Adding Significance to the Implicit Association Test

Peter Stüttgen

Joachim Vosgerau

Claude Messner

Peter Boatwright

Draft March 31st 2011

Peter Stüttgen (pstuettg@andrew.cmu.edu) is a doctoral candidate in Marketing, and Joachim Vosgerau (vosgerau@cmu.edu) and Peter Boatwright (boatwright@cmu.edu) are Associate Professors of Marketing at the Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Claude Messner (claude.messner@imu.unibe.ch) is Professor of Marketing at the University of Bern, Engehaldenstrasse 4, 3012 Bern, Switzerland.

Abstract

The Implicit Association Test has become one of the most widely used tools in psychology and related research areas. The IAT's validity and reliability, however, are still debated. We argue that the IAT's reliability, and thus its validity, strongly depends on the particular application (i.e., which attitudes are measured, which stimuli are used, and the sample). Thus, whether a given application for a given sample will achieve sufficient reliability cannot be answered *a priori*. Using extensive simulations, we demonstrate an easily calculated post-hoc method based on standard significance tests that enables researchers to test whether a given application reached sufficient reliability levels. Applying this straightforward method can thus enhance confidence in the results of a given IAT. In an empirical test, we manipulate the sources of error in a given IAT experimentally and show that our method is sensitive to otherwise unobservable sources of error.

Keywords: Implicit Association Test, Reliability, Simulation

Appropriate measurement of the unconscious has long been an important topic in psychology. Whereas early accounts such as Freud's psychoanalysis were marred by the difficulty of valid assessment and post-hoc interpretations, Fazio, Sanbonmatsu, Powell, and Kardes' (1986) seminal paper on automatic priming offered a methodology that seemed to allow reliable measurement of unconscious attitudes. Nine years later, Greenwald and Banaji (1995) formally defined these as 'implicit attitudes', "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald & Banaji, 1995, p. 8). Implicit attitudes are thought of as co-existing with explicit attitudes about the same attitude object but potentially differing in their evaluative component, accessibility, and stability (Wilson, Lindsey, & Schooler 2000). The Implicit Association Test (IAT), introduced by Greenwald, McGhee, and Schwartz (1998), is the most widely used tool for their measurement. In fact, the IAT has become one of the most applied psychological methods ever used; more than 1700 articles have been published with this method to date (source PsycINFO). As of March 2011, the keyword search "Implicit Association Test" in Google yields approximately 79,000 hits (for comparison, the keyword search "big five personality test" yields 'only' 38,000 hits, indicating the current popularity of the IAT).

The IAT findings and far-reaching policy implications have triggered a vibrant discussion regarding the reliability and validity of the IAT (e.g., Arkes & Tetlock, 2004; Banaji, Nosek, & Greenwald, 2004; Blanton & Jaccard, 2006a, 2006b, 2006c, 2006d; Blanton, Jaccard, Christie, & Gonzales, 2007; Blanton, Jaccard, Gonzales, & Christie, 2006; Greenwald et al., 2002; Greenwald, Rudman, Nosek, & Zayas, 2006; Kang &

Banaji, 2006; Mitchell & Tetlock, 2006; Nosek & Sriram, 2007). Complicating matters is the fact that both validity and reliability are difficult to determine since there are no other sufficiently validated measures of implicit attitudes that would allow for benchmarking. Other explicit criterion measures (e.g. behavior, judgment, and choice) are equally problematic, as there is some debate under which conditions implicit attitudes will guide behavior, judgment, and choice (Blanton et al., 2006; Messner & Vosgerau, 2010; Mitchell & Tetlock, 2006).

In this paper, we argue that the amount of error contained in the IAT varies from application to application, depending on which attitudes are measured, the selection of stimuli, and the sample at hand. As a consequence, some IATs will exhibit satisfactory levels of reliability and validity whereas others will not. We present a method based on standard significance tests that allows researchers to distinguish between applications plagued by too much error and applications with little error. Thus, our method provides confidence in the results of any given IAT that passes our significance test.

The reminder of the paper is organized as follows. First, we give a short description of the IAT and review the potential sources of measurement error in the IAT. We then simulate implicit attitudes and manipulate measurement error from different sources, and show that the overall level of error is reliably related to the number of significant pairs of IAT-scores in a given IAT. Based on our simulations, we determine a cutoff above which IATs can be confidently interpreted as they contain sufficiently low error. Finally, we test our method on three empirical IAT applications.

The Implicit Association Test

In the IAT, participants see stimuli (words or photos) that are presented sequentially in the center of a computer screen. For example, in one of Greenwald et al.'s (1998) original IATs, the stimuli consisted of pleasant (e.g., peace) and unpleasant (e.g., rotten) words, and of words representing the two target concepts: flowers (e.g., rose) and insects (e.g., bee). Participants have two response keys. In the first part of the IAT, participants are instructed to press the left response key (we will denote this as *R1*) whenever a pleasant word or a flower name is presented on the screen; whenever an unpleasant word or an insect name is presented, they are instructed to press the right response key (*R2*). Importantly, participants are asked to respond as fast as possible without making mistakes. Participants perform this categorization task until all stimuli have been presented several times. Typically, there are 40 trials within a block, so that respondents are asked 20 times to press a key for flowers and pleasant words, and 20 times to press another key for insects and unpleasant words. This is the first critical block of the IAT.

In the second critical block, participants' task is the same; however, now the allocation of the response keys is switched. The left response key is now pressed for pleasant words and insects (*R3*), and the right response key is pressed for unpleasant words and flowers (*R4*). So in contrast to the first block, flower names now share a response key with unpleasant and insects share a response key with pleasant. Again, this block typically consists of 40 trials, with 20 responses for insects and pleasant words and 20 responses for flowers and unpleasant words.

The time it takes participants to respond in each trial of the two blocks is interpreted as a measure of the strength with which flowers are associated with pleasant (first block) or unpleasant (second block), and insects are associated with unpleasant (first block) or pleasant (second block). Response latencies are averaged within the first and the second block. The block with shorter average response latencies is called the compatible block, and the block with longer average response latencies is called the incompatible block. The IAT-effect is computed by subtracting the mean response latency of the compatible block from the mean response latency of the incompatible block, i.e.,

$$IAT = \frac{\overline{R3} + \overline{R4}}{2} - \frac{\overline{R1} + \overline{R2}}{2} \qquad (1)$$

A positive IAT-effect is typically interpreted as an implicit preference for flowers over insects. The more positive the IAT-effect, the stronger the implicit preference.

In 2003, Greenwald et al. (2003) introduced a new scoring method, the so-called D-score. In the D-score, individual IAT-scores given by equation 1 are divided by the individual's standard deviation of all response latencies in both blocks. The D-score is aimed at correcting for variability in the difference scores due to differences in general processing speed (GPS) across participants.

However, Blanton and Jaccard (2008) show that an individual's standard deviation can be written as an additive function of (1) half of the difference between blocks (i.e., of the original IAT-score) and (2) the variance of the within-block latencies (i.e., measurement error). Thus, in the absence of random measurement error (generally a desirable condition) it will equal exactly half of the original IAT-score since the variance of the within-block latencies will be zero. Thus, no matter what the original difference between blocks is (representing strong or weak attitudes), the D-score will assign a value of ±2.0 to every respondent, which is typically interpreted as a very strong implicit attitude. The more measurement error is contained in the response latencies, the lower will be the resulting D-score. Thus, the D-scoring removes meaningful variance in individual IAT-scores by using an individual standardization that assigns everybody an extreme implicit attitude in the absence of measurement error.

As a consequence, individual IAT D-scores can no longer be meaningfully compared as both mean and variance of a score are individually standardized. Our proposed solution, in contrast, depends on the meaningful comparisons of individual IAT-scores; we thus employ the original IAT-scoring method. In light of the psychometric problems of the D-score, we consider this to be an advantage of our method.

Validity and Reliability of the IAT

Validity and reliability of the IAT have been assessed by various researchers. Doubts about the IAT's reliability were fueled by findings of unsatisfactory levels of test-retest reliabilities (e.g., Bosson, Swann, & Pennebaker, 2000; but see also Cunningham, Preacher, & Banaji, 2001) whereas the IAT's validity was threatened by reports of low correlations between the IAT and other measures of implicit attitudes (e.g., Bosson et al., 2000; Sherman, Presson, Chassin, Rosem & Koch, 2003; Olson & Fazio, 2003; for an overview see Messner & Vosgerau, 2010).

Reliability is typically regarded as a necessary condition for validity. However, Cunningham et al. (2001) have argued that low reliability (or high measurement error) need not be a threat to construct validity as low reliabilities only impose an upper limit on the possible correlations with other measures of implicit attitudes (Bollen, 1989). The authors employ a latent variable model to analyze the results from several measures of implicit attitudes to explicitly model the effect of measurement error. They conclude that the IAT assesses the same fairly stable implicit construct as do other implicit attitude measures albeit with large amounts of measurement error, i.e., the IAT is a valid but potentially not reliable measure of implicit attitudes. Specifically, the authors state that "on average, more than 30% of the variance associated with the measurements was random error" (p.169). Since the reliability of a measurement instrument is defined as $r = 1 - \frac{\sigma_{error}^2}{\sigma_{total}^2}$ (Mellenbergh, 1996), it implies that the reliability, on average, is less than .7. Nunnally (1978) suggests that reliability levels for instruments used in basic research be above .7, and that reliability for instruments used in applied research be at least .8. Where important decisions about the fate of individuals are made on the basis of test scores, Nunnally recommends reliability levels above .9 or .95. We will calibrate our proposed method such that IATs that are judged to be satisfactory will have a reliability of at least .8. If the IAT is to be used for basic research or as a diagnostic test of individual differences, the method can easily be changed to reach a threshold of .7 or of .9 or .95.

