A Comparative Investigation of Seven Implicit Measures of Social Cognition

Yoav Bar-Anan

Ben-Gurion University, in the Negev, Be'er Sheva

Brian A. Nosek

University of Virginia

Abstract

The present research compared the psychometric qualities of seven indirect attitude measures (implicit measures) across three attitude domains (race, politics and self-esteem) with a large sample (n = 23,413). We compared the seven measures on internal consistency, sensitivity to known effects, relationship with implicit and direct measures of the same topic, and sensitivity to outlier scores, and robustness to the exclusion of misbehaved or well-behaved participants. The overall psychometric performance from strongest to weakest was: Implicit Association Test (IAT), Brief IAT, Go-No go Association Task (GNAT), Single-target IAT, Affective Misattribution Procedure (AMP), Sorting Paired Features task, and Evaluative Priming. Further, the AMP showed a steep decline in its psychometric qualities when outlier scores were removed, and the GNAT's qualities were highly dependent on the inclusion of well-behaved participants and the exclusion of misbehaved participants. Relations among implicit measures and with criterion variables were weak for self-esteem, moderate for race and strong for politics. This suggests that some of the source of variation in reliability and predictive validity of implicit measures is a function of the concepts rather than the methods.

A comparative investigation of seven implicit measures of social cognition

The emergence of implicit social cognition in the 1990's and 2000's was accelerated by the invention of measurement methods that assessed social cognitions without requiring an act of introspection (Gawronski & De Houwer, in press; Gawronski & Payne, 2010; Greenwald & Banaji, 1995; Nosek, Hawkins, & Frazier, 2012). These measures share a signature feature of assessing social cognitions indirectly wherein the behavioral response is not a direct indicator of the evaluation. The social evaluation is inferred by comparing behavioral responses across two or more conditions. For example, in evaluative priming tasks (EPT; Fazio, Sanbonmatsu, Powell, & Kardes, 1986) target words appear one at a time and are evaluated as good or bad as quickly as possible. Immediately preceding the target words are primes that might automatically activate a positive or negative evaluation – such as images of prominent U.S. Democratic or Republican politicians. The indirect assessment of social evaluation is the average difference in time required to categorize good target words as good (and bad target words as bad) when preceded by a Democratic prime versus a Republican prime. Democrats may be faster to categorize good words (and slower to categorize bad words) when preceded by Democratic primes, whereas Republicans may be faster to categorize good words (and slower to categorize bad words) when preceded by Republican primes.

A substantial research literature using these indirect, or implicit, measures has emerged. This is particularly true for the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007), which through 2010 accounted for approximately half of the research use of implicit measures, and EPT, which accounted for about a fifth of the research applications (Nosek, Hawkins, & Frazier, 2011). The accumulated research literature shows considerable progress, particularly with the IAT and EPT, in establishing construct validity, identifying extraneous influences, demonstrating predictive validity, and identifying the component psychological processes contributing to measurement (for reviews see Cameron, Brown-Iannuzzi, & Payne, in press; De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Fazio & Olson, 2003; Gawronski & De Houwer, 2012; Gawronski & Payne, 2010; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Nosek, Smyth, et al., 2007; Nosek et al., 2011, 2012; Teige-Mociemba, Klauer, & Sherman, 2010; Wentura & Degner, 2010).

Despite increasing diversity of implicit measurement methods, there is much less knowledge about the psychometric properties of measures other than the IAT and EPT. Moreover, there is little systematic knowledge of the comparative psychometric qualities and performance of different implicit measures. Most research applications use just one implicit measure. If more than one implicit measure is used, it is usually for the purposes of conceptual replication in separate studies. There are relatively few studies that use multiple implicit measures in the same study and experimental setting thus creating a vacuum of comparative knowledge between implicit measures.

In this article, we report the results of a large investigation of the psychometric properties of seven measures of implicit social cognition across a variety of evaluation criteria. For most of the measures, this investigation is the most comprehensive test of reliability and validity conducted to date. For all of the measures, no prior research has compared them with as many other implicit measures. The present investigation allowed a direct comparison of the psychometric properties of the seven implicit measures with the same sample, same setting, and same criterion variables across three different content domains – race, politics, and self-esteem. The results provide insight into the psychometric qualities of the implicit measures and are

grounds for practical decisions about choice of implicit measures in research applications. The results also provide evidence regarding the variation of the relations among different implicit measures across different content domains.

Psychometric Properties of Implicit Measures

While use of implicit measures is dominated by the IAT and EPT, there are a variety of newer implicit measures that are gaining in popularity and have unique features making them attractive for research applications. Further, a diverse pool of effective implicit measures can facilitate theoretical development in implicit social cognition that is not constrained to the procedural idiosyncrasies of a single or small number of measurement methods. However, having multiple measures available is of limited use without knowledge of their comparative qualities. As such, there is both theoretical and practical benefit to enhancing knowledge of the psychometric properties of a variety of implicit measures.

In addition to the IAT and EPT, we investigated the Affect Misattribution Procedure (AMP; Payne, et al., 2005), Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009), Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), Single-Target Implicit Association Test (Karpinski & Steinman, 2006; Wigboldus, Holland & van Knippenberg, 2004), and the Sorting Paired Features task (SPF; Bar-Anan, Nosek, & Vianello, 2009). These were selected because (a) they are adaptable for measuring different types of associations making them applicable to a variety of research applications (unlike, for example, the name-letter task that is only for self-esteem; Nuttin, 1985), (b) outside of the IAT and EPT, three of them – GNAT, AMP, and ST-IAT – were the most frequently cited implicit measures in 2010 (Nosek et al., 2011), and (c) the other two, SPF and BIAT, are new measures – the first having unique measurement qualities, and the second being a derivative of the IAT has similar characteristics but shortened administration format and perhaps some unique psychometric qualities (Nosek, Bar-Anan, Sriram, & Greenwald, 2012; Sriram & Greenwald, 2009). Beyond the emerging evidence from accumulating research applications, systematic evaluation of the psychometric properties of these five additional measures is scarce.

Besides adding to the psychometric evaluation of individual measures, a key contribution of this article is inter-relations assessment. Considering the substantial progress in other aspects of implicit social cognition it is surprising how little is known about the relations among implicit measures. The fact that two measures are called "implicit" does not itself guarantee that they are influenced by similar psychological processes, predict similar behaviors, measure the same construct, or even correlate with one another. Indeed, the most cited comparative investigation found weak relations among multiple measures of implicit self-esteem (Bosson, Swann, & Pennebaker, 2000). Part of the lack of relationship was attributable to weak reliability of some of the measures. For example, moderately strong relations have been observed between IAT and EPT measures of racial attitudes after accounting for measurement error with latent variable analysis (Cunningham, Preacher, & Banaji, 2001). And, subsequent investigations that used more reliable implicit measures have demonstrated stronger interrelations among the two or three measures investigated (Bar-Anan, Nosek, & Vianello, 2009; Greenwald Smith, Sriram, Bar-Anan, & Nosek, 2009; Karpinski & Steinman, 2006; Krause, Back, Egloff & Schmukle, 2011; Payne, Govorun & Arbuckle, 2008; Ranganath, Smith, & Nosek, 2008; Rudolph, Schröder-Abé, Schütz, Gregg & Sedikides, 2008). Even relations among self-esteem measures in the original Bosson study were stronger when accounting for measurement error (Lane, Brescoll, & Bosson, 2001). On the other hand, experimental investigation of the psychological processes influencing implicit measures suggests that there are idiosyncratic influences that will likely influence their interrelations (Deutsch & Gawronski, 2009; Gawronski,

Cunningham, LeBel & Deutsch, 2010; Payne, Burkley & Stokes, 2008; Sherman, 2009). Some of the idiosyncrasies may be extraneous "methodological" influences. Others may reflect unique components of implicit social cognition showing that no one measurement method assesses the whole construct. A future investigation of structural relations in the present data will examine whether relations among implicit measures indicate a shared latent variable that is attitudinal and unique to some or all implicit measures (Bar-Anan, Spies, & Nosek, 2012). In the present research we provide evidence whether (and which) implicit measures relate to each other in three attitude domains.

The present research is the first test of the effect of attitude domain on relations among different implicit measures. A large study that used only a single implicit measure found variation in the relationship between the IAT and explicit attitude measures as a function of the attitude domain (Nosek, 2005). The present research repeats Nosek's test with fewer topics (three instead of 57) but with more implicit and explicit measures. Based on previous research and theory (Nosek, 2005; Bosson et al., 2000), we predicted that politics would elicit the strongest implicit-explicit relations, and that self-esteem would show the weakest implicit-explicit relations. However, there was no consistent pattern of results in previous research that allowed a strong prediction regarding variations in the relations among implicit measures. Nosek (2005, 2007) interpreted the effect of attitude domain on implicit-explicit relations as revealing insights about the interplay between implicit and explicit cognition. However, if the same variation would be found for relations among implicit measures, then Nosek's insights might apply to relations among attitude measures in general, rather than implicit-explicit relations.

Finally, besides information about individual measures and relations among implicit measures, there is much to learn about the comparative properties of implicit measures as attitude measures. For instance, which measures perform better in criteria such as sensitivity to detecting known effects, predicting criterion variables, and robustness to data exclusion and outlier analyses. For example, De Houwer and De Bruycker (2007) reported evidence that the IAT outperformed the Extrinsic Affective Simon Task (EAST) on internal consistency and relationship with other measures. Similarly, Payne, Govorun and Arbuckle (2008) and Spruyt, Hermans, De Houwer, Vandekerckhove, and Eelen (2007) compared implicit measures on prediction of criterion variables. However, each investigation was constrained by examining just two or three implicit measures, on one or two psychometric criteria, with just a single focal topic, and using relatively small samples. The present investigation is more expansive on all four of these constraints. We examined seven implicit measures, on seven evaluation criteria, across three topics, using a very large sample.

Study Overview

One reason for the paucity of comparative research on implicit measures is likely practical – it requires substantial resources to conduct such studies. Each implicit measure requires a non-trivial amount of time to administer and is mentally taxing to complete. For example, in a laboratory data collection participants were required to complete more than a dozen implicit measures across two sessions of one hour each (Smyth & Nosek, 2005). The most notable finding was that participants found the experience exhausting, and the data quality declined rapidly over the course of the sessions. Further, large samples are necessary to obtain reliable estimates for comparing across many measures and topics. These constraints may be prohibitive for ordinary laboratory resources.

We had a unique opportunity to address these practical challenges via a public website that attracts a high volume of participants. As such, we pursued a systematic evaluation of implicit measures across multiple topics with multiple evaluation criteria. And, we were able to

avoid over-taxing individual participants by implementing a planned incomplete design. More than 24,000 participants each completed a random subset of the available implicit measures. This allowed for comparison among all measures despite the fact that any given participant completed only a few of the possible measures. Parallel explicit measures and criterion variables for each topic were also included for psychometric evaluation.

Evaluation Criteria

We evaluated the measures on multiple criteria: internal consistency, test-retest reliability, sensitivity to known-groups effects, relations with other implicit measures of the same topic, relations with direct measures of the same topic, relations with other criterion variables, and sensitivity to outliers and data exclusion.

Internal consistency. It is desirable for a measure to have little random error during task performance to elicit strong internal consistency. A presumption of this criterion is that the internal consistency is not due to an extraneous influence, but rather reflects assessment of the construct of interest. On its own, it is not possible to tell whether stronger internal consistency is indicative of greater construct sensitivity. However, the use of multiple criteria helps to clarify this. For example, if strong internal consistency is due to extraneous influences, then that implicit measure will not also show strong relations with other measures of the same construct. Whereas, if the strong internal consistency is due to effective construct measurement, then the increased reliability will facilitate relatively stronger relations with criterion measures.

Test-retest reliability. Similar principles apply to test-retest reliability. If a measure assesses stable elements of a construct, then stronger test-retest reliability is a positive indicator that the measure is subject to less random error[1]. This is a relatively weak criterion in the present investigation because test-retest reliability was only assessed among those participants that completed multiple sessions and were randomly assigned to complete the same measure again. The average sample size of the same participant completing the same implicit measure of the same topic was 116. Even though this is probably above the average sample size for psychological science, meeting such a standard is a rather weak achievement for obtaining highly reliable estimates. Moreover, most of the retest data was collected within an hour of the original test. Interest in test-retest reliability, especially for distinguishing stable and transient components, requires a longer average time between tests. Even so, because even the short time scale has some information value, we report test-retest reliability, but do not include it as a primary evaluation criterion.

Sensitivity to known effects and group differences. All else being equal, better measures will be more sensitive to detecting known effects. For example, measures of self-esteem – both direct and indirect – consistently demonstrate a main effect of positivity toward the self (Nosek, Banaji, & Greenwald, 2002; Yamaguchi et al., 2007). Holding characteristics of the setting and sample constant, measures that are maximally sensitive to detecting the association of self with good will elicit the strongest effect size. Two topics – self-esteem and racial attitudes – elicit consistent main effects across measures. The latter indicating an easier time associating good with white people compared to black people.

In addition to detecting main effects, better measures are more sensitive to detecting known differences between social groups. For example, theory and evidence support the contention that Black and White participants should differ in their racial attitudes – Whites being

relatively more favorable to White people, Blacks being relatively more favorable to Black people, even implicitly (Fazio, Jackson, Dunton, & Williams, 1995; Nosek, Smyth, et al., 2007; Payne et al., 2005). So, better measures ought to be more sensitive to detecting this difference. And, with political attitudes, differences in implicit evaluations between Democrats and Republicans for their political parties are observed (Lindner & Nosek, 2009; Smith, Ratliff, & Nosek, in press). However, the status of group differences with self-esteem is less clear. Theory and evidence is equivocal about whether groups – such as people from Eastern versus Western cultures – will demonstrate mean differences in self-esteem when measured implicitly (Yamaguchi et al., 2007). As such, we included only racial and political attitudes for evaluation of known-group differences.

A challenge for the main effects criterion is that any particular measure may show a larger main effect than another either because of sensitivity to the target construct, or because of greater sensitivity to an extraneous influence that covaries with the construct main effect. This concern is less relevant (but not completely irrelevant) for comparison of known group differences. However, as with reliability, our use of multiple criteria offers a means of addressing this ambiguity. If a measure's stronger main effect is a function of construct-irrelevant influences, then it will not be accompanied by stronger performance on other evaluation criteria.

Correlation with other implicit measures of same topic. To the extent that indirect measures are influenced by the same construct(s), better measures should be more related to other implicit measures. This criterion is straightforward, but with an important qualification. Two implicit measures could both be valid but assess different components or qualities of the construct. For example, Olson and Fazio (2003) suggested that EPT is more sensitive to the individual stimulus items, and the IAT is more sensitive to their superordinate categories. As such, if the items lead to a different evaluation than the categories, then a lack of relationship may indicate assessment of different constructs, not lack of validity of one or both. In the present design, this can be addressed directly because each measure can be compared to many other measures – both direct and indirect – each with unique features to provide more confidence in construct validation.

Another challenge is that two (or more) measures could have a shared extraneous influence that produces covariation between them that has nothing to do with the construct. This is most obviously a possibility for the measures based on response latency because of the potential impact of average response latency and its associated constructs – cognitive fluency, task-switching ability (Mierke & Klauer, 2003; Sriram, Greenwald, & Nosek, 2010). The AMP is the only indirect measure that does not use response latency as a dependent variable, and the AMP and EPT are the only measures that do not require categorization of stimuli into superordinate categories. These factors could disadvantage these measures in particular on comparisons of intercorrelations among measures. We can, however, address this challenge with the present design. For example, if naming the superordinate category is a critical feature for eliciting distinct evaluative processes, then the AMP and EPT should relate more strongly with each other than they do with the other measures.

