.
- Cunningham, W.A., Preacher, K.J., and Banaji, M.R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, 12, 163-170.
- Greenwald, A.G., Banaji, M.R., Rudman, L.A., Farnham, S.D., Nosek, B.A., and Mellott, D.S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3-25.

- Greenwald, A.G., McGhee, D.E., and Schwartz, J. (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., and Banaji, M.R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.
- Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., and Banaji, M.R. (in press). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*.
- Hu, L.T., and Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.
- Hummert, M.L., Garstka, T.A., O'Brien, L.T., Greenwald, A.G., and Mellott, D.S. (2002). Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging*, *17*, 482-495.
- Lane, K.A., Banaji, M.R., Nosek, B.A., and Greenwald, A.G. (2007). Understanding and using the Implicit Association Test: IV. What we know (so far) about the method. In B. Wittenbrink and N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59-102). New York: Guilford Press.
- Lane, K.A., Mitchell, J.P., and Banaji, M.R. (2005). Me and my group: Cultural status can disrupt cognitive consistency. *Social Cognition*, *23*, 353-386.
- Lemm, K.M. (2000). *Personal and social motivation to respond without prejudice: Implications for implicit and explicit attitude and behavior*. Unpublished doctoral dissertation, Yale University.
- Lowery, B.S., Hardin, C.D., and Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842-855.
- Mausfeld, R. (2003). Conjoint representations and the mental capacity for multiple simultaneous perspectives. In H. Hecht, R. Schwartz, and M. Atherton (Eds.), *Looking into pictures: An interdisciplinary approach to pictorial science* (pp. 17-60). Cambridge, MA: MIT Press.
- Mierke, J., and Klauer, K.C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology*, *85*, 1180-1192.
- Nosek, B.A., Banaji, M.R., and 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., and Banaji, M.R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*, 166-180.
- Nosek, B.A., Greenwald, A.G., and 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). Hove, UK: Psychology Press.
- 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., and Banaji, M.R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36-88.

- Sinclair, S. Lowery, B.S., Hardin, C.D., and Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583-592.
- Sriram, N., Greenwald, A.G., and Nosek, B.A. (2007). *Correlational biases in mean response latency differences*. Unpublished manuscript.
- Teachman, B., and Brownell, K. (2001). Implicit associations toward obese people among treatment specialists: Is anyone immune? *International Journal of Obesity*, *25*, 1525-1531.
- Teachman, B.A., Gapinski, K.D., Brownell, K.D., Rawlins, M., and Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. *Health Psychology*, *22*, 68-78.
- Vargas, P.T., Sekaquaptewa, D., and von Hippel, W. (2007). Armed only with paper and pencil: "Low-tech" measures of implicit attitudes. In B. Wittenbrink and N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 103-124). New York: Guilford Press.