Error in the IAT

We start with a general measurement-model of the IAT (Figure 1). This model consists of the following four components: first, an individual's true association strengths (*Tj$_i$*) for the four implicit attitudes (say, flowers/positive, insects/positive, flowers/negative, insects/negative); second, the observed reaction times for each response key and each individual (*Rj$_i$*); third, random measurement error (*ME$_{ji}$*); and fourth, potential systematic error (*SE$_{ji}$*). *Sj$_i$* denotes the latent construct actually measured by the observed reaction times, consisting of both the implicit attitudes and systematic error.

Thus, if we were to calculate an IAT-score at each of the three steps in Figure 1, the correlation between the first two, *v*, would reflect the IAT's validity, whereas the correlation between the latter two, *r*, would reflect the IAT's reliability. Random measurement error then impacts the reliability of the IAT, whereas systematic error would reduce the validity of an IAT. When observed IAT-scores are correlated with behavior or other predictor criteria (thought to reflect the true implicit attitudes), as is standard practice in the literature, the resulting correlation actually reflects both validity and reliability (*t*) and is therefore difficult to interpret in terms of trying to assess the IAT's validity and/or reliability.

The standard approach to estimate correlations involving latent constructs (i.e., *r* and *v*) would be to use a structural equations model (SEM). However, SEM is not helpful in this particular application. Since each latent construct in Figure 1 is connected to only a single observed construct without any cross-connections, the latent constructs *S1* through *S4* are not separately identified from the means of the observed constructs *T1* through *T4*. Since the calculation of the IAT-scores also only uses the means of the observed reaction times SEM will always result in estimates of *r* equal to 1.

Thus, the aim of our paper is to develop a practical post-hoc method that allows for estimating *r* alone in any given IAT. Applying this method will enable researchers to ensure that the reliability *r* of a given IAT is sufficiently high. To do so, we first start by reviewing the different sources of systematic and random error in the IAT.

Systematic Error in the IAT

Systematic error can be interpreted as adding a constant intercept to the reaction times. Thus, systematic error changes what exactly the reaction time measurement is centered on. For example, some people are generally faster to respond than others. The construction of IAT-scores is aimed at eliminating the influence of such nuisance factors by subtracting the average response latency of the compatible block from that of the incompatible block. As long as the added intercept is constant across the two blocks, the difference IAT-score is free of such nuisance factors. If the intercept differs between blocks, but is constant across participants, systematic error will only shift the neutral point of the IAT-scores away from zero. Blanton and Jaccard (2006a, 2006b; but see also Greenwald, Nosek, & Sriram, 2006) therefore concluded that researchers should not assume that the IAT-metric has a meaningfully defined zero-point, and urge researchers not to test IAT-scores against zero. Our methodology (which we will introduce later on) takes this caution into account, and instead of testing individual or aggregated IAT-scores against zero, will test individual IAT-scores against each other.

When in addition systematic error in the IAT also varies between subjects, the validity of the IAT will be affected. The extant literature has identified a couple of such potential sources of systematic error, namely cognitive inertia (Messner & Vosgerau, 2010), general processing speed (Blanton et al., 2006), familiarity with the stimuli (Kinoshita & Peek-O'Leary, 2005), and potential "cross-category associations" between the stimuli (Steffens & Plewe, 2001).

Cognitive inertia refers to the difficulty of switching from one categorization rule in the first block to an opposite categorization rule in the second block (Messner & Vosgerau, 2010), leading, *ceteris paribus*, to slower reaction times in the second block relative to the first block. This leads to the well-documented order effect, i.e. IAT-effects are typically stronger when the compatible block precedes the incompatible block (e.g., Greenwald, Nosek, & Banaji, 2003; Hofmann, Gawronski, Gschwender, & Schmitt, 2005). Not only does it seem plausible that people are heterogeneous in the extent to which they exhibit cognitive inertia, but due to the standard procedure of counter-balancing the order of blocks across participants the effect of cognitive inertia is certainly not constant across participants.

The effect of general processing speed is due to the fact that some people are generally faster to respond than others (Blanton et al., 2006). Likewise, the more familiar the stimuli, the faster participants will be able to respond to them. Individual differences in general processing speed only manifest themselves for tasks that are moderate to high in difficulty, but not for tasks that are easy (Ackerman, 1987). Since the categorization task is supposedly easy in the compatible block, but harder in the incompatible block, differences in general processing speed will manifest themselves more in the incompatible block. Therefore, differencing the two blocks will not subtract out the effect of general processing speed.

Finally, cross-category associations distort IAT-effects if some or all of the stimuli used in a particular application are strongly associated with one of the target categories as well as with one of the evaluative poles (Steffens & Plewe, 2001; cf., also Govan & Williams, 2004; DeHouwer 2001). For example, IAT-scores in a Germans-Turks IAT will differ when the category German is represented by photos of Hitler versus photos of Claudia Schiffer. Because Hitler not only represents the category German but also the category of most evil dictators, respondents will be faster to categorize Hitler with unpleasant words than with pleasant words. Such cross-category associations lead participants "to complete the task with sorting rules different from those intended for the design" (Nosek, Greenwald, & Banaji, 2007, p. 269) and therefore distort what is being measured in the IAT.

The presence of any of these sources of systematic error will reduce the validity of the IAT, but would not necessarily affect the IAT's reliability.

Random Measurement Error in the IAT

In order to understand the reliability of (i.e., the amount of random error contained in) the observed IAT-effects as well as how different factors affect it, we need to analyze the components of the variance of the observed IAT effects (all variances and covariances in the following derivations are across individuals, not across keystrokes):

$$\begin{aligned}\operatorname{var}(IAT) &= \operatorname{var}\left(\frac{\overline{R3}+\overline{R4}-\overline{R1}-\overline{R2}}{2}\right)\\ &\propto \operatorname{var}(\overline{R1})+\operatorname{var}(\overline{R2})+\operatorname{var}(\overline{R3})+\operatorname{var}(\overline{R4})\\ &\quad +2\begin{pmatrix}\operatorname{cov}(\overline{R1},\overline{R2})+\operatorname{cov}(\overline{R3},\overline{R4}) \\ -\operatorname{cov}(\overline{R1},\overline{R3})-\operatorname{cov}(\overline{R1},\overline{R4})-\operatorname{cov}(\overline{R2},\overline{R3})-\operatorname{cov}(\overline{R2},\overline{R4})\end{pmatrix}\end{aligned}$$

Following the notation introduced in Figure 1, we can substitute $\overline{Rj_i} = Tj_i + SE_{ji} + \overline{ME_{ji}} = Sj_i + \overline{ME_{ji}}$. (Since systematic error has no direct impact on reliability and any indirect impacts are identical to those of the true attitudes, we neglect systematic error in our discussion of reliability and refer to Sj_i as the true attitudes to be measured rather than as the sum of the true attitudes and systematic error.) Given that random measurement error is independent of attitudes and systematic error, all resulting covariances involving measurement error are equal to zero. Therefore, we have

$$\begin{aligned}\operatorname{var}(IAT) \propto\ & \operatorname{var}(S1) + \operatorname{var}(S2) + \operatorname{var}(S3) + \operatorname{var}(S4) \\ & +2\begin{pmatrix}\operatorname{cov}(S1,S2) + \operatorname{cov}(S3,S4) \\ -\operatorname{cov}(S1,S3) - \operatorname{cov}(S1,S4) - \operatorname{cov}(S2,S3) - \operatorname{cov}(S2,S4)\end{pmatrix} \\ & + \operatorname{var}\left(\overline{ME_1}\right) + \operatorname{var}\left(\overline{ME_2}\right) + \operatorname{var}\left(\overline{ME_3}\right) + \operatorname{var}\left(\overline{ME_4}\right)\end{aligned} \qquad (2)$$

From this equation, we can split the variance of the observed IAT-effects in two parts: the first part reflects the variance of the true IAT-effects, *IAT'* (i.e., IAT-effects computed using the true attitudes rather than the reaction time measures), whereas the last line reflects the variance of measurement error. It is obvious, then, that all influences increasing the amount of measurement error but not affecting the variance of the true IAT-effects will increase the relative amount of error contained in the observed IAT-effects (or equivalently, decrease the IAT's reliability). Similarly, all influences decreasing the variance of the true IAT-effects but not affecting the amount of measurement error in the reaction times will have the same effect.

In addition to the most obvious factor, the amount of measurement error contained in the response latencies, two other factors influence the amount of error contained in the IAT-effects for any given application: the variance of the true implicit attitudes and the correlation of target evaluations (Blanton et al., 2006). These two factors vary from application to application, thereby making it impossible to assess the general reliability of the IAT with conventional test-retest procedures. What is needed instead is a post-hoc methodology for assessing the reliability of a given IAT. We discuss both factors, the variance of the true implicit attitudes and the correlation of target evaluations, in detail before developing our post-hoc methodology.