Finally, because the relationships among implicit measures has been underinvestigated, this criterion offers an opportunity to examine the interrelations of seven implicit measures across three topics.

Correlation with direct measures of same topic and other criterion variables. To the extent that there is a meaningful relationship between direct and implicit measures of the same

topic, better measures will be more sensitive to detecting it. This holds even if the measures are assessing distinct constructs. For example, height and weight are distinct constructs that are positively correlated. The best measures of each will minimize error and maximize the correlation between height and weight closest to its true relationship in that sample (see also Greenwald, Nosek, & Banaji, 2003; Nosek et al., 2012).

Evaluation models (e.g., Gawronski & Bodenhausen, 2006; Fazio, 2007; Wilson, Lindsey, & Schooler, 2000) and existing psychometric evidence (e.g., Nosek & Smyth, 2007) suggests that implicit and direct (self-report) measures assess distinct, but related constructs. As such, like assessments of height and weight, the measures that are best able to measure the constructs will elicit the strongest relationship between the variables – closest to its "true" relationship. No theory anticipates that implicit and explicit social cognitions are exclusive of one another – that is entirely unrelated intra- and inter-individually. Indeed, it would be quite surprising if, for example, implicit and explicit racial evaluations were entirely unrelated. Theory and evidence anticipate that both would be influenced by factors such as one's own racial identity and one's experience with the culture and members of each racial group. Indeed, with a sample of more than 700,000 self-reported racial attitudes were moderately correlated (r = .31) with the racial attitude IAT (Nosek, Smyth, et al., 2007), and in samples subjected to latent variable analysis, that estimate exceeds r = .40 (Nosek, 2007).

Nonetheless, there is an important challenge with using direct measures as an evaluation criterion across multiple implicit measures. Each measure has a unique procedure and may engage distinct psychological processes. As a consequence, it is possible that the indirect, implicit measures vary in the extent to which they are influenced by *direct* evaluation. If, for example, participants in the AMP evaluate the primes instead of the targets, then it is actually behaving as a direct measure instead of an indirect one (Bar-Anan & Nosek, in press).

So, on its own, variation in correlations with direct measures is ambiguous as a criterion. However, there is a relatively straightforward solution. If an implicit measure is actually a direct measure in disguise, then the stronger correlations with direct measures could be accompanied by weaker correlations with other implicit measures. If, on the other hand, the implicit measure is simply a more effective measure then the stronger correlation with direct measures will likewise be accompanied by stronger correlations with the other implicit measures than they have amongst themselves.

Sensitivity to outliers. A standard presumption of measurement is that each participant assessment contributes equally to the observed effects. Regression diagnostics, for example, identify individual data points that exert inordinate impact (leverage) on observed effects qualifying the interpretation of the result as characterizing the whole sample rather than a few outliers. As such, the psychometric properties of better measures will be less sensitive to the removal of outliers. If outliers are exerting an inordinate influence on the properties of a measure, then the effects are not indicative of the respondents as a whole, but of a subset of them.

Effects of data exclusion. With performance assessments, respondents must follow task instructions or else interpretation of the assessment may be compromised (e.g., Nosek & Hansen, 2008). If, for example, participants in the Stroop task (Stroop, 1935) ignore the instructions and name the color words rather than their ink color, then the results will not indicate the response interference of word meaning on color naming. Ideally, instructions are clear and easy-to-follow so that these events do not occur. And, in the event that respondents fail to follow instructions, ideally there is a means of identifying the misbehavior so that the assessment can be removed

prior to analysis. We expect the psychometric performance of the measures to improve when removing participants that are suspect of misbehavior. Yet, it is desirable to have measures that provide interpretable data from the largest proportion of respondents as possible to avoid (a) reducing power and (b) potentially biasing the sample if exclusion is more likely among some participants more than others (e.g., high versus low intelligence; high versus low conscientiousness).

A challenge with this criterion is that across tasks, the relevant exclusion criteria may vary as a function of standard practice for each measure. As such, we examined the effect of applying task relevant exclusion criteria for each measure such that none, 1%, 2%, 5%, 10%, 15%, 20%, 25%, and 50% of the sample was excluded. Better measures would show stronger psychometric performance overall, and also be more robust to the applied exclusion rule even when only a small portion of data is excluded. For example, if the SPF performs relatively poorly until 30% of the sample is removed, whereas the IAT performs consistently strongly across exclusion levels, then it would suggest that the IAT is robust to exclusion criteria, but that the SPF is dependent on dropping a significant proportion of the sample to show strong performance. The change in psychometric performance based on the data exclusion levels indicates the extent to which each measures' performance is reliant on the exclusion practices.

Method

Participants

24,015 participants started at least one of the measures in the study. Of those, 23,413 (97.5%) completed at least one measure, 8.7% completed only one measure, 4.9% completed 2 measures, 7.7% completed 3 measures, and 31% completed 4 measures. 45.1% completed more than four measures, of which 10% completed more than 10 measures. Among the participants who completed at least one measure (63% women, 36% men, 1% unknown; mean age = 29.1, SD = 12.0) the reported racial origins were: 0.6% American Indian, 3.3% Asian or Asian American, 6.2% Black (not of Hispanic origin), 7.8% Hispanic or Hispanic American, 70% White (not of Hispanic origin), 6.5% multi-racial, 1.8% other, and 3.2% did not identify. 79% reported US citizenship and 20% reported citizenship of other nations.

Materials

Stimuli

Attitude objects stimuli. The same stimuli appeared in all the implicit measures (the exemplars in the IAT, BIAT, GNAT, ST-IAT and SPF; the primes in the AMP and EPT). The race stimuli were 6 pictures of white people (3 females, 3 males), and 6 pictures of black people (3 females, 3 males). The pictures were taken from 1998-99 NBA and WNBA basketball player and coach image repositories, selecting individuals who were unlikely to be recognized by most people (Nosek & Banaji, 2001). For those measures that used category names (IAT, BIAT, GNAT, ST-IAT and SPF), the race category labels were *White People* and *Black People*.

The politics stimuli were pictures of American politicians: 5 Democrats (Barack Obama, Hillary Clinton, Bill Clinton, Al Gore, and John Kerry) and 5 Republicans (George W. Bush, George H. W. Bush, Ronald Reagan, Condoleezza Rice, and Rudy Giuliani). The category labels were *Democrats* and *Republicans*. The self-esteem stimuli were words pertaining to the two category labels Self (*I*, *Me*, *Mine*, *Myself* and *Self*) or *Others* (*They*, *Them*, *Their*,

Theirs, and *Others*). The AMP also included a control prime stimulus—a gray rectangle when the primes were pictures, and the letters XXXXX when the primes were words.

Attribute stimuli. The category labels for the attribute categories in the IAT, BIAT, GNAT, ST-IAT and SPF were *Good Words* (items: *Paradise, Pleasure, Cheer, Wonderful, Splendid, Love*) and *Bad Words* (items: *Bomb, Abuse, Sadness, Pain, Poison, Grief*). In the EPT, the attribute category labels were *Good* (items: *Paradise, Pleasure, Cheer, Friend, Splendid, Love, Glee, Smile, Enjoy, Delight, Beautiful, Attractive, Likeable, Wonderful*) and *Bad* (items: *Bomb, Abuse, Sadness, Pain, Poison, Grief, Ugly, Dirty, Stink, Noxious, Humiliate, Annoying, Disgusting, Offensive*). In the AMP, the target stimuli were 72 Chinese Pictographs, and a black and white noise stimulus was used as a mask (all from Payne et al., 2005).

Implicit Measures

All the implicit procedures were tested prior to the study to make sure that they showed psychometric qualities similar to published reports. We used the best available design features based on the present knowledge and the practical constraints of the study (time, accuracy, and need for clear and succinct instructions; see Table 1 for the number of trials contributing to the calculated score for each measure).

Common procedural features. Before each task, instructions oriented participants to the measure and performance rules. The first page of instructions explained the key features and included an image illustrating one trial with dogs and cats as the attitude objects (excluding the AMP, EPT because these had more than one display per trial). The categories and the stimuli that belonged to each category were presented for all measures in which the stimuli had to be categorized into their superordinate category (all except AMP, EPT).

The background screen color was always black. Each response block started with brief instructions. In tasks that used two keys (all but the SPF and the GNAT), the keys were always 'e' for a left response and 'i' for a right response. The names of the categories appeared on the top-left and top-right corners of the screen throughout each block. Attribute category labels and stimulus words appeared in green, attitude object labels and words in white. The inter-trial interval (ITI) was 250ms (275ms in the EPT). In each trial of the IAT, BIAT, ST-IAT, and the SPF, the target stimulus appeared until correct response, and an incorrect response was followed by a red X that remained on the screen until the participant responded correctly. In the IAT, BIAT and the GNAT, the sequence of the trials alternated between attitude objects and attribute stimuli: a trial presenting an attitude object stimulus was always followed by a trial presenting an attribute word. In all the tasks, the stimuli for each category were selected randomly until all the stimuli of that category were displayed. Then, each category stimulus "pool" was refilled and the stimuli were randomly selected again one by one. The unique features of each task are summarized next, and all tasks – exactly as they were administered in the study - can be experienced at: https://implicit.harvard.edu/implicit/user/yba/mtmmd/tasks2.htm.

Table 1

Number of Trials that Contributed to the Calculated Score

Measure	IAT	BIAT	GNAT	ST-IAT	SPF	EPT	AMP
N trials	120	128	160	192	120	180	48

Notes. The 48 trials of the AMP do not include 24 trials with the neutral prime; The 160 trials of the GNAT included 64 "No-go" trials that did not provide any latency data (but error-rate was combined to the score).

Implicit Association Test. The IAT procedure followed the one described in Nosek et al. (2007). Words and images were presented one at a time at the center of the screen with category labels at the top-right and top-left corners. Participants were instructed to respond as fast they could while making as few mistakes as possible. In the first response block (20 trials)

participants categorized items representing the two attitude objects (e.g., Democrats vs. Republicans). In the second block (20 trials) participants categorized good and bad words. The third block (20 trials) was a combination of blocks 1 and 2: for example, participants categorized Democrats and good words with one key and Republicans and bad words with the other key. The fourth block (40 trials) was the same as the third block. In the fifth block (40 trials), the attitude objects switched sides: the object that was categorized with the left key in block 1-3 was now categorized with the right key, and the object that was categorized with the right key was now categorized with the left key. The sixth block (20 trials) and the seventh block (40 trials) were a combination of blocks 2 and 5: for example, participants now categorized Republicans and good words with one key and Democrats names and bad words with the other key. The order of the pairing (which pairing appeared on blocks 3-4 and which appeared on blocks 6-7) was randomized between participants.

Brief Implicit Association Test. The BIAT procedure followed the one described in Sriram and Greenwald (2009), but with a different block sequence. In the instructions slide, the BIAT was presented to the participants as "the IN-or-OUT game." Words and images were presented one at a time with the two "IN" categories at the top of the screen. Participants hit the right-response key when they saw an item belonging to the "IN" categories and hit the left-response key when the item did not belong to those categories. The stimuli and response format was the same as the IAT. The key difference between the IAT and the BIAT is that only two "focal" categories are named and appear on screen. The "OUT," non-focal stimuli always belonged to the two contrasting categories. For instance, if the focal categories were *Good words* and *Republicans*, the items that did not belong to these categories were the items representing *Bad words* and *Democrats*.

The BIAT used only four stimuli from each category (as in Sriram & Greenwald, 2009). The BIAT sequence included nine blocks of trials. In each block, the first four trials were selected from the target categories (e.g., Democrats, Republicans). The remaining trials for each block alternated between target categories and attributes (good, bad items). The first block was a practice round of 16 total trials with *mammals* and *birds* as target categories and *good* and *bad* as attribute categories. The other eight blocks began with the four category-only warm-up trials, and then presented 16 category-attribute alternating trials. The 2nd through 5th blocks had the same focal attribute (e.g., *Good words*) and alternated the focal category (e.g., *Democrats, Republicans*) such that one appeared in blocks 2 and 4, and the other appeared in blocks 3 and 5. The 6th through 9th blocks had the other attribute focal (e.g., *bad*) and likewise alternated the focal category between blocks. To iterate an example for one of the possible block sequences, the focal categories were: *Republicans* and *Bad words* in blocks 2 and 4, *Democrats* and *Bad words* in blocks 3 and 5, *Republicans* and *Good words* in blocks 6 and 8, and *Democrats* and *Good words* in blocks 7 and 9. The order of attributes and categories as focal was randomized between subjects resulting in four between-subjects conditions (good or bad first; Democrats or Republicans first) for each topic.

Go/No-go Association Task. The GNAT procedure was based on Nosek and Banaji (2001), designed for scoring based on response latencies rather than error rates. The GNAT was presented as the "HIT-or-HOLD game." Words and images were presented one at a time with the two "HIT" categories presented at the top of the screen. Participants hit the space-bar key when they saw items belonging to the "HIT" categories," and did nothing when they saw an item that did not belong to those categories. The only differences between the GNAT and the BIAT is that there was a response deadline for the items, participants did nothing for items that did not belong to the categories on screen, and participants did not need to correct their errant responses.

The GNAT sequence included 9 blocks (with 20 trials each) that paralleled the block sequence of the BIAT. The first block was a practice round with dogs and cats as the attitude objects. The 2nd through the 5th block had the same focal attribute (e.g., *Good words*), and alternated the focal category (e.g., *Democrats, Republicans*) such that one was focal in blocks 2 and 4, and the other was focal in blocks 3 and 5. The 6th through 9th blocks had the other attribute focal (e.g., *Bad words*) and likewise alternated the focal category between blocks. The order of attributes and categories as focal was randomized between subjects resulting in four between-subjects conditions for each topic (e.g., in politics: good or bad first X Democrats or Republicans first). When the target item belonged to one of the non-focal categories, it was presented for 950ms. If participants erroneously responded during such a trial, the red X appeared for 150ms. When the target item belonged to one of the focal categories, the response deadline was 1200ms. If participants did not hit the space until the deadline expired, a red "Please respond more quickly!" message appeared for 150ms. Because we planned the scoring to rely exclusively on latency data on responses to focal categories, each block presented more focal trials (12) than non-focal (8) trials.

Single-Target Implicit Association Test. There is currently no consensus on one ST-IAT procedure (e.g., Bluemke & Friese, 2007; Karpinski & Steinman, 2006; Wigboldus, Holland & van Knippenberg, 2004). We used Karpinski and Steinman's procedure with modifications that are used to maximize reliability and validity in IAT formats (Cunningham, et al., 2001; Greenwald, et al., 2003; Nosek, Greenwald, & Banaji, 2005, 2007). The main modifications: no response deadline, no correct-response feedback, and participants were required to correct errant responses. Prior to this study, we compared this procedure with the one used in Karpinski and Steinman (2006) in two studies with large samples (N's = 2087, 1064), and found no difference in the psychometric qualities. The instruction slides were identical to those used for the IAT, except that the image illustrated an ST-IAT trial.