1. Measurement Error in Response Latencies

The most obvious source of error in the IAT is measurement error associated with response latencies. Reaction times to the same stimuli will obviously vary when measured on the millisecond level. Not surprisingly, more measurement error in the individual reaction times will lead to more error in the IAT-effects. However, since equation (2) includes the variances of the means of the measurement error of the reaction times, including more responses per key assignment will increase the reliability of the reaction times (though this is a trade-off with possible effects of fatigue).

2. Variance of True Implicit Attitudes

Similarly obvious is the influence of the variance of the true implicit attitudes. Increasing this variance (for one or more of the true attitudes) while holding the covariances and the error constant (say, by multiplying Sj by $k>1$) increases the percentage of meaningful variance in the observed IAT-effects, and thereby increases the IAT's reliability.

In the other extreme, consider administering an IAT to a perfectly homogenous sample, i.e. the true attitudes are *exactly* the same for all participants. In this case, all variances and covariances will equal zero, leaving only error in the observed IAT-effects.

3. Correlation of the Target Evaluations

As can be seen from equation (2), the variance of IAT-scores depends not only on the variance of the true attitudes, but also on their correlations. As Blanton et al. (2006) have shown, this is due to the way IAT-scores are constructed. The more the positive (and/or negative) associations towards the two target constructs are correlated across participants (i.e., the greater cor($S1$,$S3$) and/or cor($S2$,$S4$)), the smaller is the amount of meaningful variance in the observed IAT-scores. Consider, for example, a Germans-Turks IAT. The more respondents prefer Turks over Germans, the faster they will respond in the block that pairs Turkish names with pleasant words (and German names with unpleasant words), and the slower they will be in the block that pairs Turkish names with unpleasant words (and German names with pleasant words). Likewise for respondents who prefer Germans over Turks. The more they do so, the faster they will respond in the block that pairs German names with pleasant words, and the slower they will respond in the block that pairs German names with unpleasant words. What both groups of respondents, Turk-lovers and German-lovers, share is the underlying bipolar attitude structure: liking one target implies disliking the other (Pratkanis, 1989). In other words, their evaluations of Turks and Germans are negatively correlated. The construction of IAT-scores takes advantage of this negative correlation by subtracting the average response latency of one block from that of the other block. The two blocks thus serve as repeated but reversed measures of the same construct, and taking the difference score maximizes the difference in implicit attitudes relative to measurement error.

The situation changes dramatically, however, if the evaluations of the two targets are positively correlated. Imagine, for example, that respondents do not care about nationalities but only differ in their degree of misanthropy to philanthropy. In this case, the more respondents like Turks, the more they also like Germans. In the terminology of Pratkanis (1989), this is a unipolar attitude structure: liking one target implies liking the other target. The evaluations of the two targets are positively correlated. In this case both IAT-blocks serve as repeated (but not reversed!) measures of the same construct, and subtracting one from the other removes all attitude information. What is left is mainly measurement error.

A similar argument holds if respondents have ambivalent attitudes towards one or both of the target categories. Ambivalent attitudes result from harboring both positive *and* negative associations towards a category (De Liver, van der Plight, & Wigboldus; 2007; see also Cacioppo & Berntson, 1994 for a review on the separability of positive and negative associations). Thus, the correlations between the true attitudes would also be positive, resulting in a lower reliability of the IAT.

Note that the underlying structure of the implicit attitudes, that is, whether target evaluations are negatively correlated or positively correlated, is unobservable. As the IAT is designed to measure implicit attitudes, but the relative amount of error in IAT-scores depends on the correlation of the target evaluations, the amount of relative error resulting from the underlying attitude structure cannot be determined.

In summary, the variance of the true implicit attitudes and the correlation of target evaluations affect the reliability of a given application of the IAT. As both factors vary from application to application, it is impossible to get an estimate of the general reliability of the IAT. Above, we determined the directional effects of these factors, *ceteris paribus*, and illustrated some extreme cases. However, in order to understand the strengths and possible interactions of those effects better as well as, more importantly, to derive and illustrate our proposed solution to the problem, we conduct a simulation analysis.

Simulation Procedure

Our simulation procedure directly mirrors the data generating process of the IAT, while allowing us to simulate the different determinants of the IAT's reliability. Specifically, we simulate the amount of relative error in the IAT by varying the amount of measurement error while holding the variance of true implicit attitudes constant. We also manipulate the correlation between the target evaluations from -1 to +1, and vary the number of keystrokes per combination. We utilize a simulation rather than an analytical approach because the analytical solution would entail a multivariate distribution, defined over the positive definite space, conditional on a subset of the covariance terms. Although the unconditional distribution in this case would be the Wishart distribution, to our knowledge the required conditional distribution has not been derived.

First, we simulate the underlying true attitudes to be measured by the IAT (again, we choose to ignore systematic error, i.e. we effectively simulate *Sj*). For each participant (we choose n = 50 for the simulations reported below), we simulate the four parts of the implicit attitudes corresponding to the four measures of the IAT procedure (i.e., *S1*, *S2*, *S3*, and *S4*) from standard normal distributions. This allows us to manipulate the correlations of the target evaluations. For instance, for the Germans-Turks IAT mentioned above, we manipulate the correlation between the associations German/positive and German/negative and the associations Turks/positive and Turks/negative, respectively. We denote this correlation by ρ (cf. Figure 1), and let ρ vary from -1 (extreme bipolar attitude structure) to +1 (extreme unipolar attitude structure).[1]

We then simulate the individual keystrokes' response latencies by adding random, normally distributed measurement error to the true implicit attitudes. We do simulations with k = 20 and k = 40 keystrokes per combination. As the true amount of measurement error is unobservable, we vary the amount of measurement error in three levels (low, medium, and high) to cover a wide range of potential amounts of measurement error. The measurement error associated with the individual response latencies and the number of keystrokes together determine the error variance associated with the means of the response latencies (henceforth error variance of means = EVM); in our simulations, the EVM for 20 keystrokes ranges from very small amounts (10%) to very large amounts (90%), for 40 keystrokes it ranges from 5% to 45%, relative to the variance of the true implicit attitudes. In addition to being unobservable, the true amount of measurement error again depends on the sample at hand (as the error variance is relative to the true variance). For a sample with almost identical implicit attitudes, IAT-scores will contain almost exclusively error, whereas the same amount of error will have less of an impact in a highly heterogeneous sample. Thus, we make no claim about the true amount of measurement error contained in reaction times, but simply attempt to understand its effect on the resulting IAT-effects.

In the final step, we compute the true IAT-effect from the simulated true attitudes as well as the observed IAT-effect from the simulated reaction times. By comparing the true and the observed IAT-effects, we can then calculate the percentage of error variance in the total (i.e., true + error) observed variance.

We repeat this simulation 10,000 times for each combination of the varying factors. Thus, we conduct 2 (keystrokes: 20 vs. 40) x 3 (measurement error: low vs. medium vs. high) x 25 (ρ: from -1 to .9 in steps of .1, .925, .95, .975, .99, .9999) x 10,000 = 1,500,000 simulations.

Simulation Results

We briefly summarize the results of the simulations, confirming the above analysis, before proposing our solution to ensure sufficient reliability for the IAT.

1. Effect of Random Measurement Error and Number of Keystrokes

The amount of measurement error associated with the individual keystrokes and the number of keystrokes per key have the expected effects on the amount of error in IAT-scores. The more error the individual keystrokes contain, the less precisely the four underlying attitudes can be measured. This then leads to more error contained in the IAT-effects, all else equal. As mentioned above, though, this can be mitigated by increasing the number of keystrokes. Keeping the variance associated with the individual keystrokes constant, increasing the number of keystrokes per key decreases the amount of error contained in the IAT-effect (see Figure 2).

<< Insert Figure 2 about here >>

2. Effect of Correlation between Target Evaluations

More interestingly, though, the simulations also confirm the expected relationship between the correlation of the target evaluations and the overall amount of error contained in IAT-effects. Figure 3 displays the percentage of error variance contained in the IAT as a function of the correlation of the target evaluations. The three panels are, from left to right, for low, medium, and high measurement error (all with 20 keystrokes, resulting in EVMs of 10%, 50%, and 90%, resp.). Within each panel, the correlation of the target evaluations (ρ) varies from -1 at the left end to +1 at the right. The solid line depicts the median of the simulated error percentages, whereas the dashed lines are the 5th and 95th percentiles (i.e., 90% of the simulated error percentages fall between the two dashed lines).

<< Insert Figure 3 about here >>

As mentioned above, it is not surprising that the percentage of error in IAT-effects is higher for higher amounts of measurement error. Note, however, that for each of the three levels, the average error contained in the IAT-effects is well below the respective EVM if ρ is highly negative. This is due to the fact that these negative correlations add to the meaningful variance contained in the observed IAT-scores (cf. equation (2)).

If ρ is highly positive, on the other hand, the percentage of random error in the observed IAT-effects is above the respective EVM. If the targets are perfectly positively correlated, the observed IAT-effects contain *only* random error, independent of how well response latencies measure implicit attitudes. Again, this is an artifact of the way the IAT-effect is computed. The more positive ρ is, the smaller is the variance of the true IAT-effects and, therefore, the greater the role measurement error plays.

Thus, the measurement error in the individual response latencies only affects how well the IAT can measure in the optimal case and how quickly the situation worsens (i.e., the curvature). Even if response latencies are extremely good measures of implicit attitudes, the amount of random error contained in IAT-effects quickly reaches unacceptably high values (or equivalently, reliability reaches unacceptably low values) once target evaluations are positively correlated.