The ST-IAT is similar to the IAT, but instead of two attitude object categories, only one attitude object is presented. The task consisted of four blocks (48 trials in each block). In the first block, participants sorted items that belonged to three categories – Good words (using the right key), Bad words (left key) and an attitude object (e.g., Democrats). The attitude object shared a key with one of the two attribute categories. In Block 2, the attitude object category changed side. For instance, if in the first block the attitude object was categorized with the same key as good words, then in the second block, it was categorized with the same key as bad words. Block 3 was identical to Block 1, but the attitude object was replaced with the contrasting attitude object (e.g., Democrats was replaced with Republicans). Block 4 was identical to Block 2, but with the same attitude object as Block 3. In other words, these were two ST-IATs, one for each attitude object. There were four block order conditions for each topic, randomized between participants: which attribute shared a key with attitude object in block 1 and 3 (Good words vs. Bad words) and which attitude object was presented in blocks 1-2 (e.g., Democrats vs. Republicans). Each block presented 14 trials with the attitude object items, 14 trials of the attribute category items that were sorted with the same key response as the attitude object, and 20 trials of the other attribute category items. The goal of this imbalance was to decrease response bias influences for the left versus right key (28 trials were categorized with one key, 20 trials with the other).

Sorting Paired Features. The SPF procedure followed the one described in Bar-Anan et al. (2009). Item pairs appeared in the middle of the screen, and they were sorted into category pairs appearing in the four corners. The response keys were 'W', 'C', 'O' and 'M' for the top-left, bottom-left, top-right and bottom-right, respectively. The four category pairs were all the possible combinations between the attitude object categories and the attribute categories (e.g.,

Good words+Democrats, Bad words+Democrats, Good words+Republicans, Bad words+Republicans). For instance, a trial could present the stimuli "Awful" and the image of John Kerry (presented one below the other), and the correct response would be to categorize these two stimuli, with one of the 4 key responses, to the corner showing Bad word+Democrats labels. The two pairs that included good words were always presented at the top-left and bottom-left corners, and bad word pairs at the top-right and bottom-right. Two pairs that included one of the attitude objects (e.g., Good words+Democrats, Bad words+Democrats) were presented at the top-left and top-right corners, and the other two were at the bottom-left and bottom-right. Which category was presented at the top and which at the bottom was randomized between participants. The task consisted of three identical blocks, each with 40 trials – 10 trials for each of the four object-attribute pairs.

Evaluative Priming Task. The procedure followed the one described by Fazio et al. (1995). In the instructions slide, participants were informed that they would be presented with a set of words to evaluate as either "Good" or "Bad." They were instructed to categorize the words as quickly as they can while making as few mistakes as possible. In the first block, primes did not appear, and participants categorized each of the 28 target words. The target words stayed on the screen until participants responded or until 1500ms had passed, whichever came first. The left key was used to classify targets as *Bad words*, and the right key as *Good words*. If participants failed to respond before the 1500ms response deadline expired, a message "Please respond faster!" appeared in red font for 300ms, and the trial ended without allowing a response. If participants responded incorrectly, a red X appeared for 300ms.

Primes appeared in blocks 2-4. At the onset of the second block, participants were informed that before each green target word, an image (a white word in the self-esteem task) would be presented. Participants were instructed to try to remember the primes for a later memory test. The primes appeared for 200ms, followed by a blank screen for 50ms, and then the target. The other durations were the same as in the first block. The three blocks had 60 trials each (15 trials for each prime category/target category combination). A final block tested participants' memory with 16 trials – presenting 8 primes and 8 new stimuli that belong to the prime categories – and having participants identify “old” or “new”.

Affective Misattribution Procedure. The procedure followed the one described by Payne et al (2005). Instructions clarified that, for each trial, two or three images would appear one after the other in rapid succession. Three images illustrated the sequence from left to right, a white man (who did not appear in the task), a Chinese pictograph and the mask. Participants were instructed to ignore the first image and evaluate whether the Chinese drawing is more or less pleasant than the average Chinese drawing. Targets were rated as pleasant with the left key and unpleasant with the right key. The first block consisted of three practice trials. In the practice, the prime was presented for 125ms, immediately followed by the target stimulus that was presented for 150ms before it was replaced with the mask stimulus. After the practice, instructions reminded participants of the rules and also cautioned them to "evaluate each Chinese drawing and not the image that appears before it. The images are sometimes distracting." Then, two blocks followed, each with 36 trials (12 for each of the two prime categories and 12 for the control prime category). In those blocks, the prime exposure duration was 75ms and the target exposure duration was 100ms.

Direct Attitude Measures

Self-reported preference. Participants were asked: "Which statement best describes your personal feelings toward U.S. Democrats [Black people][yourself] and Republicans [White people][other people]?" There were 7 response options, ranging from strong, moderate, or slight

preference for one attitude object over the other to no preference between the two objects in the middle, to a slight, moderate, or strong preference of the opposite direction.

Feeling thermometers. Participants were asked: "Please rate how warm or cold you feel toward the following groups (0 = coldest feelings, 5 = neutral, 10 = warmest feelings)." The groups in each self-report measure were: Race: Black People and White People; Politics: Democrats and Republicans; Self-esteem: myself and others.

Item ratings. There were two item ratings questionnaires: one for race and one for politics. Participants were asked to rate how warm or cold they feel toward each person represented in the stimulus items used in the implicit measures (0 = coldest feelings, 4 = neutral, 8 = warmest feelings). The people were presented together on the same page, and participants rated each of them separately.

Speeded Self-report (SR). In the speeded self-report, participants rate attitude objects very rapidly. Although this is a direct measurement, participants may have reduced ability to control it, which might make it more sensitive to automatic evaluation (Ranganath et al., 2008). The procedure was based on the one described by Ranganath and colleagues, with some modifications to allow easier responding. Items were presented for one second, requiring participants to rate them very quickly on a scale of 1 (most unfavorable) to 4 (most favorable) using the keys 1, 2, 3, and 4. Participants evaluated each item as accurately and rapidly as they could. To do so, participants were encouraged to go with their gut response. The task consisted of a 16-trial practice block followed by a 60-trial block. In the practice block, each of the practice items (the words *Weddings, God, Science, Apples, Lemons and Orange juice,* and images of a chair, a lamp and an umbrella) was presented for 1200ms, right after a 125ms "XXXXXXXX" mask (some of the practice items appeared twice). If participants did not respond before the 1200ms deadline expired, a red "Please respond faster!" message appeared for 600ms. In the second block, the response deadline was shortened to 1000ms. Twenty trials presented filler stimuli, and 20 trials presented stimuli of each attitude object (the same stimuli used in the indirect measures). Each stimulus appeared 2-3 times.

Modern Racism Scale (MRS). The MRS (McConahay, 1983, 1986) is a popular self-report measure of racial attitudes. While it was designed to be indirect, most interpretations of the scale suggest that its goal is transparent and, therefore, likely direct (e.g., Fazio et al., 1995). Because not all participants were U.S. citizens, the two last words in the statement "Discrimination against Blacks is no longer a problem in the U.S." were replaced with the words "My country." For this and the next two scales, participants rated their agreement with each item from 1 (strongly disagree) to 6 (strongly agree). The items were presented in random order.

Rosenberg Self-Esteem (RSE). The Rosenberg Self-Esteem scale (Rosenberg, 1965) measures people's feelings of global self-worth with 10 items. It is the most widely-used measure of self-esteem.

Right-wing Authoritarianism (RWA). The RWA (Altemeyer, 1981, 1996) is a 15-item measure that is strongly related to conservatism and self-reported identification with Republicans over Democrats (Jost, Glaser, Kruglanski, & Sulloway, 2003).

Other Criterion Measures

Reported Contact with Black people. Participants were asked: "think about the time you spend interacting closely with others (NOT including immediate family members, and NOT including passing, casual interactions). How much of this time (say, over the last month) includes close interactions with Black people?" There were 10 response options ranging from *All* to *None*.

Voting behavior. Participants reported whether they had voted and which candidate they voted for in the most recent past U.S. presidential election (2004).

Voting intention. Participants reported which candidate they would vote for in the 2008 elections, if "all the candidates listed below were on the ticket." The list included all the politicians that had declared their candidacy during the primary season in late 2007, with 8 Democrats and 11 Republicans.

Exploratory criterion measures. We also included a few novel measures of self-esteem that we developed in an exploratory attempt to add more criterion measures to this topic. However, we found very little evidence that these measured self-esteem, so we excluded them from all the analyses.

Procedure

The study was administered via the research Web site for Project Implicit (https://implicit.harvard.edu; see Nosek, 2005, for more information) between November 6, 2007, and May 30, 2008. It was open to the Internet public, and participation was voluntary. Participation in research at the Project Implicit website required identity registration with a demographic questionnaire. Each time they logged in, participants were randomly assigned to a study in the Project Implicit study pool, including this study. Unlike other studies in that pool, it was possible to be randomly assigned to this study more than once (up to 32 times).

The procedure was constructed such that each session should be approximately 15 minutes. Measures were randomly selected for each session with a constraint that there were always 2 "long-duration" measures and 2 "short-duration" measures, and the same measure (i.e., same method and topic) could not be selected twice in a single session. Figure 1 presents these two groups of measures and illustrates their selection for a session. Participants could initiate additional sessions, and could receive identical measures from previous sessions to facilitate test-retest comparisons. At the end of each session, the purpose of the study was explained to the participants, and they receive result feedback for the implicit measures they performed.

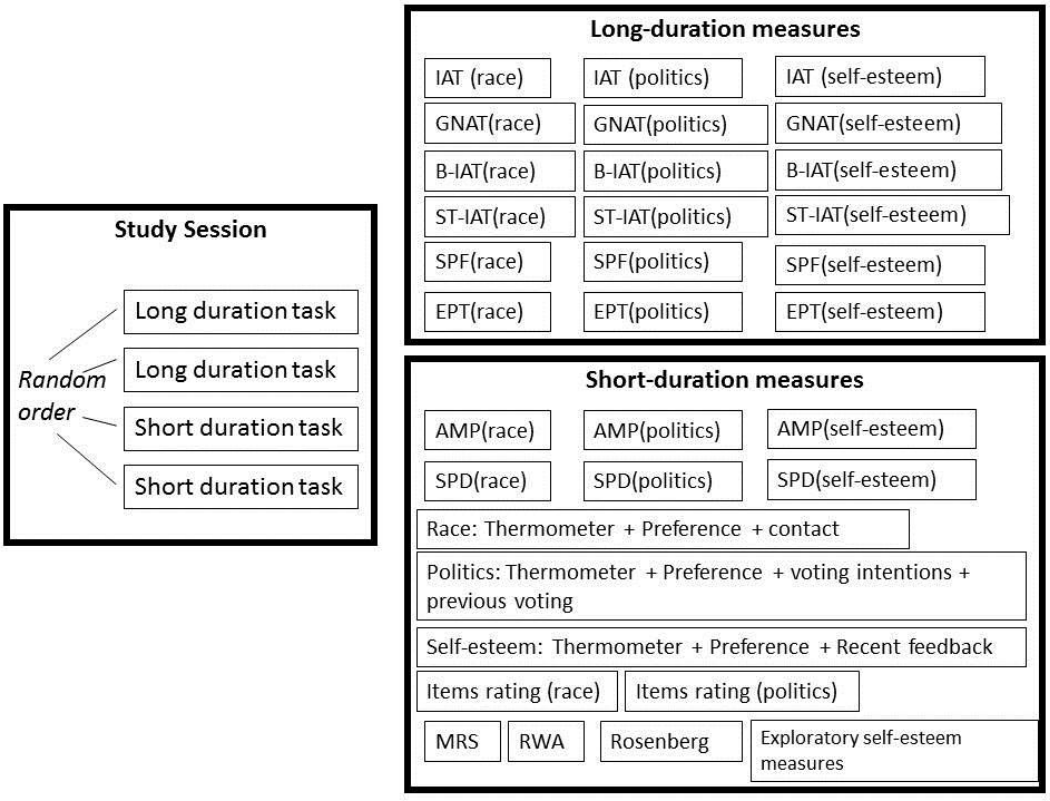


Figure 1. Illustration of the study procedure. Each study session included 2 long and short tasks, selected and ordered randomly. The tasks are listed on the right. Measures that share a rectangle were presented together in the same questionnaire.

Results

Data processing

Given the large samples, the emphasis in this report is on effect size rather than significance testing. Positive scores in the comparative preference measures represented preference for White people over Black people, Democrats over Republicans, or the self over others depending on the task content.

The data produced by the implicit measures can be processed in various methods, especially those that rely on response latency and error-rates. Past research tested several scoring algorithms of the IAT (Greenwald, et al., 2003), and the logic of the *D*-measure appears to have general application for procedures that compare contrasted conditions, especially with response latency as a dependent variable (Sriram, et al., 2010; Sriram, Nosek, & Greenwald, 2012). There have not yet been similar systematic investigations for scoring the other measures. Therefore, we used *D* for measures that—like the IAT—were based on response latency, did not use a response deadline, and required correction of error responses (the BIAT, ST-IAT and the SPF). For other implicit measures (the EPT, GNAT and the AMP), we tested a few scoring algorithms and chose the one that produced the best psychometric qualities (internal consistency and relationship with other measures) with the present data to maximize the performance of each measure for this comparative analysis[2]. As detailed below, all the latency-based measures used a variation of *D* – standardization of an average latency difference score by dividing it by the standard deviation of the participant's response latencies across both critical performance conditions.

IAT. The scoring followed previous recommendations (Greenwald, et al., 2003). Trials slower than 10000ms or faster than 400ms were removed. The trial response latency was the latency from stimulus onset until the correct response regardless of whether there was an error. Participants with more than 10% trials faster than 300ms were removed from the IAT analyses. For each participant, the IAT *D* score was the average of *D* scores calculated separately on the two IAT halves: one half was the difference between the average response latency of Blocks 3 and 6 divided by the standard deviation of all the trials in these two blocks. The other half was the same calculations using Blocks 4 and 7. The result, termed the IAT *D*, has the theoretical ranges of -2 and 2, with 0 reflecting no difference in average response latency between the two pairing conditions.

BIAT. The only changes compared to the IAT scoring were: the first four trials of each block were removed from the analyses (following recommendation of Sriram & Greenwald, 2009), and the *D* was the average of the *D* scores of the two halves Blocks 2-5, and Blocks 6-9.

GNAT. When the response deadline in the GNAT is very short, the score is based on error-rates, and is computed using a signal detection analysis (Nosek & Banaji, 2001). However, because we used a relatively long response deadline, there were low error rates (less than 10% in all the three GNAT measures), making latency-based scoring more appropriate. Nevertheless, when we compared various alternative scores using the data of the present study, the best score was an average of the standardized error-based and latency-based scores. The error-based score was the difference score between the error-rates in the two pairing conditions. The latency-based score was very similar to the IAT *D* score with the following changes: only correct "Go" trials were included, and, like the BIAT, separate scores were calculated and averaged for the first four blocks and the last four blocks. Before averaging the standardized error-based and latency-based scores we rescaled the standardized scores such that a zero score would reflect no preference (e.g, the z-score of the zero score in the error-based race score was 0.75. Therefore, we subtracted 0.75 from all the error-based scores before averaging them with the latency-based scores).

ST-IAT. We computed an ST-IAT *D* score for each attitude object. For each ST-IAT (i.e., for each attitude object), two scores were calculated and averaged, one using trials 1 to 24 of each block, and the other using trials 25 to 48. The preference score was the difference between the two single-category ST-IAT *D* scores.