Adding Significance to the IAT

As pointed out previously, the correlation between the true (implicit) target evaluations ρ is unobservable. Likewise, the amount of measurement error relative to the variance of the true implicit attitudes is unknowable. Thus, one cannot determine *a priori* whether the IAT is reliable for a certain application/sample or not. However, we will show that the percentage of error contained in the observable IAT-effects is reliably related, irrespective of the source of the error, to the number of significant pairwise comparisons between individual IAT-effects (i.e., the number of respondent-pairs whose IAT-effects are significantly different from each other). Thus, calculating significance tests between the IAT-scores for all possible pairs of participants can be used as a proxy for the amount of error contained in a particular application of the IAT. Our simulations show that higher amounts of error lead to less significant pairwise comparisons. Based on our simulations, we determine a minimum cutoff for the number of significant IAT-score pairs as a proxy for satisfactorily low levels of error contained in the IAT.

Calculating Significant Pairs

With *k* keystrokes per response key, we can interpret an individual's IAT-effect as the average of *k* repeated measurements of that individual's true IAT-effect (where each of the repeated measurements is calculated as $\frac{R3+R4}{2}-\frac{R1+R2}{2}$ using 1 keystroke from each of the 4 response keys rather than the average). Thus, significance testing of the difference between two individual IAT-effects (say, IAT_1 and IAT_2) is essentially a significance test between two means.

Testing against the null hypothesis of no difference between the means, the *t*-statistic for this hypothesis test is given by

$$t=\frac{IAT_1-IAT_2}{\sqrt{sd_p^2\cdot\left(\frac{1}{k_1}+\frac{1}{k_2}\right)}}$$

where sd_p^2 is the pooled variance of the two IAT measurements given by

$$sd_p^2 = \frac{(k_1 - 1)\cdot sd_1^2 + (k_2 - 1)\cdot sd_2^2}{(k_1 - 1) + (k_2 - 1)}$$

where sd_1 (sd_2) and k_1 (k_2) are the standard deviation and the number of draws for the first (second) IAT-effect of the pair to be compared.

Since k_1 and k_2 are typically the same for all IATs within one study, the calculation of the t-statistic reduces to

$$t = \frac{IAT_1 - IAT_2}{\sqrt{\frac{sd_1^2 + sd_2^2}{k}}} \qquad (3)$$

where $k = k_1 = k_2$. This t-statistic is distributed according to a Student-t distribution with $(k_1\text{-}1)+(k_2\text{-}1)=2k\text{-}2$ degrees of freedom and can then be used for a standard hypothesis test.

In order to calculate this t-statistic, one needs to compute the standard deviation of an individual IAT-effect, sd_{IAT}. This is given by

$$sd_{IAT} = \sqrt{\operatorname{var}\left(\frac{R3 + R4 - R1 - R2}{2}\right)}$$

Since the k response times per response key are seen as k independent repeated measures of the same true attitude (rather than being in a specific order), we set the covariances (of the response times for one participant) to zero. Thus, the following equation can be used to estimate the standard deviation for an individual IAT-effect from the observed reaction times:

$$sd_{IAT} = \frac{1}{2}\sqrt{\operatorname{var}(R1) + \operatorname{var}(R2) + \operatorname{var}(R3) + \operatorname{var}(R4)} \qquad (4)$$

We choose the typically used value of α=.05 as the significance level for our significance test applied to all possible pairwise comparisons.

Relationship Between Error in the IAT and Number of Significant Pairs

Since we simulate individual keystrokes, we can use the same formula to conduct the significance tests in our simulations. It is expected that more measurement error in the individual reaction times should lead, *ceteris paribus*, to fewer significant pairs, as this would reduce the value of the *t*-statistic. Since more measurement error in the individual reaction times also increases the percentage of error contained in the IAT-effect, this would suggest a negative relationship between the number of significant pairs and the error percentage.

Likewise, the more positive the correlation of the target evaluations ρ is, the smaller the variance of the true IAT-scores will be without affecting their standard deviations. Thus, the greater ρ is, the higher the error percentage is in individual IAT-effects, and the fewer pairwise comparisons will be significant.

Concluding, the more significant pairwise comparisons we observe, the lower is the error percentage in the IAT-effects (or equivalently, the higher the reliability; see Figure 4). This relationship can be used to determine whether a particular application of the IAT can safely be interpreted, or whether it should be disregarded because it likely contains too much error.

<< Insert Figure 4 about here >>

Proposed Solution

As mentioned above, we would like to ensure that the reliability of IATs judged to be interpretable is at least .8 (Nunnally, 1978). Since the relationship between number of significant pairs and error percentage is not one-to-one, but includes some variance, we can never be 100% sure that every interpreted IAT has a reliability of .8, irrespective of how high we set the minimum threshold for the number of significant pairs. Instead, we choose the standard significance level of 5% to define the threshold, i.e. we want to make sure that at least 95% of the interpreted IATs have a reliability of.8 or greater. Thus, for each amount of significant pairs, we calculate the 5th percentile of the distribution of reliabilities resulting from IATs with the respective number of significant pairs. As this is an increasing function of the number of significant pairs, we can find the minimum number of significant pairs such that the 5th percentile of reliabilities is .8, and be certain that for all IATs with at least that many significant pairs at most 5% have reliabilities of less than .8.

While the effect on the relationship between the number of significant pairs and the error percentage goes in the same direction for both changes in random measurement error and changes in ρ, the strength of the effect differs for the two. In particular, an increase in random measurement error resulting in a certain increase in the error percentage is associated with a larger decrease in the percentage of significant pairs than a change in ρ resulting in the same increase in the error percentage. To explore potential consequences of this relationship for a proposed cutoff, we analyze the relationship between the percentage of significant pairs and the 5th percentile of the error percentage for different levels of measurement error. Figure 5 displays this relationship for 5%, 15%, and 25% of EVM, as well as a fitted (negative) exponential curve.[3] The more measurement error, the stronger the curvature (and therefore the larger the rate parameter of the exponential distribution).

<< Insert Figure 5 about here >>

Based on these estimated exponential functions, we propose that the threshold be set at 50% of all possible pairwise comparisons. For EVMs up to 20%, the 5th percentile reaches .8 between 48% and 50%; thus, a threshold of 50% is appropriate for these levels. As can be seen in the right-most panel of Figure 5, the 5th percentile of reliabilities *never* reaches .8 at EVM of 25%.

Thus, we make sure that in the best case, we do not exclude more IATs than necessary. On the other hand, this approach runs the risk of accepting more IATs than appropriate if the random measurement error is large. However, two reasons justify our approach: (1) At 25% EVM only 8 of the 10,000 repetitions of the best case ($\rho = -1$) have at least 50% significant pairs. This is due to the fact that larger amounts of measurement error lead to fewer significant pairs. Thus, for large amounts of measurement error the conditional probability of the reliability being at least .8 (conditional on at least 50% significant pairs) is lower than one would like, but the unconditional probability is still far below 5%. (2) Even in the case of larger amounts of measurement error, using this conservative threshold is still better than not using a threshold at all.

Figure 6 shows the percentage of our simulated IATs that would be accepted using different threshold levels (i.e., IATs with more significant pairs than the respective threshold). It can easily be seen how increasing the threshold leads to rejecting more IATs with low reliabilities while retaining the ones with high reliabilities. The graph also visualizes that even a threshold that is too low is better than no threshold at all. For instance, applying a threshold of 30% or 40% rejects most IATs with very low reliabilities, providing a significant improvement over using no threshold at all. However, in order to achieve reliabilities of .8 and higher, a threshold of 50% is needed.

If the IAT is used for basic research (i.e., reliability should be .7 or higher; Nunnally, 1978), we recommend a threshold of 40%. However, if the IAT is to be used as a diagnostic test of individual differences, calling for a reliability of at least .9 (Nunnally, 1978), the threshold should be increased to at least 65% of significant pairs.

<< Insert Figure 6 about here >>

While our cutoff of 50% (or 40% or 65%) ensures that IATs will have a minimum reliability level of .8 (or .7 or .9, respectively), it does not take into account that cognitive inertia artificially increases the variance of IAT-scores if the order of blocks is counterbalanced across participants. If order-effects due to cognitive inertia are present, the variance of IAT-scores is artificially inflated, which can lead to an increase of significant pairwise comparisons of IAT-scores. To eliminate the confounding effect of cognitive inertia, our proposed method must be applied within each block-order condition.

Empirical Test

Manipulating the Correlation of Target Evaluations by Cross-Category Associations

We conducted three IATs to empirically illustrate the effects of the different determinants of error and demonstrate the effectiveness of our proposed solution of pairwise significance tests. The study was conducted in Germany and assessed participants' attitudes towards Germans and Turks. We used 80 keystrokes per block, i.e. 40 keystrokes per response key. Thus, we can calculate IAT-effects and the number of significant pairs including all keystrokes as well as including only the first 20 keystrokes per response key (i.e., the standard number of keystrokes). This enables us to illustrate the effect of EVM on the reliability of the IAT.