SPF. The trial exclusion and participant exclusion rules were identical to the IAT rules. The scoring was a modification of the IAT scoring (as recommended by Bar-Anan et al., 2009). In the SPF, there are four different trial conditions (e.g., *Democrats-Good words*, *Democrats-Bad words*, *Republicans-Good words* and *Republicans-Bad words*). A *D* score for each of the four trial conditions was computed within each block. The *D* score was the difference between the average latency of the trial condition and the average latency in the block, divided by the overall standard deviation of that block. Then, single-attitude object scores were computed as the difference between the good and bad *D* scores of each category (e.g., the difference between the *D* scores of *Democrats-Good words* and *Democrats-Bad words*). A preference score for each block was computed as the difference between the two single-category scores. The final SPF preference score was the average of the preference scores of the three blocks.

EPT. The scoring algorithm was selected after a comparison of a number of possible algorithms that were used in other published studies, and also a few modifications and combinations of those algorithms. The eventual algorithm that outperformed the other algorithms is a novel one, although it was only slightly better than the more common algorithms. We excluded trials with incorrect responses and trials with response latency that was more than 2 standard deviations away from the average response latency in the trial's condition (each trial condition was a prime-target combination; e.g., *Democrats-Bad*). We also excluded EPT sessions with more than 40% incorrect responses.

For each trial condition within each block, we computed the average of the log transformed response latencies. We then computed a single-category *D* score as the difference between those averages (e.g., *Democrats-Bad* minus *Democrats-Good*) divided by the overall standard deviation. For each block, we computed a preference score as the difference between the two single-category scores. The EPT preference score was the average preference score across the three blocks.

AMP. Following previous usage of the AMP, we did not exclude any trials (Payne et al., 2005; Payne et al., 2008). In some of the studies that used the AMP, all the participants were included (e.g., Payne et al., 2010), whereas in others, participants who always used the same response were removed (e.g., Payne et al., 2005). We compared four different response-bias cutoffs (95%, 99%, 100% of trials with the same response and no-exclusion), and used the one that produced the best psychometric qualities (a 95% cutoff). The single-category score was computed as the difference between the rate of pleasant responses after primes of that category and the rate of pleasant responses after the neutral primes. The preference score was the difference between the two single-category scores (Payne et al., 2005, Experiment 6).

Other measures. The rest of the scoring was straightforward. In the SR and items rating, the single-category score was the average rating of the category items. The preference score was the difference between the single-category scores. The scale scoring was the average rating of each item (reversed when needed).

Table 2 presents the summary of the main results of the performance of the seven implicit measures on the criteria tested in this study.

Table 2
A Summary of the Results

			Reliability				**Correlations**		**Outliers exclusion**		
			Internal Consistency		Test-retest		With implicit measures	With direct measures	% shared variance lost after dropping outliers (2 standard deviations)		
Overall			Average alpha		Average correlation		Average correlation	Average correlation	Internal consistency	Corr. with implicit measures	Corr. direct measures
IAT			**.88**		.45		.39	.35	*5.5*	2.7	2.2
BIAT			.83		**.63**		**.41**	**.38**	7.7	2.4	2.7
GNAT			.74		.42		.40	.33	10.0	1.3	2.0
ST-IAT			.77		.48		.36	.31	9.9	**2.8**	2.3
SPF			*.53*		.46		.31	.27	12.5	2.2	*0.8*
EPT			*.57*		*.33*		*.25*	*.24*	16.0	2.1	1.1
AMP			.69		.50		.26	.32	**29.2**	2.7	**4.7**
Race	Mean Effect-size	Known groups effect		95% CI	Correlation	95% CI					
IAT	0.75	**1.12**	**.86**[a]	.85-.87	.40[bc]	.22-.55	**.36**	.27	*6.7*	**3.5**	*1.2*
BIAT	0.73	0.77	.81[b]	.80-.82	**.63[a]**	.50-.73	.34	.27	7.9	2.9	*1.2*
GNAT	**0.83**	0.67	.70[d]	.69-.73	.30[bc]	.13-.45	.35	.27	11.8	1.1	2.3
ST-IAT	0.20	*0.33*	.74[c]	.72-.76	.34[bc]	.16-.49	.30	.24	11.2	2.2	2.2
SPF	0.24	0.76	*.52[f]*	.49-.55	.38[bc]	.21-.53	.24	.24	11.8	1.4	1.8
EPT	0.07	0.57	.54[f]	.51-.57	*.18[c]*	.00-.36	*.20*	*.19*	15.4	1.1	1.5
AMP	*-0.23*	0.39	.66[e]	.64-.68	.33[bc]	.18-.47	*.21*	**.31**	**34.0**	1.5	**7.4**
Politics											
IAT	0.49	**1.49**	**.93[a]**	.92-.93	.65[abc]	.54-.74	.58	.60	*1.9*	3.8	3.7
BIAT	**0.63**	1.40	.89[b]	.88-.90	**.78[a]**	.69-.85	**.60**	**.63**	5.3	3.3	4.2
GNAT	0.47	1.38	.84[c]	.83-.85	.72[ab]	.63-.79	.59	.59	4.9	2.0	3.4
ST-IAT	0.42	1.04	.84[c]	.83-.85	.54[c]	.40-.66	.55	.56	8.1	5.1	5.8
SPF	*0.21*	1.08	*.59[f]*	.57-.61	.58[bc]	.44-.70	.52	.48	13.7	4.0	*2.4*
EPT	0.27	*0.73*	.63[e]	.61-.65	*.51[c]*	.34-.63	.45	*.42*	18.5	5.0	4.0
AMP	0.31	0.88	.81[d]	.80-.82	.73[a]	.65-.79	*.43*	.48	**27.2**	**5.7**	**10.2**
Self											
IAT	**1.31**		.82[a]	.81-.83	*.26[a]*	.09-.41	.21	.14	*7.9*	0.8	0.8
BIAT	1.03		.76[b]	.74-.78	**.42[a]**	.25-.57	**.25**	**.18**	1.2	0.9	1.2
GNAT	1.23		.65[d]	.63-.67	.34[a]	.14-.51	.21	.08	13.3	0.9	*-0.1*
ST-IAT	0.57		.70[c]	.68-.72	.36[a]	.19-.51	.20	.09	1.6	1.0	0.4
SPF	0.96		*.48[f]*	.45-.51	.39[a]	.23-.53	.14	*.06*	12.2	**1.1**	0.1
EPT	0.43		.54[e]	.51-.56	*.29[a]*	.36-.43	*.07*	.08	14.0	0.1	0.1
AMP	*0.16*		.55[e]	.53-.57	.35[a]	.20-.48	.10	.16	**26.6**	0.9	**1.7**

Notes. The effect sizes were computed from the preference for White people, Democrats and the Self. Mean correlations and internal consistencies were averaged after applying Fisher's transformation and then transformed back to correlations; **Bold** font = best performance in the relevant criterion; *Italic* font = worst performance in the relevant criterion; The sample sizes for the test-retest analyses were between 83 and 158 (average = 116); The Cronbach alphas were compared using Feldt test (1969); The test-retest values are correlations between the first and the second tests.

Mean Effects and Known-Group Differences

All else being equal, better measures will be more sensitive to known effects and to detecting known group differences. We first examined mean effects and known-groups differences for each of the topics separately, and then considered the overall task results in the aggregate.

Racial attitudes. All the implicit measures except the AMP (d = -0.26) indicated an implicit preference for White people compared to Black people (see Table 2 for effect sizes, and Table A1 in the appendix for more details). Effects for the GNAT (d = 0.83), IAT, and IAT were substantially stronger than for ST-IAT, SPF, and EPT (d = 0.07). As we discuss later in more details, it is possible that the AMP showed a preference for Black over White people because it measures attitudes toward the items more than toward their social groups. Indeed, although participants self-reported preference for *White people* over *Black people* in the preference (d = 0.39) and the thermometer self-report measures (d = 0.31), they showed a preference for the Black people stimuli over the White people stimuli in the items ratings (d = -0.79) and the speeded self-report (d = -0.14). Therefore, it is possible that with race attitudes, the effect-size criterion did not only reflect the sensitivity of the measures to attitudes, but also reflected the sensitivity of the measures to the social groups and insensitivity to the specific race stimuli.

We also tested the extent to which the implicit measures were sensitive to the difference between racial evaluations of White and Black participants. Theory and evidence shows that White people tend to show the strongest preference for White over Black and Black people tend to show the strongest preference in the opposite direction (e.g., Fazio et al., 1995; Nosek, Smyth, et al., 2007). Table 2 summarizes the comparison of Black and White participants for all racial attitude measures (See Table A2 in the appendix for more details). The IAT, BIAT, and SPF showed highest sensitivity to the participant's social group (Cohen's d = 1.12, 0.77, and 0.76, respectively). The GNAT and the EPT came next (Cohen's d = 0.67 and 0.57, respectively). The least sensitive to detect known group differences were the AMP and the ST-IAT, with effects about half the size as the strongest ones (Cohen's d = 0.39 and 0.33, respectively).

Political attitudes. All the direct (mean d = 0.54) and implicit (mean d = 0.33) measures indicated a preference for Democrats over Republicans. This is not surprising as the sample was more liberal than conservative on average. The tasks did vary in the strength of this effect, but because the known-groups difference between liberals and conservatives should be so strong, the more direct test of sensitivity to political attitudes is indicated by the main effects separately for liberals and conservatives, and the magnitudes of the difference between these groups. As presented in Table 2, the scores of the IAT, BIAT and GNAT were the most sensitive to participant's political identity (ds = 1.49, 1.40 and 1.38, respectively). The SPF and ST-IAT were next on that criterion more than 20% weaker (ds = 1.08, 1.04, respectively), and the AMP and the EPT were the least sensitive about 35 to 50% weaker than the strongest measures (ds = 0.88, 0.73, respectively).

Self-esteem. All the direct (mean d = 0.41) and implicit (mean d = 0.64) measures indicated a preference for the self over others. The magnitude of the effect varied substantially in the following order from strongest to weakest: IAT, BIAT, GNAT, SPF, ST-IAT, EPT, and AMP (Table 2). The first four were similarly large effect sizes; ST-IAT and EPT elicited moderate effect sizes that were more than 40% and 50% smaller than the first four respectively. The AMP elicited a weak preference for self over others that was more than 80% smaller than the first four.

Comparative conclusions on mean effects and known-group differences. In summary, the IAT and the BIAT showed the best sensitivity to detect the expected preference and the

expected effect of participants' social identity. The GNAT also showed good sensitivity – once better than the IAT and the BIAT, and once better than the BIAT. These three measures were usually followed by the ST-IAT and the SPF. The AMP and the EPT usually showed the weakest sensitivity.

One of the originating rationales of developing indirect, or implicit, measures was that people may be reluctant to self-report preferences for some groups, particularly racial groups, but that these effects could be measured if the attitude content was more subtly assessed or elicited automatically. In that context, it is notable that the two self-report measures of preferences for White people compared to Black people showed stronger effect sizes favoring Whites than four of the implicit measures – AMP, ST-IAT, SPF and EPT.

However, the pattern as a whole illustrates a potential inferential challenge of using main effects as a criterion in isolation. If, for example, AMP, ST-IAT, SPF, and EPT are more influenced by stimulus characteristics than category evaluations, then the fact that the individual Black stimuli were evaluated relatively more favorably explicitly than the category "Black people" might have produced the differences in main effect. This can be tested, in part, by looking at the correlations between implicit measures and the self-reported category and individual evaluations. If the difference in effect magnitude is attributable to the difference in influence of category versus items, then the correlations between measure and item should be stronger with category ratings for high effect size measures and with item ratings for weak effect size measures. The pattern of correlations between implicit measures and the direct measures in Table 4 is somewhat supportive of this possibility. For both EPT and AMP, the correlation is slightly stronger with the aggregate preference rating of individual stimulus items than with the group preference and thermometer ratings. This holds for both race and politics. This pattern is less consistent for ST-IAT and SPF. And, the pattern is partially reversed – stronger correlations with group preference and thermometer than with aggregated individual items – for the IAT, BIAT, and GNAT, but inconsistently so. However, for politics, the IAT, BIAT, and GNAT actually had the strongest relations with the aggregated individual items. Therefore, the larger effect for the IAT, BIAT and the GNAT in measuring politics probably suggests better sensitivity for these measures in assessing political attitudes, in comparison to the other measures. In sum, the pattern of relations suggests that (a) some measures – particularly EPT and AMP – are more sensitive to individual stimuli than to the concept categories, consistent with prior demonstrations for the EPT in comparison to the IAT (Olson & Fazio, 2003; Payne et al., 2008), and (b) differences in effect magnitudes *also* indicate differences in sensitivity to the construct, with the IAT and the BIAT as the most sensitive measures, and the GNAT not far behind. The latter point will be affirmed more clearly following accumulation of evidence across the subsequent evaluation criteria.

Reliability

All else being equal between measures, higher internal consistency is considered more desirable than lower internal consistency (John & Benet-Martinez, 2000), particularly for maximizing the power to detect relations with other measures. We computed Cronbach's alpha (Cronbach, 1951) from three data parcels for each measure as our assessment of internal consistency. The first parcel included the first trial of each triplet of consecutive trials, and the third parcel included the third trial of each triplet for each response block. For tasks requiring calculation of scores across response blocks or trials, those scores were computed separately with each parcel of data.

The average internal consistencies are presented in Table 2. Almost all of the 95% confidence intervals were non-overlapping. The IAT was the most internally consistent measure,

the second best measure was always the BIAT, and the GNAT and the ST-IAT shared the third and fourth places. For each of the three topics, the AMP, EPT, and SPF were consistently the 5th, 6th, and 7th respectively. Comparatively, SPF and EPT were notably less reliable than the others, and the IAT did particularly well. Squaring the reliability correlations gives an estimate of the shared variance of a measure with itself to illustrate the size of the reliability gap. IAT and BIAT had R-squared values of 77% and 69% for average internal consistency whereas EPT and SPF 32% and 28%, less than 1/2 the magnitude.

Test-retest reliability. Participants were not assigned to the same measure twice in the same study session. However, participants who completed more than one session could be assigned to the same measure again (~100 participants per measure). Only about 10% of the retests were completed more than 24 hours after the time of the first test, and about 50% of the retests were completed less than an hour after the first test. Therefore, the test-retest correlation is not so different from internal consistency of the measures rather than their stability over time. Table 2 presents test-retest correlations for each topic and averaged across topics. The BIAT showed the strongest test-retest reliability, and all of the other measures clustered closely together behind the BIAT, except for EPT which has the weakest test-retest reliability. We caution that with just 100 participants per test-retest for each topic (300 per measure combined across topics), these averages and rankings have relatively wide standard errors compared to other estimates.

Comparative conclusions for reliability. The reliability results suggest that IAT and BIAT have the strongest reliability. And, SPF and especially EPT, have the weakest. However, reliability is not redundant with validity. An unreliable measure that is particularly effective or well suited for measuring a particular construct may be less efficient in its assessment (i.e., requiring more participants to detect effects), but not necessarily less valid for drawing inferences. On the flip side, a measure could be perfectly reliable, but not assess the construct of interest.

Additional convergent evidence is accumulated if correlations with criterion measures – other indirect measures and direct measures of the same construct – follow a similar rank-order as the differences in internal consistency across measures. This would increase confidence that the differences in internal consistency reflect variation in the quality of psychometric performance of the measure, rather than being an artifact of extraneous influences. That evidence is provided in the next sections.

Relationships with Other Implicit Measures

Assuming that the implicit attitude measures are valid to some degree, all else being equal, more valid measures will be more strongly related to other implicit measures than less valid measures will be. Of course, this general statement is qualified by the possibilities that (a) subsets of implicit measures assess different components of implicit cognition constructs – each valid but distinct, (b) subsets of implicit measures share a confounding influence that create spuriously strong relations that are not relevant to the construct, and (c) the measures do not *en masse* have shared confounding influence that spuriously inflate their covariation. Concern (a) can be addressed by examining the possibility of distinct covariation with other criterion measures, as we pursue in the next section. Concerns (b) and (c) can be addressed by examining whether there are clusters of strong covariation and, particularly because the AMP is not a response latency based procedure, it can address issues related to (b) and minimize the likelihood of (c). An in-depth examination of the structural relations among implicit measures goes beyond the scope of this article, but is taking up in detail with these data by Bar-Anan, Spies, and Nosek (2012).

Table 3
Correlations among the Implicit Measures

	IAT	BIAT	GNAT	ST-IAT	SPF	EPT	AMP
Avg. N	395	374	365	387	254	393	509
Range N	294-558	304-496	278-520	278-548	313-524	298-539	414-558
Average							
IAT		.51	.50	.41	**.38**	.24	**.30**
BIAT	**.51**		**.53**	.46	**.38**	**.29**	.28
GNAT	.50	**.53**		**.48**	.33	*.27*	.28
ST-IAT	.41	.46	.48		.31	*.23*	.26
SPF	.38	.38	.33	.31		.25	*.19*
EPT	*.24*	.29	*.27*	*.23*	.25		.23
AMP	.30	*.28*	.28	.26	*.19*	*.23*	
Race							
IAT		$.49^{a}$	$.48^{a}$	$.33^{bc}$	.28	$.29^{a}$	$.27^{a}$
BIAT	$.49^{a}$		$.42^{a}$	$.40^{ab}$	.28	$.22^{ab}$	$.23^{ab}$
GNAT	$.48^{a}$	$.42^{a}$		$.47^{a}$	.25	$.19^{ab}$	$.24^{ab}$
ST-IAT	$.33^{b}$	$.40^{ab}$	$.47^{a}$		.24	$.15^{b}$	$.16^{b}$
SPF	$.28^{b}$	$.28^{bc}$	$.25^{a}$	$.24^{cd}$		$.20^{ab}$	$.21^{ab}$
EPT	$.29^{b}$	$.22^{c}$	$.19^{b}$	$.15^{d}$	.20		$.16^{b}$
AMP	$.27^{b}$	$.23^{c}$	$.24^{ab}$	$.16^{d}$	.21	$.16^{b}$	
Politics							
IAT		$.65^{a}$	$.70^{a}$	$.62^{a}$	$.60^{a}$	.41	.45
BIAT	$.65^{ab}$		$.70^{a}$	$.63^{a}$	$.63^{a}$	.52	.43
GNAT	$.70^{a}$	$.70^{a}$		$.65^{a}$	$.53^{ab}$	.47	.43
ST-IAT	$.62^{ab}$	$.63^{ab}$	$.65^{a}$		$.48^{bc}$	.42	.47
SPF	$.60^{b}$	$.63^{ab}$	$.53^{ab}$	$.48^{b}$		.45	.38
EPT	$.41^{c}$	$.52^{b}$	$.47^{bc}$	$.42^{b}$	$.45^{bc}$		.43
AMP	$.45^{c}$	$.43^{c}$	$.43^{c}$	$.47^{b}$	$.38^{c}$	.43	
Self							
IAT		$.37^{a}$	$.29^{ab}$	$.21^{ab}$	$.22^{a}$	.00	$.14^{a}$
BIAT	$.37^{a}$		$.36^{a}$	$.32^{a}$	$.18^{ab}$	.10	$.16^{a}$
GNAT	$.29^{ab}$	$.36^{a}$		$.24^{ab}$	$.18^{ab}$	.03	$.16^{a}$
ST-IAT	$.21^{bc}$	$.32^{a}$	$.24^{ab}$		$.18^{ab}$	.11	$.12^{a}$
SPF	$.22^{bc}$	$.18^{b}$	$.18^{bc}$	$.18^{ab}$		.10	$-.03^{b}$
EPT	$.00^{d}$	$.10^{b}$	$.03^{c}$	$.11^{b}$	$.10^{ab}$		$.06^{a}$
AMP	$.14^{c}$	$.16^{b}$	$.16^{bc}$	$.12^{b}$	$-.03^{b}$	.06	

Notes. The average correlation was calculated after applying Fisher's transformation, and then was transformed back to a correlation coefficient; In the overall section: **Bold** = the strongest correlation of that column; *Italics* = the weakest correlation of that column; In the by-topic sections: in each column, correlations that do not share superscript are significantly different than each other;

The average correlation of each measure with the other indirect measures is presented in Table 2, and the correlation matrices are presented in Table 3. Average correlations might be skewed by extreme individual correlations. Therefore, for each measure we rank-ordered the correlations of the other measures with it, and then we averaged those ranks for each of the

measures to detect cases in which the average did not reflect the frequent quality of the measure. The average rankings were very consistent with the average correlation results suggesting that there were no inordinately influential correlations (see Table A3 in the appendix).

The BIAT was the most related to the rest of the indirect measures (average r = .41). The other measures, from highest to lowest: GNAT (mean r = .40), IAT (mean r = .39), ST-IAT (mean r = .36), SPF (mean r = .31), AMP (mean r = .26), and EPT (mean r = .25). The worst measures on this criterion, AMP and EPT, might be considered distinct from the other indirect measures because of procedural features, such as not requiring categorization of the primes. If that is the case, there might be two clusters of implicit measures – IAT and ostensibly-related derivatives – and the AMP and the EPT. If so, then the AMP and EPT would be related to each other more strongly than to the other measures. This was not the case. The average AMP-EPT relation was .23, which was weaker than the relation of the AMP with four other measures (IAT = .30, BIAT = .28, GNAT = .28, ST-IAT = .26) and weaker than the relation of EPT with four other measures (BIAT = .29, GNAT = .27, SPF = .25, and IAT = .24).

Comparative conclusions for relations among indirect measures. The AMP and EPT are more distinct from other implicit measures, including each other, than are the rest of the measures. The most likely explanation for this pattern of relations, coupled with the similar rank ordering for internal consistency, is that AMP and EPT are both relatively distinct, and *also* less effective in reliably assessing the target evaluation than are the other measures. However, it could still be the case that both measures assess unique components of evaluation that are not assessed by the other implicit measures (including each other). The SPF similarly did not perform particularly well in the combination of internal consistency and relation with other indirect measures. The next section examines a third feature – relations with direct measures and criterion variables – to provide converging evidence for understanding the comparative qualities of the implicit measures.

Correlations with Direct Measures and Other Criterion Variables

Table 4 presents the average relationship of each implicit measure with the direct measures of the same topic and the other criterion variables. For each criterion measure, the correlations of the implicit measures were compared statistically, and also ranked. The aggregated correlations are presented in Table 2 (see Table A3 for the average rankings).

Across topics, the ranking for correlations with direct measures was mostly similar to the other evaluation criteria: BIAT, IAT, GNAT, AMP, ST-IAT, SPF, and EPT showing the weakest relations. Like reliability, squaring the correlation provides an estimate of shared variance between the implicit measure and the direct measures. Illustrating the magnitude of the differences, the IAT and BIAT had, on average, more than double the variance relation with the direct measures than did the SPF and EPT.

The main difference between the performance of the implicit measures in this criterion in comparison to the previous criterions is that AMP showed the strongest average correlation with direct measures for racial attitudes, and the second-strongest average correlation with direct measures for self-esteem. Bar-Anan and Nosek (in press) reported evidence that the AMP's positive psychometric qualities are a function of a subset of participants who report that they intentionally rated the primes and not the targets. If participants are accurate with this self-perception, it would explain why the AMP performed somewhat better in comparative correlations with direct measures than it did with implicit measures and internal consistency (but see Cameron et al., in press, for evidence against the possibility that the AMP measures direct evaluation). The possibility that a minority of the participants influence the AMP's good

relations with direct measures is examined in more detail in the next section examining the influence of outliers.

Table 4

Correlations of the Implicit Measures with Direct Attitude Measures and other Criterion Variables

Race	Preference	Thermometer	Items	MRS	Contact with Black people	Speeded Report
Effect-size	0.39	0.31	-0.79	--	--	-0.14
Range N	480-630	494-653	464-684	480-671	492-647	421-569
IAT	.32^{ab}	.32^{a}	.21^{b}	.29^{ab}	-.14^{ab}	.31^{ab}
BIAT	.29^{ab}	.32^{a}	.27^{b}	.29^{a}	-.13^{ab}	.28^{ab}
GNAT	.31^{ab}	.25^{ab}	*.20*^{b}	**.32**^{a}	-.18^{ab}	**.37**^{a}
ST-IAT	.23^{bc}	.28^{a}	.27^{b}	.24^{ab}	-.11^{ab}	.29^{ab}
SPF	.28^{ab}	.28^{a}	.21^{b}	*.18*^{b}	-**.20**^{a}	.27^{ab}
EPT	*.15*^{c}	*.17*^{b}	.26^{b}	.22^{ab}	-*.09*^{b}	*.24*^{b}
AMP	**.35**^{a}	**.33**^{a}	**.41**^{a}	.29^{ab}	-.13^{ab}	.33^{ab}

Politics	Preference	Thermometer	Items	RWA	Voted	Voting intentions	Speeded Report
Effect	0.62	0.60	0.46	--	--	--	0.47
Avg. N	554	561	431	559	284	523	547
Range N	468-600	516-607	396-453	472-593	234-316	444-572	459-589
IAT	.64^{ab}	.60^{a}	**.69**^{a}	-.43^{bc}	**.66**^{a}	.51^{a}	**.64**^{a}
BIAT	**.66**^{a}	**.65**^{a}	**.69**^{a}	-**.57**^{a}	.65^{a}	**.54**^{a}	.61^{a}
GNAT	.65^{ab}	.60^{a}	**.69**^{a}	-.49^{ab}	.65^{a}	.46^{abc}	.58^{a}
ST-IAT	.58^{b}	.59^{ab}	.56^{bc}	-.44^{bc}	.64^{a}	.49^{ab}	.61^{a}
SPF	.48^{c}	.46^{c}	.58^{b}	-.39^{cd}	.51^{b}	.41^{bc}	*.47*^{c}
EPT	*.43*^{c}	.43^{c}	*.46*^{c}	-*.33*^{d}	*.43*^{b}	*.36*^{c}	.49^{bc}
AMP	.48^{c}	.51^{bc}	.59^{b}	-.36^{cd}	.49^{b}	.35^{c}	.52^{bc}

Self	Preference	Thermometer	Rosenberg	Speeded Report
Effect	0.44	0.37	--	0.44
Avg. N	601	604	591	534
Range N	459-691	462-693	494-667	450-579
IAT	.11^{ab}	.13	.17^{a}	.14^{ab}
BIAT	.14^{a}	.16	**.18**^{a}	**.24**^{a}
GNAT	*.00*^{b}	.11	.06^{abc}	.13^{ab}
ST-IAT	.05^{ab}	.12	.11^{ab}	*.07*^{b}
SPF	.10^{ab}	*.06*	.00^{bc}	.09^{b}
EPT	.13^{a}	*.06*	-*.04*^{c}	.16^{ab}
AMP	**.16**^{a}	**.17**	.07^{abc}	**.24**^{a}

Notes. In each column of each topic section, correlations that do not share superscript are significantly different from each other; The thermometer and the items columns refer to a difference score.

Another exception in this criterion in comparison to previous criteria was that the SPF score was the most related to contact with Black people. However, the SPF's better performance on self-reported contact was not substantially larger than some other measures suggesting that the aggregate trend may be a more valid indicator of the SPF's performance.

Comparative conclusions for relations with direct measures and other criterion variables. The comparative relations with direct measures were largely consistent with the variation in internal consistency and comparative relations with indirect measures. These

mutually reinforcing findings provide greater confidence that each finding can be attributed to the psychometric performance of the measure in assessing its target construct. In particular, the IAT, BIAT, GNAT, and ST-IAT performed consistently better than the SPF and EPT across topics and across the evaluation criteria. The AMP's performance was more similar to the SPF and the EPT in most criteria, with the exception of relationship with direct measures, in which the AMP switched places with the ST-IAT.

That said, there may be particular contexts and criterion variables in which the SPF, EPT, and AMP perform particularly well. Such evidence would facilitate the identification and clarification of distinct components of evaluation that are assessed by the different measures. At present, there is no evidence beyond what is presented here to suggest conditions under which one of these measures is going to be particularly advantaged in predicting criterion variables, so the possibility that each measure will have a domain of superior performance is entirely speculative.

Influence of Outliers

Outliers are of concern for interpretation of relations between measures because a small subset of observations has an inordinate influence on the observed results. If the psychometric qualities of a measure rely heavily on outliers, it may be considered a less useful measure because the overall effects and relations with other variables are heavily influenced by an extreme subset of the sample. Therefore, better measures will have psychometric qualities that are comparatively insensitive to removal of outlier scores.

Further, because participants were randomly assigned to tasks, we can assume that individuals that are outliers because of having extreme attitudes (a construct-relevant reason to be an outlier) are evenly distributed across the measures. As such, if there are differences in the number of outliers and in the impact of their removal, it suggests that the measurement procedures themselves are differentially likely to produce influential outliers.

To examine the influence of outlier scores, for each measure we removed scores that were in excess of 2 standard deviations away from the mean of the measure. As a consequence, outlier scores were around 5% of the sample in each measure (see Table A4 in the appendix). The only exception was the AMP-politics measure, in which the outliers were 9% of the sample. This indicates that the AMP-politics measure violated normal distribution assumptions. In this case, the AMP-politics measure was particularly likely to elicit very extreme evaluations preferring one party over the other. Because other measures did not show that non-normal distribution, it is likely that the pattern is due to the AMP, not the assessed attitude. One possible explanation is that unlike the latency-based measures, the AMP has a limited range of possible scores. This increases the likelihood that a sample of participants would produce a skewed distribution.

Another possible explanation for the AMP's skewed distribution is if some participants ignored the instructions and directly evaluated the politician primes instead of evaluating the Chinese pictographs. If a minority of participants who completed the AMP politics measure rated the primes directly, those with polarized explicit preferences could produce extreme scores and have strong influence on the overall mean, internal consistency and the correlation with other measures. As mentioned earlier, there is evidence elsewhere that participants who report intentionally evaluating primes in the AMP indeed have a large influence on the overall mean, internal consistency and the correlation with other measures (Bar-Anan & Nosek, in press).

An important piece of evidence for this possibility, and more general performance of the measures, is the impact of removing the outliers on internal consistency and correlations with other measures. Table 5 presents the influence of removing the outliers on the psychometric

qualities of each of the seven indirect measures. The table also shows the average percentage of variance decline in internal consistency and relations with measures.