To manipulate the correlation between the target evaluations ρ, we manipulated cross-category associations between the target concepts German – Turkish, and the attribute stimuli positive and negative. Specifically, we ran three IATs in which we varied the attribute stimuli representing the evaluative poles positive and negative. In the neutral IAT, the positive (e.g., joke or love) and negative (e.g., anxiety or fear) stimuli were not associated with either the Turkish or German stereotype. In the pro-Turkish IAT, the positive stimuli (e.g., bazaar or belly dance) were associated with the Turkish stereotype, and the negative stimuli (e.g., Hitler or Nazi) were associated with the German stereotype. In the pro-German IAT, the positive stimuli (e.g., poet or Easter) were associated with the German stereotype, and the negative stimuli (e.g., death penalty or macho) were associated with the Turkish stereotype. Both cross-category associations, whether they favor Turkish or German, make the correlation between the target evaluations more negative (or less positive). Consequently, we expect more pairwise comparisons of IAT-scores to become significant than in the neutral IAT.

Method

Pretest of the Stimuli. To find positive and negative attribute stimuli that are associated with either the Turkish or German stereotype, we created a list of 92 words. 47 participants were asked to rate each word on two 95mm line scales. The first scale assessed how strongly a word was associated with positive (left anchor) or negative (right anchor), the second scale assessed how strongly a word was associated with German (left anchor) or Turkish (right anchor). Participants indicated association strengths by marking the lines at the corresponding point of their judgment. On the basis of these ratings, we selected 10 positive and 10 negative words such that 5 of each were strongly associated with the stereotype German and the other 5 were strongly associated with the stereotype Turkish. The positive stimuli were rated significantly more positive (*pro-Turkish* M = 15.08mm, SD = 5.33; *pro-German* M = 14.2mm, SD = 2.8) than the negative stimuli (*pro-Turkish* M = 87.6mm, SD = 1.4; *pro-German* M = 78.89mm, SD = 7.49; t (8) = 29.49 $p < .001$; t (8) = 18.09, $p < .001$, respectively). Likewise, in the pro-Turkish set the positive stimuli were more associated with Turkish (M = 77.59mm, SD = 5.0) than the negative stimuli (M = 11.37mm, SD = 6.59; t (8) = 17.94, $p < .001$), and in the pro-German set the positive stimuli were more associated with German (M = 15.55mm, SD = 4.13) than the negative stimuli (M = 62.52mm, SD = 7.89; t (8) = 12.3, $p < .001$).

The positive and negative stimuli for the neutral non-cross-category associated IAT were taken from the Handbook of German Language Word Norms (Hager &

Hasselhorn, 1994).

Participants and Design. 84 female and 39 male participants aged 14 to 31 (*M* = 20) volunteered for a 30-minute study on word categorization. Participants were approached in the pedestrian zone of the city of Heidelberg, Germany, and were paid approximately 2.50 USD for participation. Participants were randomly assigned to one of the three experimental conditions of cross-category associations (pro-Turkish IAT vs. neutral IAT vs. pro-German IAT). The order of the two critical blocks was counterbalanced between subjects.

Procedure. Participants were run in groups of up to four persons per experimental session. Upon arrival at the laboratory, participants were seated in separate cubicles equipped with a PC and a 15-inch CRT monitor with a set up viewing distance of about 60 cm. When participants had taken their seats, the experimenter started the IAT without additional instructions. The IAT program was compiled with the software package E-Prime. For all participants, the target concepts were labeled ‘Germans’ versus ‘Turks’ and the attribute concepts were labeled ‘positive’ versus ‘negative’. Depending on the cross-category associations condition, the positive versus negative stimuli were stereotypically Turkish versus German (the pro-Turkish condition), stereotypically German versus Turkish (the pro-German condition), or unrelated to group stereotypes (the neutral condition).

Five stimuli were used as instantiations of the attribute concepts in each condition (see appendix). IATs were constructed following the typical IAT-design with five practice blocks and two critical blocks. Practice blocks comprised 20 trials, whereas main blocks comprised 80 trials. The inter-stimulus interval was set to 150 ms in all blocks.

The first and second block referred to the differentiation between target concepts and attribute concepts, respectively. Depending on the order condition, the first critical block called for the same response either to German stimuli and positive stimuli or to German stimuli and negative stimuli. The second critical block called for the same response to the reverse combination of target and attribute concepts as compared to the first critical block.

Upon completion of the IAT and an unrelated questionnaire, participants were debriefed, rewarded, thanked, and dismissed.

Results and Discussion

Data reduction. The first two trials of each block were dropped from the analyses, as were all trials involving reactions of more than 3000 ms, reaction times of less than 300 ms, or wrong answers (Perkins, Forehand, Greenwald, & Maison, 2007). Because reaction time data are typically positively skewed, all analyses are based on log-transformed reaction times. For presentational concerns, however, we report raw mean values in milliseconds.

IAT effect scores were computed by subtracting the mean (log-transformed) response latency in the block pairing German and positive stimuli (German-positive block) from the mean (log-transformed) response latency in the block pairing German and negative stimuli (German-negative block). Positive IAT effect scores thus reflect that responses are faster when the same response is required for the target concept ‘German’ and the attribute concept ‘positive’.

IAT-effects. An Analysis of Variance (ANOVA) on IAT-effects with the independent variables cross-category association (pro-Turkish vs. neutral vs. pro-German) and block order (German-positive block first *vs.* German-negative block first) revealed a main effect for cross-category association, *F*(2, 117) = 39.55, p < .01, a main effect for block order, *F*(1, 117) = 6.28, p < .02, but no interaction (see Figure 7).

As expected, for participants in the pro-German IAT, response latencies were shorter in the German-positive block than in the German-negative block (M(German-negative) = 846.61 ms, SD = 166.94; M(German-positive) = 671.32 ms, SD = 145.12), *t*(39) = 9.24, *p* < .01, whereas the reverse was true for participants in the pro-Turkish IAT (M(German-positive) = 847.18 ms, SD = 211.11; M(German-negative) = 774.90 ms, SD = 139.86), *t*(41) = 2.89, *p* < .01. The response latencies of participants in the neutral condition fell somewhere in between (M(German-positive) = 690.27 ms, SD = 102.66; M(German-negative) = 761.68 ms, SD = 139.64), *t*(40) = 4.14, *p* < .01. All three mean IAT-effects are significantly different from each other, all *t*s > 4.15, *p*s < .01.

<< Insert Figure 7 about here >>

Number of Significant Pairwise Comparisons of IAT-scores

Differences among the three IATs. First, we analyze the number of significant pairs by cross-category association (pro-German vs. neutral vs. pro-Turkish), once including all valid response times (see above for exclusion criteria) and once including only the first 20 valid keystrokes per response key. See table 1 for the results of this analysis.

<< Insert Table 1 about here >>

Matched sample significance tests between 20 vs. 40 keystrokes within each cross-category association reveal a significant increase in the percentage of significant pairs when all valid keystrokes are used for the analysis (all *z*s > 2.1, all *p*s < .05). Again, this is expected as an increase in the number of keystrokes decreases the standard error associated with the measurement. It is also in accordance with our simulations.

Similarly, as expected, independent sample significance tests testing for differences among the three IATs show that cross-category-associations lead to more significant pairs. All pairwise comparisons within one row of table 1 are significantly different (all *z*s > 2.6, all *p*s < .01; except for the comparison between the pro-Turkish and the neutral IAT calculated with all responses (*z* = .905, *p* = .366)).

In conclusion, the neutral IAT without cross-category-associations proves to have insufficient reliability. The most likely reason for this is that the distribution of the true implicit attitudes is not dispersed enough; the study included only German citizens (mostly university students), who are likely to have similar attitudes towards Germans vs. Turks. If the study had also included Turkish participants we would expect it to have more than 50% significant pairs.

The Influence of Cognitive Inertia. We now look at the effect of cognitive inertia on the number of significant pairs, since we found a main effect for the order of the critical blocks on the IAT-effects in the ANOVA. The order effect is due to the fact that cognitive inertia in essence shifts all IAT-effects in one order condition to the right, and all IAT-effects in the other order condition to the left by adding/subtracting a constant to the response latencies in the respective second block. However, as long as individuals are affected equally strongly by cognitive inertia (i.e., the constant is approximately the same for all participants within one order condition), this should not affect the percentage of significant pairs within one order condition. Table 2 shows these percentages for each of the three IATs (using all valid keystrokes). Independent sample significance tests do not reveal significant differences for any of the three vertical pairwise comparisons in the table (all zs < 1.3, all ps > .2).

<< Insert Table 2 about here >>

However, theoretically it is possible that the combined analysis yields higher *or* lower percentages than the separate analyses for each block order. Thus, we recommend that analysis of the IAT and the number of significant pairs be performed within order conditions.

General Discussion

We have shown that the reliability of the IAT does not only depend on how well response latencies measure implicit attitudes, but also on factors varying from application to application, like the correlation of the target evaluations (and thus the amount of cross-category associations), the number of keystrokes per combination, and the specific sample at hand. Thus, trying to evaluate the reliability of the IAT procedure *per se* is impossible; instead, each application has to be evaluated individually. Based on our simulations, we propose a method to ensure that applications of the IAT containing too much error are not interpreted.

Implementation of this method is straightforward. It consists of the following steps:

1. Clean the data according to Perkins et al. (2007).
2. Calculate the IAT-scores and associated individual standard deviations using equations (1) and (4), resp.
3. For all possible pairs of participants (for *p* participants, one will have $\frac{p \cdot (p-1)}{2}$ pairs), compute the *t*-statistic according to equation (3) and compare it to the critical value of the *t*-distribution with 2*k*-2 degrees of freedom.
4. Calculate the percentage of pairwise comparisons that turn out to be significant.
5. For reliability to be .8 or greater, the percentage must be greater than 50% (for reliability of .7 the cutoff is 40%, for reliability of .9 the cutoff is 65%). Otherwise, the reliability of this specific application is not sufficient for analysis.

An important point is that this method is only concerned with the *reliability* of the IAT measurements, not with its *validity*. While reliability is a necessary condition for validity, it is not sufficient. Our method only ensures that the IAT measures whatever it is measuring (i.e., *Sj*) reliably. As mentioned in the introduction, there is still some debate on whether response latencies truly tap into implicit attitudes or not with arguments going both ways (e.g., Cunningham et al. (2001) for the pro-side and Karpinski and Hilton (2001) for the con-side). Our proposed pairwise significance tests ensure sufficient reliability, thus enabling researchers to conduct better tests of the IAT's validity.

Finally, note the difference between our method using significant *pairwise* comparisons to the standard IAT interpretation involving significance test of *average* IAT-effects against zero. Comparing IAT-effects to zero implies that one assumes zero is the true dividing point between which target category a certain participant prefers. Yet, related to the issues of validity, there is debate whether this is an appropriate assumption (Blanton & Jaccard, 2006a). Differential effects of general processing speed (or other factors) on response latencies in the two critical blocks cast doubt on whether IAT-scores have a meaningful zero point (Blanton & Jaccard, 2006a, 2006b; but see also Greenwald, Nosek, & Sriram, 2006). In contrast, pairwise comparisons do not make any claims about whether a certain person prefers one target over the other, but only whether one person has a stronger preference for one over the other than another person. This circumvents the problem of the true zero point. Yet, our method ensures that an interpreted IAT does not contain too much error. Therefore, correlations between the IAT-effects and explicit measures, other implicit measures, and/or observed behavior can be confidently calculated. Moreover, changes due to experimental interventions can be confidently analyzed if IATs are conducted both before and after the intervention (and both of them are judged to be interpretable). Of course, if at any point the comparison of IAT-effects between individuals is of interest, only significant differences should be interpreted. It can be shown that the probability of a type I error (i.e., of concluding that person A has a stronger preference than person B when in fact this is not true) can be significantly reduced by using the significance tests. For instance, in a known-group IAT (Greenwald et al., 1998), we would expect most comparisons of individuals from different groups to be significant, while comparisons of individuals within one group may not prove significant. While the IAT overall may then be judged to be reliable (and be used to calculate correlations etc.), one should still refrain from accepting individual IAT-effects at face value.

In summary, our proposed method provides confidence in the IAT measure by adding a safeguard against IATs containing too much error. We believe that this is an important step for the continued use of the IAT in applied research.

Footnotes

[1] To be precise, we choose .9999 as the upper limit, as the true IAT-effects all reduce to exactly 0 for a perfect positive correlation.

[2] k_1 and k_2 may not be equal if some response latencies are taken out of the analysis due to an error in the categorization task.

[3] In particular, we fit a function of the form $-\left(a+\lambda\exp^{-\lambda(x-b)}\right)$ (where x is the percentage of significant pairs and λ is the rate parameter) to the non-truncated part of the observed line of 5th percentiles. We do so by minimizing the squared errors relative to the observed values.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*(1), 3-27.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15*(4), 257-278.

Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15*(4), 279-310.

Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist, 61*(1), 27-41.

Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist,* Blanton, *61*(1), 62-71.

Blanton, H., & Jaccard, J. (2006c). Tests of multiplicative models in psychology: A case study using the unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 113*(1), 155-165.

Blanton, H., & Jaccard, J. (2006d). Postscript: Perspectives on the reply by Greenwald, Rudman, Nosek, and Zayas (2006). *Psychological Review, 113*(1), 166-169.

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, 34, 277-297.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*(2), 192-212.

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology, 43*(3), 399-409.

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley & Sons.

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality & Social Psychology, 79*(4), 631-643.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*(3), 401-423.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science. Special Issue, 121*(2), 163-170.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*(12), 806-813.

De Houwer, J. (2001). A structural and process analysis of the implicit association test. *Journal of Experimental Social Psychology, 37*(6), 443-451.

De Liver, Y., van der Plight, J., Wigboldus, D. (2007). Positive and negative associations underlying ambivalent attitudes. *Journal of Experimental Social Psychology, 43*(2), 319-326

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229-238.

Gladwell, Malcolm (2005), *Blink: The Power of Thinking without Thinking*. New York, NY: Little, Brown and Company.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40*(3), 357-365.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3-25.

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., Nosek, B. A., & Sriram, N. (2006). Consequential Validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist, 61*(1), 56-61.

Greenwald, A. G., Rudman, L. A., Nosek, B. A., & Zayas, V. (2006). Why so little faith? A reply to Blanton and Jaccard's (2006) skeptical view of testing pure multiplicative theories: Postcript. *Psychological Review, 113*(1), 180.

Hager, W., & Hasselhorn, M. (Eds.). (1994). *Handbuch deutschsprachiger Wortnormen [Handbook of German Language Word Norms]*. Göttingen, Germany: Hogrefe.

Hofmann, W., Gawronski, B., Gschwender, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality & Social Psychology Bulletin, 31*(10), 1369-1385.

Kang, J., & Banaji, M. R. (2006). Fair measures: A behavioral realist revision of "affirmative action". *California Law Review, 94*, 1063-1118.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality & Social Psychology, 81*(5), 774-788.

Kinoshita, S., & Peek-O'Leary, M. (2005). Does the compatibility effect in the Race Implicit Association Test reflect familiarity or affect? *Psychonomic Bulletin and Review, 12*(3), 442-452.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293-299.

Messner, C., & Vosgerau, J. (2010). Cognitive inertia and the Implicit Association Test. *Journal of Marketing Research, 47*(April), 374-386.

Mitchell, G., & Tetlock, P. E. (2006). Antidiscrimination law and the perils of mind-reading. *Ohio State Law Journal, 67*, 1023-1121.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic Processes in Social Thinking and Behavior*: Psychology Press.

Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales, and Christie (2006). *Journal of Experimental Social Psychology, 43*(3), 393-398.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*(6), 636-639.

Perkins, A., Forehand, M. R., Greenwald, A. G., & Maison, D. (2007). Measuring the non-conscious: Implicit social cognition on consumer behavior. In *Handbook of Consumer Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pratkanis, A. R. (1989). The cognitive representation of attitudes. In A. R. Pratkanis, S. J. Breckler & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 71-98). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Sherman, S. J., Presson, C. C., Chassin, L., Rose, J. S., & Koch, K. (2003). Implicit and explicit attitudes towards cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology, 22*(1), 13-39.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the implicit association test. *Zeitschrift fuer Experimentelle Psychologie, 48*(2), 123-134.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101-126.

Table 1

# of Keystrokes	Pro-German	Neutral	Pro-Turkish
20 keystrokes	50.8%	37.9%	44.3%
40 keystrokes	55.4%	45.0%	47.2%

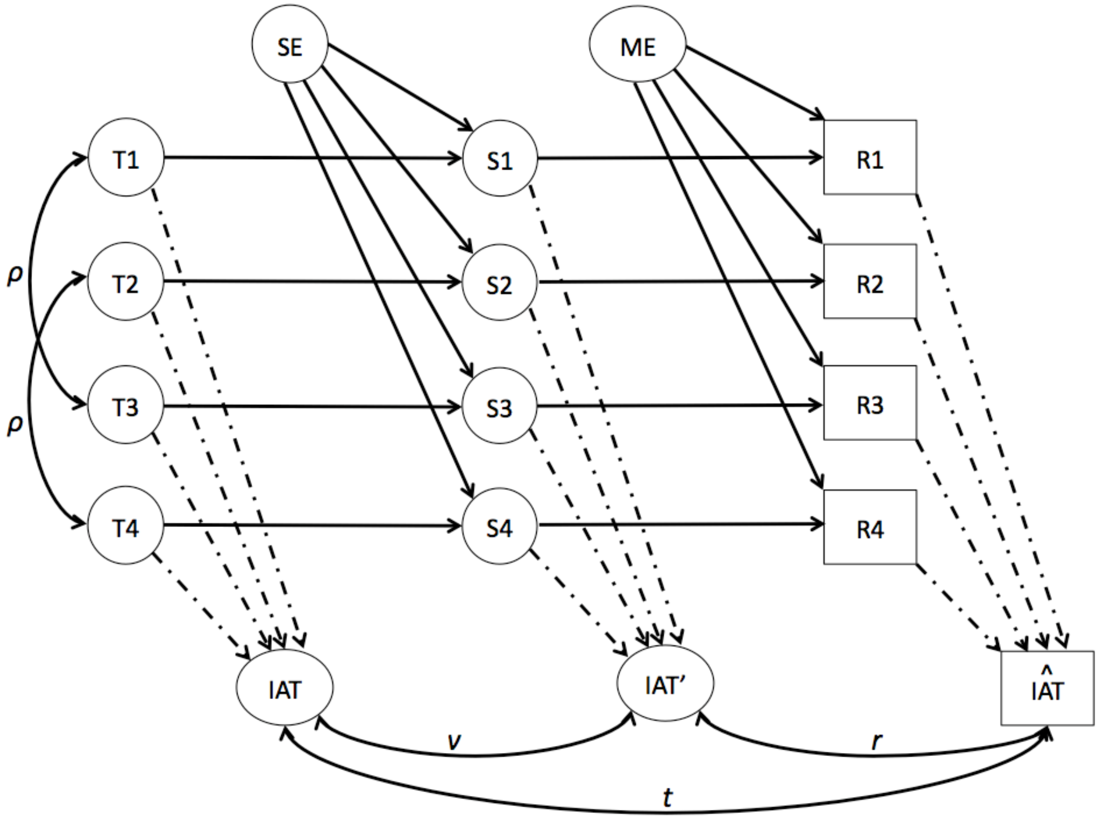
Percentage of significant pairwise comparisons by type of IAT and number of keystrokes