Table 5

The Influence of Excluding Outlier Scores on the Psychometric Qualities of the Measure

	Internal consistency (alpha Cronbach)					**Average correlation with implicit measures**					**Average correlation with direct measures**				
		Exclusion method					Exclusion method					Exclusion method			
	All cases	By STD		By percent		All cases	By STD		By percent		All cases	By STD		By percent	
	Alpha	Alpha	% loss	Alpha	% loss	R	R	% loss	R	% loss	R	R	% loss	R	% loss
Overall															
IAT	**.88**	**.85**	*5.2*	**.81**	*11.9*	.39	.35	2.7	**.34**	3.8	.35	.32	2.2	.30	3.0
BIAT	.83	.78	8.0	.73	15.6	**.41**	.37	2.4	**.34**	4.4	**.38**	**.34**	2.7	**.34**	2.8
GNAT	.77	.67	9.9	.59	19.5	.40	**.38**	1.3	**.34**	3.5	.33	.31	2.0	.29	3.2
ST-IAT	.74	.70	8.8	.62	19.5	.36	.31	**2.8**	.26	**5.4**	.31	.26	2.4	.23	4.7
SPF	*.53*	*.39*	12.9	.26	21.3	.31	.27	2.2	.23	3.9	.27	.24	*0.8*	.22	2.8
EPT	.57	.41	16.0	*.25*	25.0	*.25*	.20	2.1	.18	*3.4*	*.23*	*.20*	1.1	*.18*	*2.4*
AMP	.69	.39	**32.4**	.21	**35.6**	.26	*.19*	2.7	*.16*	3.8	.32	*.20*	**5.0**	*.18*	**7.8**
Race															
IAT	**.86**	**.82**	*6.7*	**.78**	*13.5*	**.36**	.30	**3.5**	.29	4.5	.27	.24	*1.2*	.22	2.4
BIAT	.81	.76	7.9	.70	17.3	.34	.29	2.9	.26	**4.8**	.27	**.25**	*1.2*	**.24**	*1.6*
GNAT	.71	.61	13.2	.53	21.6	.35	**.33**	1.1	**.30**	2.8	.27	.23	2.2	.22	2.6
ST-IAT	.74	.66	11.2	.58	21.3	.30	.25	2.2	.21	3.6	.24	.18	2.3	.15	3.2
SPF	*.52*	.39	11.8	.26	23.0	.24	.21	1.4	.17	2.6	.24	.20	1.8	.18	2.6
EPT	.54	*.37*	15.0	.21	24.6	*.20*	*.16*	1.1	.15	*1.6*	*.19*	.15	1.5	*.13*	2.0
AMP	.66	.31	**34.0**	*.10*	**42.7**	*.21*	.17	1.5	*.14*	2.2	**.31**	*.12*	**8.4**	.13	**8.2**
Politics															
IAT	**.93**	**.92**	*1.9*	**.90**	*6.0*	.58	.54	3.8	.52	*5.5*	.60	.57	3.7	.56	5.2
BIAT	.89	.86	5.3	.83	9.9	**.60**	**.57**	3.3	**.53**	6.5	**.63**	**.59**	4.3	**.58**	5.6
GNAT	.84	.81	4.9	.76	13.1	.59	**.57**	2	**.53**	*5.5*	.59	.56	3.4	.54	6.6
ST-IAT	.84	.79	6.5	.74	15.6	.55	.49	5.1	.42	**11.2**	.56	.51	5.7	.46	*1.5*
SPF	.59	*.46*	13.7	.32	24.5	.52	.47	4.0	.42	7.7	.48	.45	*2.4*	.42	5.6
EPT	.63	*.46*	18.5	*.34*	28.0	.45	.38	5.0	.33	8.2	*.42*	.37	3.9	.34	5.9
AMP	.81	.62	**27.2**	.56	**34.8**	*.43*	*.35*	**5.7**	*.31*	8.1	.48	*.35*	**10.3**	*.31*	**13.2**
Self															
IAT	**.82**	**.77**	*7.9*	**.72**	*16.1*	.21	.19	0.8	.17	1.3	.14	.10	0.8	.06	1.5
BIAT	.76	.69	1.2	.62	19.7	**.25**	**.23**	0.9	**.21**	1.8	**.18**	**.14**	1.2	**.14**	1.2
GNAT	.65	.55	13.3	.43	23.7	.21	.19	0.9	.16	**2.2**	.08	.09	*-0.1*	.06	0.4
ST-IAT	.65	.62	1.6	.52	21.6	.20	.18	1	.15	1.5	.09	.06	0.4	*.05*	0.5
SPF	*.48*	.33	12.2	.19	19	.14	.10	**1.1**	.08	1.4	*.06*	*.05*	0.1	*.05*	*0.1*
EPT	.54	.39	14.0	.26	22.5	*.07*	*.05*	0.1	.05	*0.3*	.08	.07	0.1	.08	*0.1*
AMP	.55	*.19*	**26.6**	*-.09*	**29.3**	.10	*.05*	0.9	*.03*	1.1	.16	.11	**1.6**	.09	**1.9**

Notes. The removal by standard deviation removed scores that were 2 standard deviations away from the mean; The removal by percentiles removed the 10% most extreme scores; The average % loss is the average loss of shared variance.

The IAT performed best in retaining internal consistency with removal of outliers, which lost just 5.5% variance on average, followed by BIAT, ST-IAT, GNAT, SPF, EPT and then

AMP, which lost 29.2% variance on average, more than 5 times as much as the IAT. For correlations with direct measures, the order was similar with much less loss overall, except that SPF and EPT lost considerably less than the others. However, those two implicit measures had weaker relations with direct measures in the first place, so had less to lose. For relations with other implicit measures, most measures showed a similar average loss of shared variance (around 2.5%), except the GNAT that lost about half of the variance compared to the other measures (1.3%). Only on this criterion the AMP usually did not lose more than the other measures, but this was probably only because it did not have good relations with implicit measures in the first place, so it had less to lose. After removing outliers, the AMP usually went from having the second weakest relations with other implicit measures to having the worst relations.

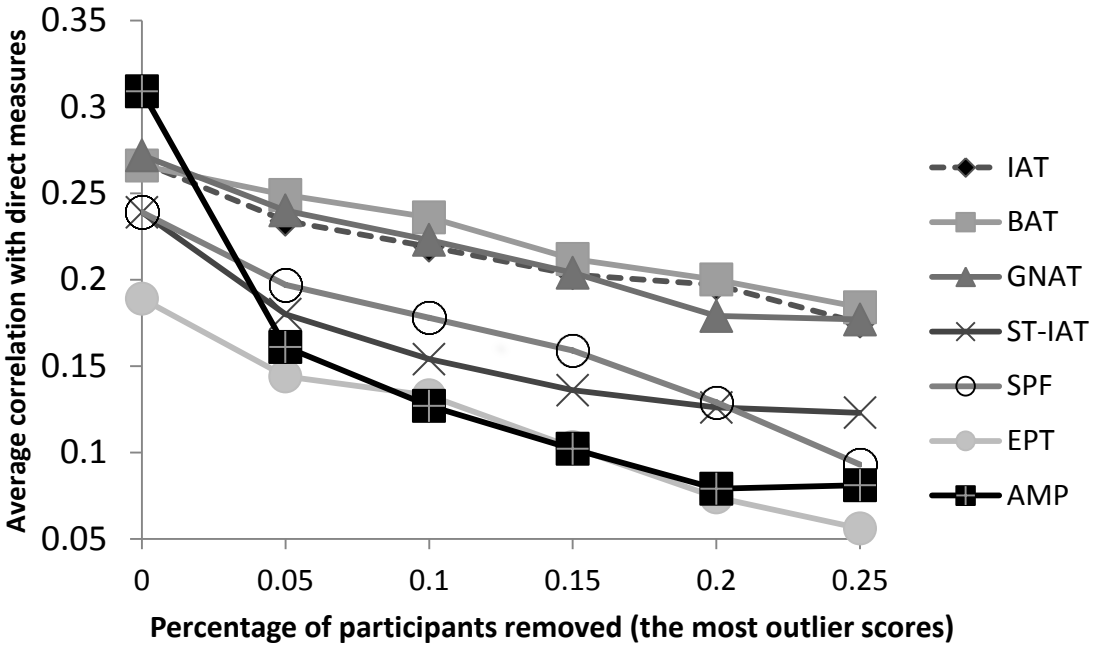
One reason for the marked reduction in the psychometric qualities of the AMP as a result of outlier removal might be the previously discussed fact that the AMP had sometimes more outlier scores than the other measures. Losing a larger portion of extreme scores is likely to cause a greater loss in the psychometric qualities. If, instead, all measures had the same percentage of scores removed, they might show similar declines. However, the AMP lost more scores than the other measures only when it measured politics (Table A4). Yet, it was the most sensitive to outlier exclusion in each of the three attitude domains.

To test further the sensitivity of each measure to the influence of outliers, we employed another method to remove outliers. For each measure, we removed the 10% most extreme scores regardless of whether the score was above or below the average score. Therefore, each measure suffered a similar reduction in sample size. This approach revealed very similar results: the AMP was the measure that suffered the most from the outlier removal (see Table 5). The race AMP provides a good example for this effect. Without the 10% most extreme cases, the internal consistency of the race AMP decreased from α = .66 to α = .10, and the average correlation between the AMP and the direct race measures declined from r = .31 to r = .13. Compare that with the resistance of the IAT race measure to the loss of 10% of the most extreme scores, with a small decrease from α = .86 to α = .78 in internal consistency, and from r = .27 to r = .22 in average correlation with direct measures.

Figure 2A plots the average correlation of each implicit race measure with the direct race attitudes measures as a function of the percentage of outlier scores removed. After losing just 5% of the outlier scores, the AMP drops from being the most strongly correlated with direct measures to being the second weakest. The other measures show a steadier and smaller decline as a function of the percentage of outlier scores removed. Figure 2B illustrates a similar pattern for the politics attitude measures, with the AMP showing a steeper decline for the most extreme 10% of outliers in comparison to the other measures. For self-esteem, the fact that all measures suffered very little loss of relations with direct measures is not surprising as only a few of the measures had any meaningful relationship with the direct measures to begin with.

Comparative conclusions for the influence of outliers. The internal consistency of the AMP and the relationship between the AMP and the direct measures were considerably reduced when the analyses did not include outlier scores. Although all measures showed some loss, the AMP loss was much larger than the other measures. For example, with racial attitudes, the AMP went from being the most strongly related to the direct measures to the most weakly related to the direct measures. The AMP's psychometric performance is mostly a function of the outlier scores. The IAT was most robust to the removal of outliers, followed by BIAT, and then by GNAT, ST-IAT, SPF and EPT, which were quite similar to one another.

A



B

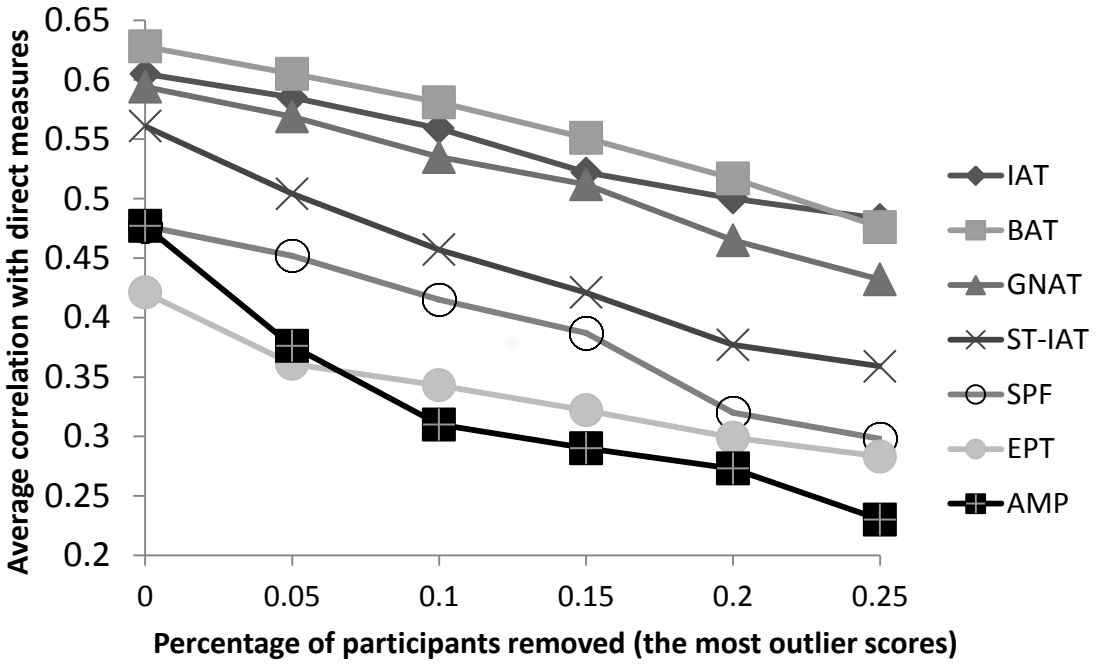


Figure 2. The effect of removing outlier scores (by percentage) from each implicit measure on its average correlation with the direct measures and other criterion variables. Panel A: Race measures; Panel B: politics measures.

Data Exclusion

Better measures will elicit interpretable performance from as many participants on as many response trials as possible. In addition, better measures will be more robust to the exclusion rules such that they maximize psychometric performance with as little data removal as possible, and are relatively insensitive to the application of different exclusion criteria. The common practice of removing participants that misbehave or do not otherwise perform the tasks as instructed reflects the belief that these participants damage the measures' psychometric qualities. We tested the effect of removing participants who were suspect of misbehaving on the psychometric qualities of each measure. Measures that show little increase in their psychometric quality after removing those participants can be considered better because it means that they are less sensitive to misbehaving participants. On the other hand, a decrease in psychometric qualities after removing suspect participants would be a negative feature of a measure because it would raise the suspicion that the measure psychometric qualities depend on participants who do not complete the measure as they are expected to. Therefore, a minor increase in quality would be the appropriate effect of removing suspect participants. Because the measures did not show good psychometric qualities when measuring self-esteem, we focus here on race and politics (the self-esteem measures do not improve substantially when removing suspect participants).

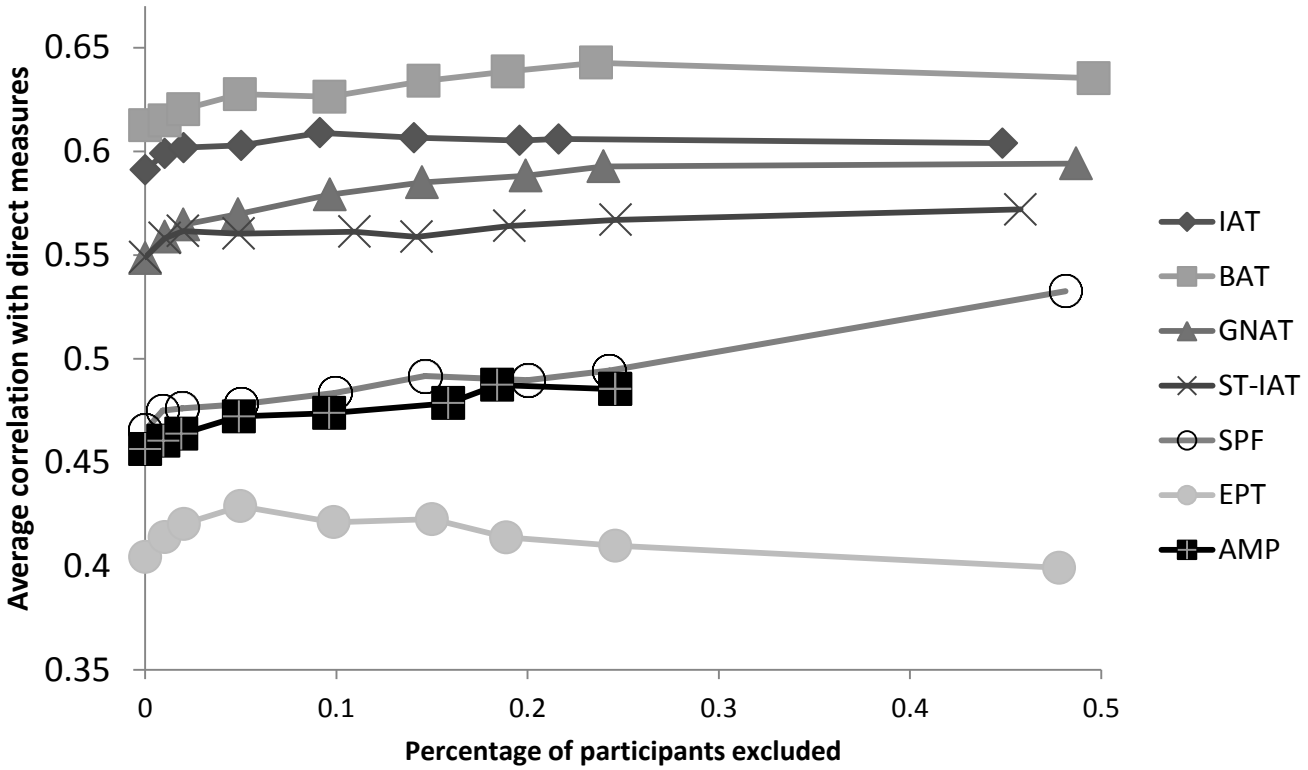


Figure 3. The effect of removing misbehaving participants on the average relationship of the implicit measures with direct measures (politics). About 75% of the participants who performed the AMP had no fast or slow trials, so we could not remove more than 25%.

For each measure, we examined the internal consistency, average correlation with implicit measures and average correlation with explicit measures as a function of removing a

certain percentage of misbehaving participants. The misbehavior score was the percentage of critical trials (i.e., trials in the blocks that were used for scoring) with response that was too fast (below 300 ms), too slow (above 10000 ms; not in the EPT and GNAT because these measures had response deadlines) or incorrect.

We examined the effect of removing the 1%, 2%, 5%, 10%, 15%, 20%, 25% or 50% most misbehaving participants on the psychometric qualities of each measure. The cutoffs were sometimes only approximations (e.g., 1.2% instead of 1%) because the different levels of misbehavior were limited by the number of trials in each measure.

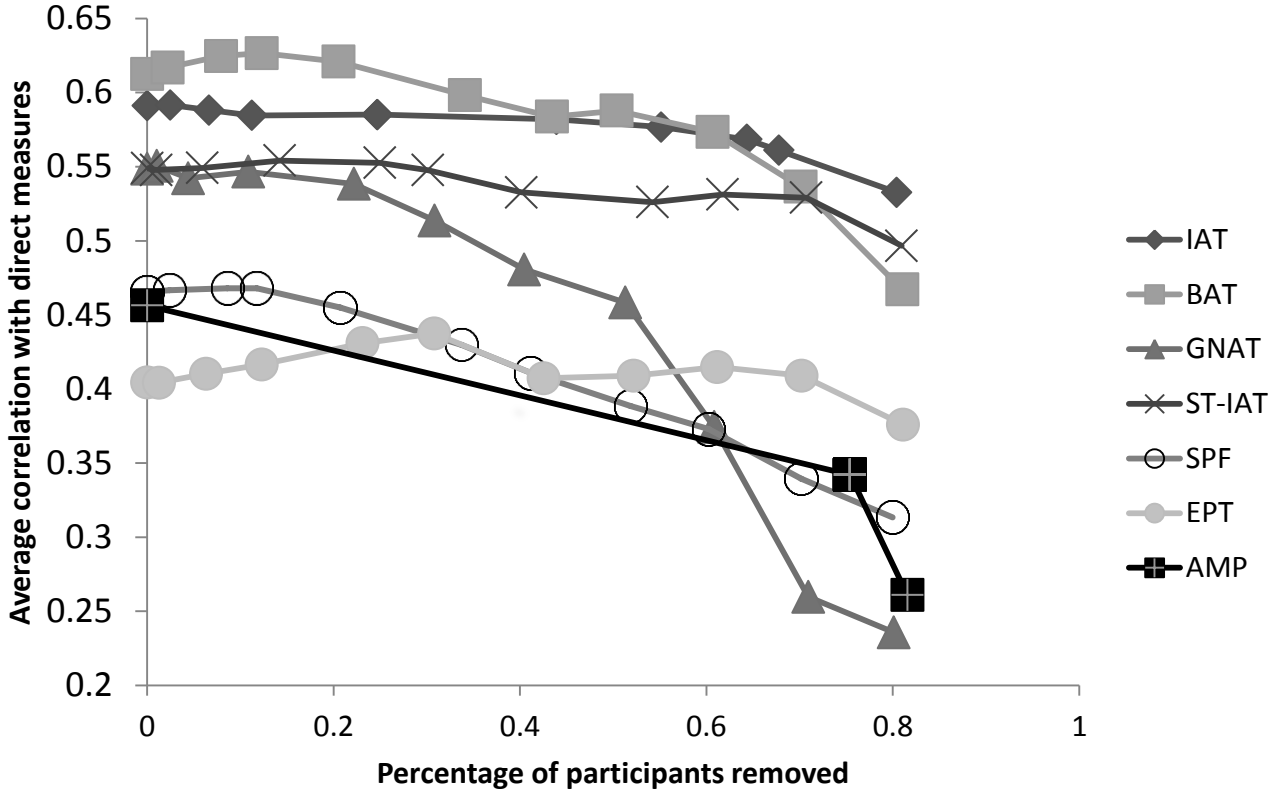


Figure 4. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with direct measures (politics). The AMP has less data points because about 75% of the participants who performed the AMP showed perfect behavior.

The psychometric qualities of most of the measures hardly changed when suspect participants were removed (For all the graphs see the web supplement). For all measures, the internal consistency improved in about .01 when removing the 1% most suspect participants, and in no more than .02 when removing 2% of the participants. The improvement in internal consistency after removing the 50% most suspect participants was always no more than .02, with the exception of the GNAT that improved in .06 when measuring race, and .08 when measuring politics. Similarly, almost all measures showed hardly any increase in their relationship with implicit and explicit measures, regardless of how many suspect participants were removed (For an example, see Figure 3). Notable exceptions were the SPF politics measure that, after removing the 50% most suspect participants, improved in almost .05 in its average correlation with implicit measures and in almost .08 in its average correlation with explicit measures; and the GNAT politics measure that improved in about .04 in both its average correlation with

implicit and explicit measures, after removing the 50% most suspect participants. However, that small decrease in the psychometric qualities after removing 50% suspect participants does not seem a particularly negative attribute.

The small influence of removing suspect participants on the measures may suggest that these participants did not damage the measures, but also that they did not contribute much to the measures' psychometric qualities. To test the latter possibility, we examined the effect of removing the participants who showed the least evidence of misbehavior. That is, we gradually removed participants who had very little suspect trials, and examined whether the psychometric qualities of the measures decreased as a result of losing these "well-behaved" participants. Figure 4 illustrates the general pattern of results by presenting the effect of removing "well-behaved" participants on the average relationship between each measure and the direct measures in the politics domain (all the graphs appear in the web supplement).

Again, for most measures, we found very little evidence that the psychometric qualities of the measures relied mostly on a minority of the participants. Even the measures that showed the most substantial loss in their psychometric qualities showed a relatively small loss. The internal consistency of the GNAT politics measure dropped from .776 to .654 when the 50% most behaved participants were removed. The GNAT politics also showed a drop of about .09 in its average correlation with implicit measures and with explicit measures without the 50% most behaved participants. The SPF politics measures decreased in its average correlation with explicit measures from .465 to .388 when removing the 50% most behaved participants.

Comparative conclusions on exclusion effects. All of the measures except the GNAT showed good insensitivity to the influence of apparently misbehaving participants. The measures' psychometric qualities did not change substantially even without the most misbehaved participants or without the most behaved participants. This suggests that if any of the measures is sensitive to a small number of participants that have different (smaller or larger) contribution to the measure's psychometric qualities, these participants cannot be detected by suspect behavior when performing the task. The only exception to these good results was the GNAT. In comparison to the other measures, the GNAT showed more substantial improvement when removing misbehaving participants, and more substantial loss of psychometric qualities when removing well-behaved participants. The GNAT also had the second-largest error rate (after the EPT). This may suggest that, because it is a response deadline task with time pressure, performing the GNAT may become frustrating for some participants producing substantial individual differences in their ability to perform it properly and provide reliable data. If so, an adaptive response deadline that calibrated the time pressure for each individual (so that it was fast but not too fast) might improve the GNAT on this performance criterion.

General Discussion

The present study compared the psychometric qualities of seven implicit measures across three topics (racial attitudes, political attitudes, self-esteem) using several criteria: internal consistency, test-retest reliability, sensitivity to known-groups effects, relations with other implicit measures of the same topic, relations with direct measures of the same topic, relations with other criterion variables, and sensitivity to outliers and data exclusion. The data provide unprecedented evidence about the psychometric qualities of individual implicit measures, comparative knowledge of psychometric qualities, practical information for the selection of measures for research application, and general knowledge about implicit measurement.

The General Validity of Implicit Measures

The present study provides support for existing claims about implicit measurement that previously have been based on evidence from just one or two implicit measures. All seven implicit measures were: (a) sensitive to known-group differences such as detecting differences in racial attitudes between Blacks and Whites or differences in political attitudes between liberals and conservatives, (b) related to other implicit measures of the same topic, (c) related to direct, explicit measures of the same topic, and (d) predicted criterion variables related to the topic (De Houwer et al., 2009; Fazio & Olson, 2003; Gawronski & Payne, 2010; Greenwald, et al., 2009; Nosek, et al., 2007, 2011, 2012; Teige-Mociemba, et al., 2010; Wentura & Degner, 2010). The evidence leaves no doubt that implicit measures are valid assessments of social cognition, affirming their usefulness for research applications.

Further, the relations between implicit and explicit measures were concordant across measures. In all cases, the implicit-explicit correlation was strongest for political attitudes and weakest for self-esteem. This suggests that theory and evidence accounting for variation in implicit-explicit relations across topics may generalize across implicit measures (e.g., Fazio & Olson, 2003; Nosek, 2005, 2007).

Variations across Attitude Domains

The present study found that the variation in implicit-*implicit* relations was concordant with the variation in implicit-*explicit* relations. Relations among implicit measures were strongest for political attitudes and weakest for self-esteem, just as they were between implicit and explicit measures of those topics. Prior research only found evidence for part of this pattern perhaps leading the theoretical interpretations astray. For example, Bosson and colleagues (2000) found weak relations among implicit measures of self-esteem. The results were presumed to say something general about implicit measurement – implicit measures are weakly related to one another. Had they used politics as the domain of evaluation, they would have arrived at the opposite conclusion. Conversely, using a single implicit measure (the IAT), Nosek (2005, 2007) interpreted variation in implicit-explicit correlations across topics as revealing insights about the interplay between implicit and explicit cognition. The present results may suggest that Nosek (2005) would have observed the same pattern of results had he used two implicit measures instead of one implicit and one explicit measure. If so, then the original evidence might not be about the interplay between implicit and explicit cognition at all.

The present evidence suggests that features of the topic determine relations among measures of the topic regardless of whether they are direct or indirect assessments. Further, the same pattern holds in the present data across topics on explicit measures relations with one another, and implicit measures relations with themselves (internal consistency). All seven measures showed the strongest internal consistency when measuring politics, and the weakest consistency when measuring self-esteem. The same pattern may also apply to relations with

criterion variables as well. For example, Greenwald and colleagues (2009) found evidence that both implicit and explicit measures were stronger predictors of behavior when they were strongly correlated with each other compared to when they were not. In other words, features of the topic may determine the degree to which *any* measure of the construct relates to any other measure of the construct or variables that are predicted by the construct. Currently, we do not have good theoretical account for our novel findings that may suggest what those features might be.

Comparative Conclusions

By conducting a single, large-scale study using the same sample, materials, and procedures with seven implicit measures, we can draw stronger comparative conclusions about the qualities of implicit measures than is possible by comparing measures across the variety of existing research applications. Of the seven implicit measures, the IAT and the BIAT showed the best psychometric qualities consistently across topics and evaluation criteria. Table 2 presents the eight main comparison criteria for each of the three topics (total of 23 because self-esteem did not have a known-groups difference criterion). The BIAT was the best measure on nine of these criteria and the IAT was the best on eight of these criteria. One of these two measures was the second best on 13 of the criteria. In total, only in 13 (22%) of possible 46 times, a measure other than the IAT and BIAT reached the 1st or 2nd place in any of the criteria. The average ranking of the BIAT in the 23 criteria was 2.04, and the average ranking of the IAT was 2.17. Next were GNAT (3.17), ST-IAT (4.26), SPF (4.56), AMP (5.48) and EPT (5.83).

On the other end, EPT had the worst psychometric qualities, and the SPF and AMP were not much better. In total, a measure other than the SPF, AMP, and EPT was ranked 5th, 6th or 7th only 17 (25%) of the 69 possible times. Of these, the AMP's relatively weak psychometric qualities were the most surprising. In particular, removing the scores that were more than two standard deviations away from the task's mean reduced the AMP psychometric qualities markedly. Had those extreme scores been excluded for all evaluation criteria, the AMP likely would have performed the worst overall.

Both AMP and EPT have already proved to be useful research tools in many past applications. However, the present results suggest that these measures may not be the best choice for research applications focusing on individual differences: EPT because of weak reliability (see also Gawronski & De Houwer, in press), main effects and relations with other measures (see also Cameron et al., in press), and AMP because of its reliance on a small number of outliers for acceptable psychometric performance.

Another main conclusion from the present study is that the measures that have received less empirical scrutiny compared to the IAT and EPT (BIAT, ST-IAT, GNAT, and sometimes the SPF and the AMP) often showed acceptable psychometric qualities, relative to what has been found with other implicit measures. Their internal consistency and correlations with other measures were similar to or not far below the strongest performers. Additionally, the psychometric qualities of most measures were not very sensitive to exclusion of outliers, or to the exclusion of well-behaved or misbehaved participants. This is good news for attitude research because it suggests that measurement diversity in the investigation of implicit social cognition is viable. We next discuss findings pertaining to each of the seven measures individually with considerations for potential research applications and innovations to improve their procedures for assessment.

IAT

Among implicit measures, the IAT has earned its status as the most popular tool because of comparatively strong internal consistency, validity, and adaptability for a variety of research applications (Nosek et al., 2011). The present study affirmed its strong psychometric qualities (see also Lane et al., 2007; Nosek et al., 2007; Teige-Mociemba et al., 2010). Despite its strong performance, the IAT may not always be the measure of choice of for research applications. In particular, the IAT is limited to measuring associations in a comparative context (e.g., preference for Democrats compared to Republicans, Blacks compared to Whites, self compared to others). This structure may contribute to its strong psychometric performance, but it is a significant constraint for research application. Analytic strategies to extract "single" evaluations from the IAT are ineffective (Nosek et al., 2005), and use of a meaningless or irrelevant comparison category reduces measurement effectiveness (e.g., Nosek & Smyth, 2011). As a consequence, other measures are needed to avoid this constraint. Also, despite its strong performance, the IAT – like all measures – is influenced by extraneous variables such as the presentation order of the response blocks (Greenwald et al., 1998; Nosek et al., 2007). Continued effort on procedural innovations may yield even stronger performance from this measure (e.g., Nosek et al., 2005; Teige-Mocigemba, Klauer & Rothermund, 2008; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009).