Table 2

Block Order	Pro-German	Neutral	Pro-Turkish
German/positive block first	56.2	43.7	48.4
Turkish/positive block first	49.7	47.1	44.6

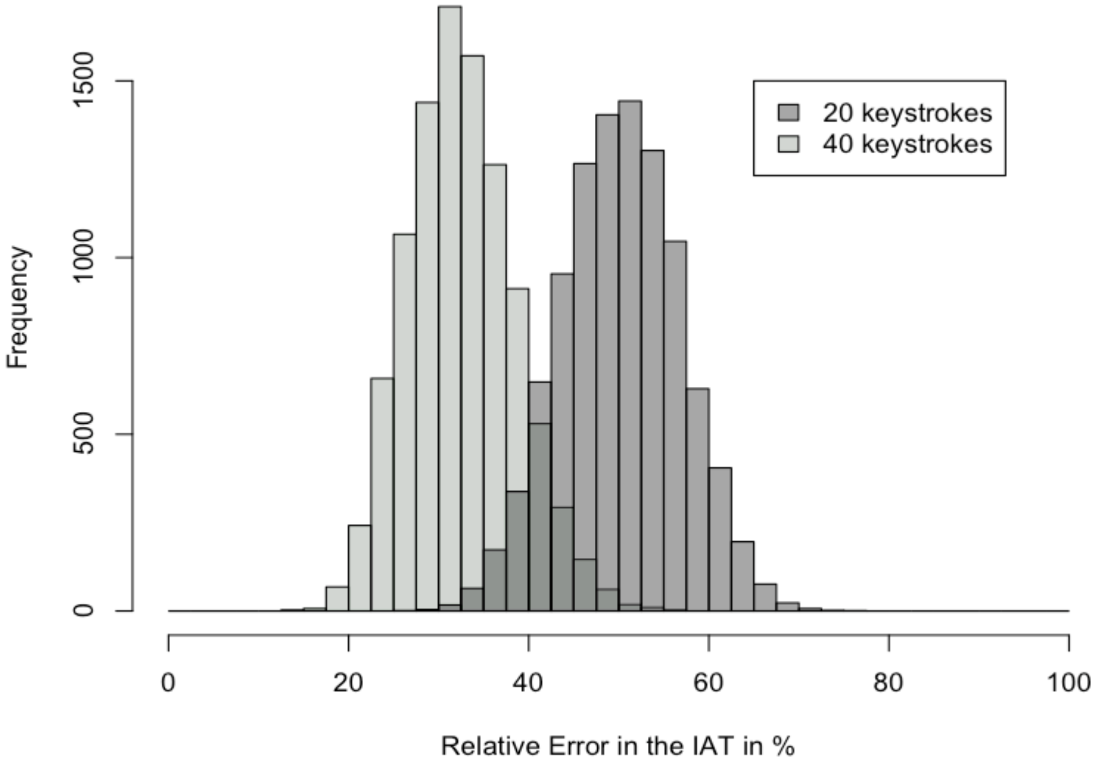
Percentage of significant pairwise comparisons by type of IAT and by order of blocks

Figure 1



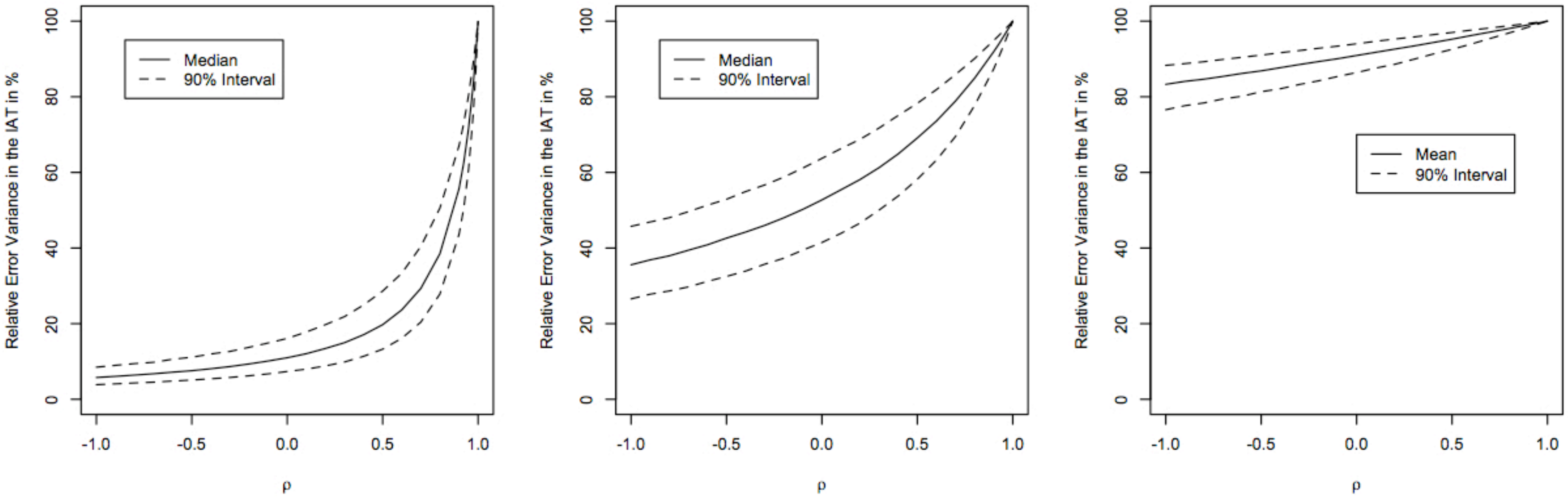
Schematic representation of implicit attitudes, IAT measurements, and different error sources

Figure 2



Histogram of the relative error contained in the IAT with 20 vs. 40 keystrokes (at medium level of measurement error and $\rho = .0$)

Figure 3



Percentage of error variance contained in the IAT for (from left to right) low (10%), medium (50%), and high (90%) levels of measurement error (EVM; all with 20 keystrokes)

Figure 4

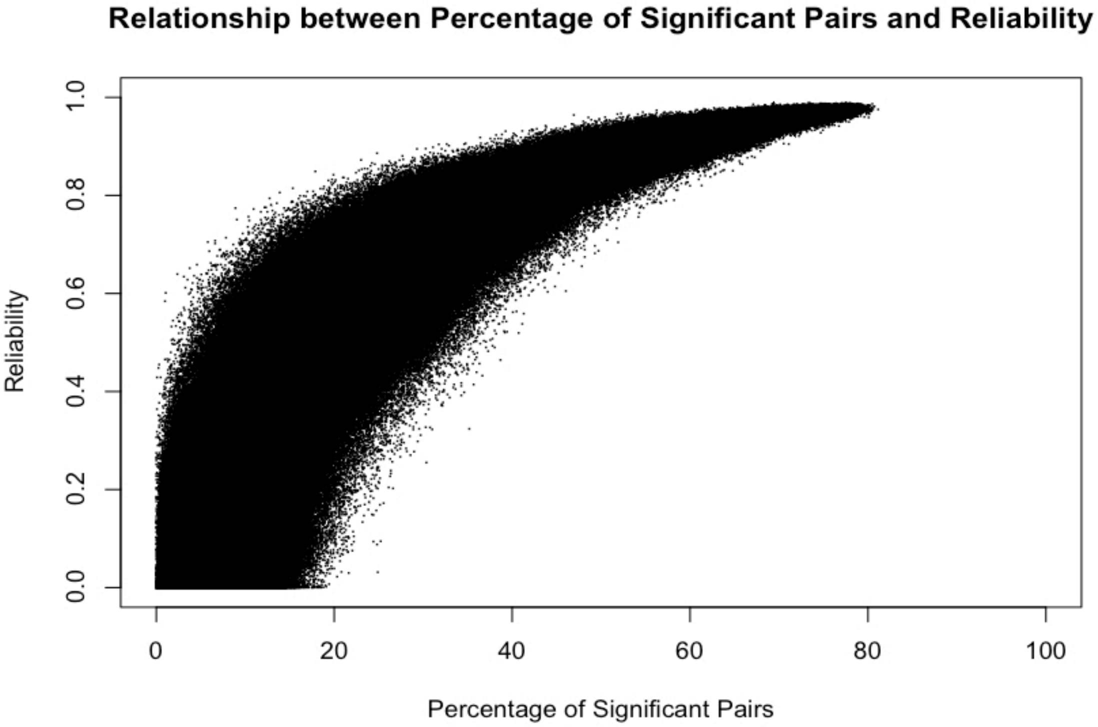
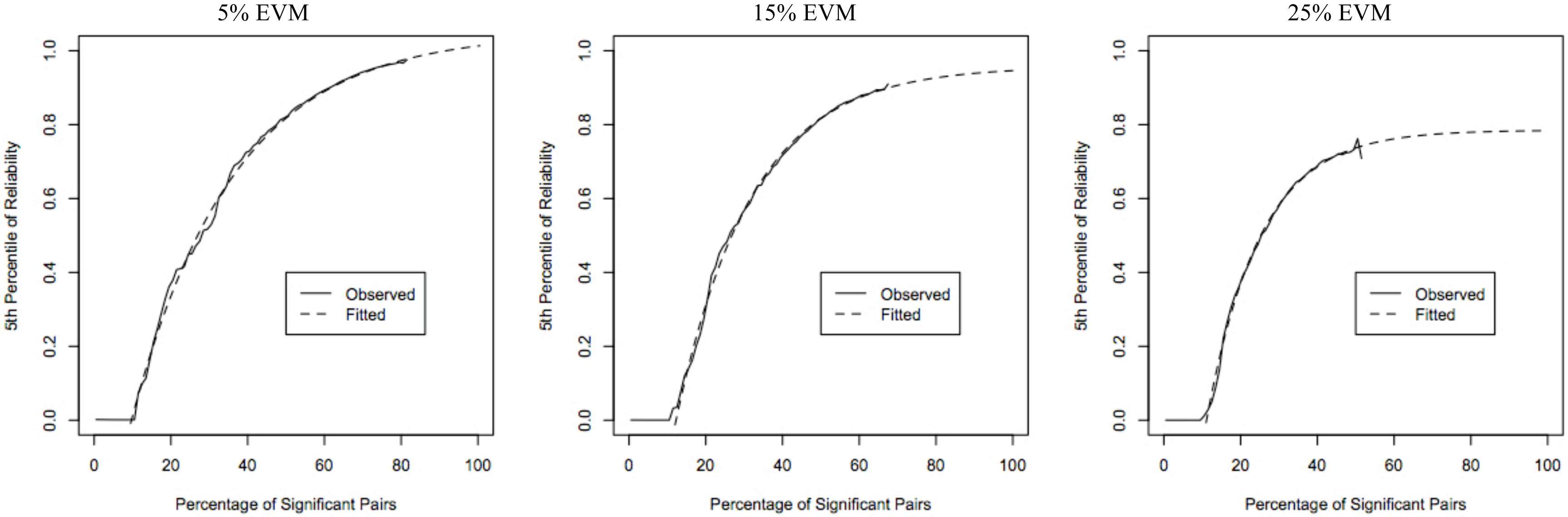
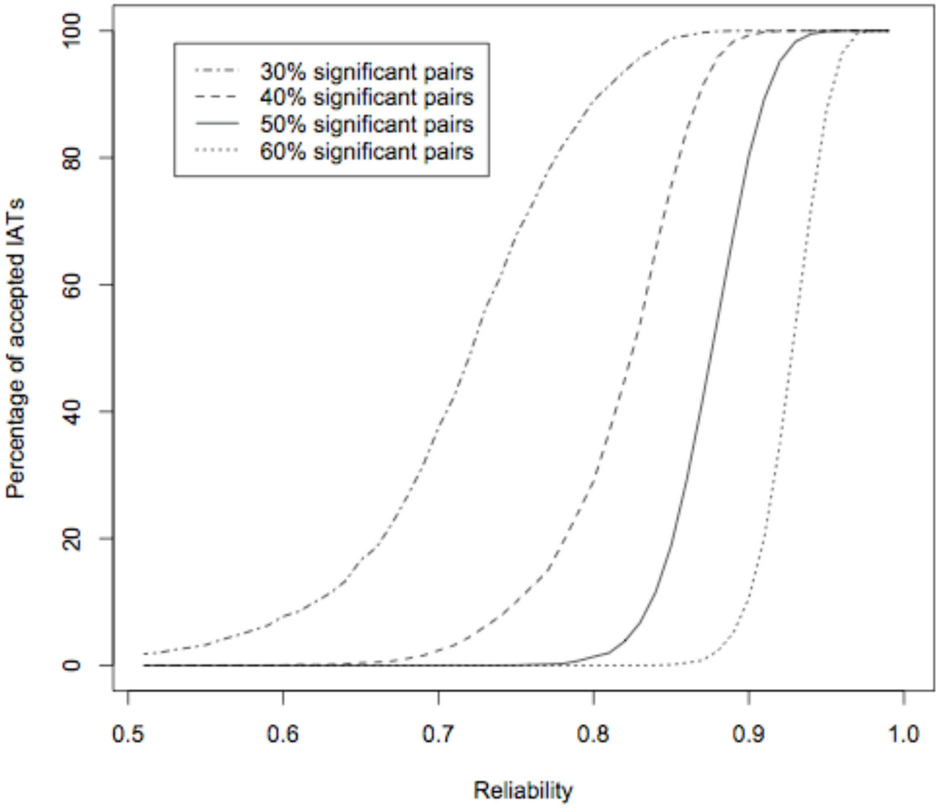

Relationship between Percentage of Significant Pairs and Reliability
Reliability
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Percentage of Significant Pairs

Figure 5



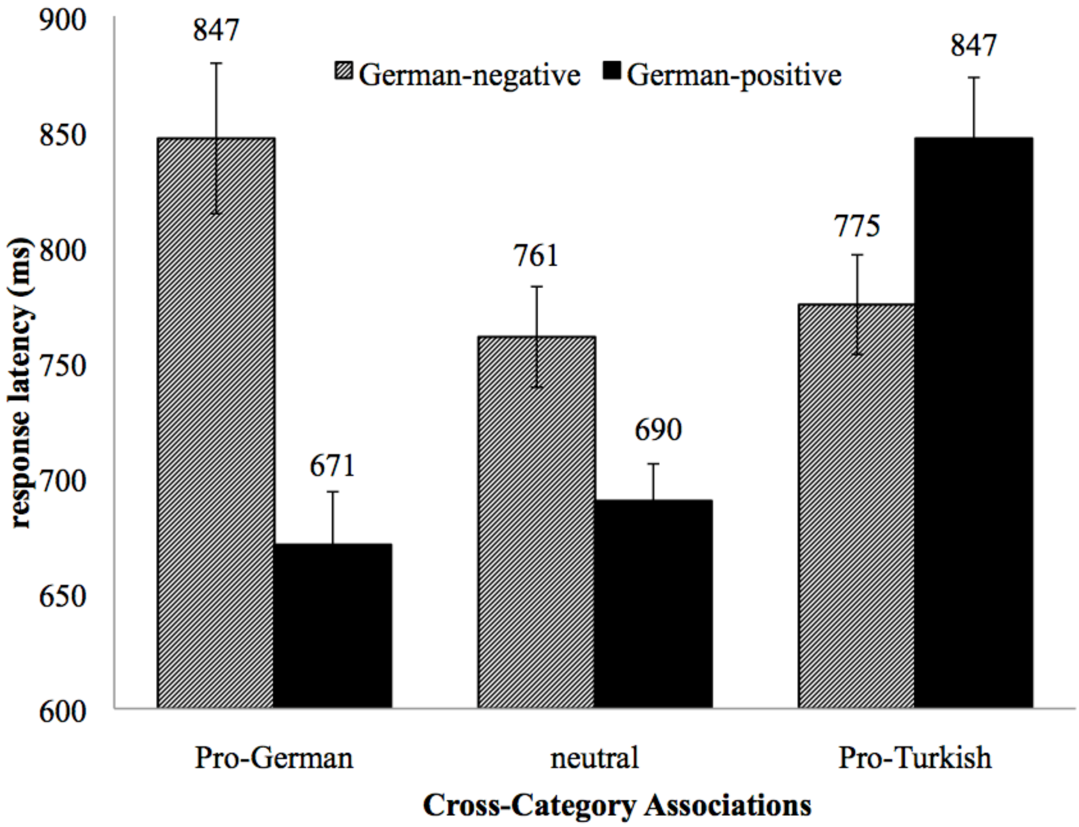
5th percentile of reliability as a function of the percentage of significant pairs for (from left to right) EVM of 5%, 15%, and 25%

Figure 6



The percentage of IATs with more significant pairs at various thresholds, as a function of reliability

Figure 7


900
850
800
750
700
650
600
response latency (ms)
▨ German-negative
■ German-positive
847
671
761
690
775
847
Pro-German
neutral
Pro-Turkish
Cross-Category Associations

Appendix

Attribute concept stimuli used in the Experiment (English translation in parentheses).

Pleasant stimuli associated with Germans: *Dichter* (poet), *Ostern* (Easter), *Weihnachten* (Christmas), *Demokratie* (democracy), *Umweltschutz* (environmental protection)

Unpleasant stimuli associated with Turks: *Folter* (torture), *Schmutz* (dirt), *Todesstrafe* (death penalty), *Macho*, *Gestank* (stink)

Pleasant stimuli (control): *Witz* (joke), *Schmetterling* (butterfly), *Liebe* (love), *Glück* (happiness), *Freude* (pleasure)

Unpleasant stimuli (control): *Angst* (anxiety), *Furcht* (fear), *Unfall* (accident), *Trauer* (mourning), *Schmerz* (pain)

Pleasant stimuli associated with Turks: *Basar* (bazaar), *Bauchtanz* (belly dance), *Dampfbad* (vapor bath), *Gastfreundschaft* (hospitality), *Süden* (south)

Unpleasant stimuli associated with Germans: *Hitler*, *Nazi*, *Skinhead*, *Herzinfarkt* (heart attack), *Fremdenhaß* (xenophobia)