BIAT

The BIAT was developed as a short form of the IAT, but evidence suggests that it may have some unique measurement qualities (Nosek et al., 2012; Sriram & Greenwald, 2009). In particular, while sharing the same structure as the IAT, participants are given just two "focal" concepts and categorize all stimuli as either belonging or not belonging to those concepts. This structure simplifies measurement, making it both easier to learn how to do the task and allowing it to be completed with fewer trials, but it also appears able to assess distinct components of evaluation (e.g., associations with good separately from associations with bad) that are not easily distinguished in the IAT (Nosek et al., 2005). Overall, the BIAT was 16% shorter than the IAT in this study and elicited similar psychometric qualities.

The BIAT's psychometric quality may improve further with refinements to the measurement procedure. Nosek and colleagues (2012) for example examined scoring procedures to maximize the psychometric performance of the BIAT. Further, response blocks with *good* as the focal attribute performed better than response blocks with *bad* as the focal attribute. It is possible that measurement innovation and research evidence will reveal unique contributions of both these associations, or will suggest additional procedural innovations to improve BIAT performance overall. Finally, the BIAT is adaptable to measuring relative association strengths among 3 or 4 different concepts such as relative preferences among Asians, Blacks, Hispanics and Whites or among Christianity, Judaism, Islam, and Hinduism (Nosek et al., 2012). This measurement flexibility may be useful in comparison to the fixed format of the IAT.

GNAT

The GNAT was developed to relax the relative comparison constraint of the IAT and, like the BIAT, has unique qualities. Its relatively good performance in the present study was surprising considering that past evidence suggested that it might be less reliable and valid than the IAT (Nosek & Banaji, 2001). On the positive side, the GNAT performed well on the psychometric criteria, often nearly as good as the IAT and BIAT. Also, it is easy to manipulate the comparative context for measuring single attitudes with the GNAT. Additionally, because the GNAT has a response deadline, the GNAT with error rates can use signal detection analysis to

distinguish sensitivity as the measure of association strengths from response bias indicating a change in response strategy between response blocks. The present research used a relatively relaxed response deadline (1200ms) eliciting fewer errors and so we did not use the signal detection approach.

On the negative side, the present research found a weakness for the GNAT on a criterion that has been never tested before: the GNAT seems to rely more than any other measure on participants performing it correctly (not responding too fast and not committing too many error responses). The GNAT's psychometric qualities were considerably better when poor performing participants were excluded and were considerably worse when the best-behaved participants were excluded. Also, after EPT, GNAT had the higher rate of error and "too fast" trials. These findings suggest that it is relatively difficult to perform the GNAT, and that the difficulty impedes the GNAT's quality as a measure. This may be particularly problematic for research applications that use people with relatively weak cognitive capacity, less experience with computers, or are less tolerant of time pressure tasks. It might also mean that the GNAT is more sensitive to extraneous influence such as individual differences in cognitive capacity making it more difficult to compare across age groups (children, young and older adults) or other groups that could differ on these variables. Whether this is the case requires additional empirical evidence. In summary, the present research found that the GNAT shows acceptable psychometric qualities, but it may require procedural innovations to improve the participants' overall performance, particularly for sensitive applications.

ST-IAT

The ST-IAT was developed to measure attitudes toward a single object in a non-comparative context. In the present research, we examined the quality of the ST-IAT as a relative measure of two categories (a separate investigation using the present data examines the relative quality of the ST-IAT as a measure of a single object, Bar-Anan & Nosek, 2012). The internal consistency of the preference score of the ST-IAT was acceptable (a range of .65-.84). The relationship of the ST-IAT's preference score to other implicit measures and to direct measures were usually better than some of the measures (SPF, EPT and sometimes AMP), and often not far behind the IAT, BIAT and GNAT.

Because it is a relatively easy task, the ST-IAT may seem more vulnerable than other measures to non-automatic processes (Stieger, Gortitz, Hergovich, & Voracek, 2011). Indeed the ST-IAT had the lowest error-rate of all the implicit measures (Table 2). An obvious strategy to perhaps avoid being influenced by association strengths is to focus on the single response (e.g., look for "bad" items) and then categorize anything that does not belong (i.e., the "good" and "Republican" items) with the other key. However, in the present research, the ST-IAT was always related to implicit measure more than to direct measures (Table 2). Additionally, the ST-IAT was usually the fourth best measure on the two main validity criteria (relationship with implicit and with direct measures). This suggests that the ST-IAT might not be heavily influenced by these validity threats in ordinary use. In summary, the present evidence suggests that the ST-IAT performs well, encouraging its further usage, mostly for its unique procedural features.

SPF

The SPF has several unique favorable features. First, all the associations are measured in the same performance block. Therefore, it is probably insensitive to the strategic influences that may affect measures that manipulate the associations between blocks (IAT, GNAT, BIAT and ST-IAT), and there will be no extraneous effects of block order as are common influences on

other tasks, particularly the IAT (Greenwald et al., 1998; Nosek et al., 2005). Additionally, it is possible to compute separate estimates for the association of each category with each attribute, although there is little evidence yet that this provides meaningful estimates of each association.

On the negative side, the present research found that the SPF has worse psychometric qualities than all the "blocked" measures. It was consistently superior only to EPT, and sometimes the AMP and ST-IAT. Because the SPF showed fair validity and reliability it can be used as a measure of association strengths, though it is not likely to be a measure of choice for general use. The present research suggests that it may be most useful for particular applications such as to rule out strategic influences related to the blocked nature of the IAT measures, to examine particular association strengths in a comparative context, or as a secondary implicit measure to replicate effects found with another implicit measure.

EPT

The EPT has a number of favorable features that contribute to its attractiveness for research use despite its comparatively weak psychometric performance. First, because the categories of the attitude object (e.g., Black and White people) are never mentioned explicitly, the EPT is a better measure for the spontaneous evaluation of individual items than any of the categorization tasks (Fazio & Olson, 2003). This feature may contribute to the EPT's weaker performance in the present study because spontaneous evaluations may be unrelated to the social category of interest. For example, using Black and White faces as primes does not guarantee that participants spontaneously evaluate those faces by race in EPT. Some participants might, whereas others might evaluate the items on attractiveness, gender, age, or any combination of features. In categorization tasks like the IAT and GNAT, participants are constrained to categorize the stimuli on a single dimension. Additionally, because the categories are not mentioned explicitly, it might be easier to disguise the EPT's purpose from the participants (although, to the best of our knowledge, there is no empirical evidence that EPT indeed has this advantage on other measures). These are important features that differentiate EPT from most other implicit measures. So, despite the fact that EPT showed the worst internal consistency, the weakest relationship to other implicit measures and the weakest relationship to direct attitude measures – there are a variety of research applications for which categorization tasks are not appropriate and EPT is the best available alternative. Because EPT has low reliability, use of this measure will be most effective by increasing statistical power via other means, such as larger samples than would be used with more reliable measures.

AMP

The AMP is attractive particularly for its procedural distinctiveness from other measures. It is the only implicit measure that has substantial measurement flexibility and widespread use that does not use response latency as a dependent variable. Previous research found that it shows good internal consistency (Payne et al., 2005), good validity (Payne, Govorun & Arbuckle, 2008; Payne et al., 2010; Cameron et al., in press), and it seems to have straightforward procedural validity: attitudes affect performance despite participants' intention to prevent it. Like EPT, the AMP does not mention the categories explicitly, which might make it more suitable for measuring associations with individual items rather than toward social categories. Like EPT, it is possible to use a number of different primes in the AMP, which might enable researchers to measure associations with a number of objects (yet, there is still no research on the effect of the number of categories on the psychometric qualities of the AMP). In summary, previous research suggests that the AMP might have many of the advantages of EPT, but without the disadvantage of poor psychometric qualities.

The present research provides support that AMP is superior to EPT in many psychometric qualities – internal consistency, and relationship with other direct measures. However, AMP was not better than EPT in relationship with other implicit measures. Additionally, AMP was very sensitive to the removal of outlier scores. Outlier scores contributed most of the AMP's positive psychometric qualities. In that respect, the AMP is worse than EPT.

Another interesting finding is that the AMP was the only implicit measure that, on average, was related to direct measures more than to other implicit measures. This may suggest that the AMP is influenced by deliberate evaluations more than by automatic evaluation[3]. In another line of research Bar-Anan and Nosek (in press) found that the AMP's psychometric qualities depend, to a large extent, on a minority of the sample (people who reported that they intentionally rated the primes instead of the target). For the rest of the sample (a range of 41%-62% in our studies), there was little evidence that the AMP measured attitudes at all. The present research suggests that this might be a unique weakness of the AMP, and not a general weakness of implicit measures. Additionally, because our results suggest that many participants are not sensitive to the AMP, perhaps procedural innovations that would target those participants could improve the AMP considerably.

The AMP's psychometric qualities were generally acceptable. Therefore, the present research cannot be considered as evidence that the researchers should not use the AMP. Rather, research applications that use the AMP should examine whether their results depend to a large extent on an outlier minority of the sample. Further, procedural innovations may further strengthen the AMP as an implicit measure,

Summary

The present study compared seven implicit measures on a variety of psychometric qualities. The results provide strong evidence of the superiority of the psychometric qualities of the IAT, but also suggest that a younger measure, the BIAT, is comparable. The worst measure was the EPT. At the very least, this suggests that studies that use the EPT should increase the statistical power (e.g., large sample size) to compensate for EPT's weaknesses, particularly its low internal consistency. The GNAT and ST-IAT had acceptable psychometric qualities encouraging future research applications that would be interested in capitalizing on the unique features of these measures. However, the GNAT was the most dependent on participants performing the task well. The AMP was superior to EPT in most psychometric qualities. However, the AMP's psychometric qualities depended to a large extent on outlier scores. This suggests that—more than any other measure—the AMP is an effective measure for only a minority of the participants. Finally, the SPF was, overall, the second-worst measure. However, because the SPF often showed reasonable psychometric qualities, it might be more suitable than the EPT and the AMP as a secondary implicit measure to test the validity of results found with a main implicit measure (e.g., the IAT) that has unique procedural idiosyncrasies. With this cumulative evidence in hand, the psychometric qualities of implicit measures are now less implicit.

References

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

Bar-Anan, Y., & Nosek, B. A. (in press). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin*.

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

Bar-Anan, Y, & Nosek, B. A. (2012). Unpublished data.

Bar-Anan, Y, Spies, J., & Nosek, B. A. (2012). Unpublished data.

Bluemke, M. & Friese, M. (2007). Reliability and validity of the Single-Target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977-997.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cunningham, W.A., Preacher, K.J., & Banaji, M.R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163-70.

De Houwer, J., & De Bruycker, E. (2007). The implicit association test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, *46*, 401-421.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368.

Deustch, R. & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on "automatic evaluations" as a function of task characteristics of the measure. *Journal of Experimental Social Psychology, 45*, 101-114.

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

Fazio, R.H., & Olson, M.A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

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

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch. R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition and Emotion, 24*, 1008-1025.

Gawronski, B., & De Houwer, J. (in press). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition). New York, NY: Cambridge University Press.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment, 24*, 218–225.

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

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

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

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

Greenwald, A.G., Smith, C.T., Sriram, N., Bar-Anan, Y., & Nosek, B.A. (2009). Race attitude measures predicted vote in the 2008 U. S. Presidential Election. *Analysis of Social Issues and Public Policy, 9*, 241-253.

John, O. P., & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology (pp. 339-369)*. New York, NY: Cambridge University Press.

Jost, J. T., Glaser, J., Kruglanski, A.W., & Sulloway, F. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*, 339-375.

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

Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2011). Reliability of implicit self-esteem measures revisited. *European Journal of Personality, 25*, 239–251.

Lane, K. A., Brescoll, V. L. & Bosson, J. Can implicit self-esteem be measured? A reanalysis of Bosson et al. (2000). Poster presented American Psychological Society Annual Conference. Toronto, Canada. June 2001.

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

McConahay, J.B. (1983). Modern racism and modern discrimination. *Personality and Social Psychology Bulletin, 9*, 551 -558.

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

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

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

Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16*, 65-69.

Nosek, B.A., & Banaji, M.R. (2001). The Go/No-Go Association Task. *Social Cognition, 19*, 625-666.

Nosek, B. A., Bar-Anan, Y., Sriram, N., & Greenwald, A. G. (2012). Understanding and using the brief Implicit Association Test: I. Recommended scoring procedures. Unpublished manuscript.

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

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

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

Nosek, B.A., & Hansen, J.J. (2008). Personalizing the Implicit Association Test increases explicit evaluation of target concepts. *European Journal of Psychological Assessment, 24*, 226-236.

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

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2012). Implicit Social Cognition. In S. Fiske & C. N. Macrae (Eds.) *Handbook of Social Cognition*. New York, NY: Sage.

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

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

Nuttin, J.M. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology*, *15*, 353–361.

Olson, M.A., & Fazio, R.H. (2003). Relations between implicit measures of prejudice. *Psychological Science*, *14*, 636 -639.

Payne, B.K., Burkley, M.A., & Stokes, M.B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*, 16-31.

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

Payne, B. K., Govorun, O., Arbukle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition and Emotion, 22*, 238-271.

Payne, B.K., Krosnick, J.A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*, 367-374.

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

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

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Eliminating the influence of recoding in the Implicit Association Test: The recoding-free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology, 62*, 84-98.

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment, 24*, 273-281.

Schmukle, S. C. and Egloff, B. (2004), Does the Implicit Association Test for assessing anxiety measure trait *and* state variance? *European Journal of Personality, 18*, 483–494.

Sherman, J. W. (2009). Controlled influences on implicit measures: Confronting the myth of process-purity and taming the cognitive monster. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new wave of implicit measures (pp. 391-426)*. Hillsdale, NJ: Erlbaum.

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

Smyth, F. L., & Nosek, B. A. (2005). Unpublished data.

Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., & Eelen, P. (2007). On the predictive validity of indirect attitude measures: Prediction of consumer choice behavior on the basis of affective priming in the picture–picture naming task. *Journal of Experimental Social Psychology*, *43*, 599-610.

Sriram, N., & Greenwald, A.G. (2009). The brief implicit association test. *Experimental Psychology*, *56*, 283-294.

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

Sriram, N., Nosek, B. A., & Greenwald, A. G. (2012). Scale invariant contrasts of response latency distributions. Unpublished manuscript.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment, 25*, 237–245.

Teige-Mociemba, S., Klauer, K.C., & Sherman, J.W. (2010). A practical guide to Implicit Association Tests and related tasks. In B. Gawronski and B. K. Payne (Eds.), *Handbook of Implicit Social Cognition (pp. 117-139)* New York, NY: Guilford.

Wentura, D. & Degner, J. (2010) A practical guide to sequential priming and related tasks. In B. Gawronski and B.K. Payne (Eds.), *Handbook of Implicit Social Cognition (pp. 95-116)* New York, NY: Guilford.

Wigboldus, D. H. J., Holland, R. W., & Van Knippenberg, A. (2004). Single target implicit associations. Unpublished manuscript.

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

Footnotes

[1] This assumes that the source of the reliability is the construct-of-interest and not a stable extraneous influence. This criterion holds even with the theory and evidence that indirect measures assess a mixture of stable and transient components of evaluation (e.g., Schmukle & Egloff, 2004). If there is a stable "trait" component, the better measures should be more sensitive at detecting it. The proportion of variance due to transient "state" factors imposes an asymptote for the maximum stability over time. However, if some measures assess more trait properties of social evaluation whereas other measures assess more state properties, then test-retest reliability might not be an effective distinguishing feature between the measures.

[2] While this may leverage chance and inflate the performance of these measures, we opted for a strategy that maximized measurement performance. Also, using "typical" scoring methods for each did not change the overall pattern of results.

[3] Alternatively, it might only reflect the fact that the other implicit measures—perhaps because they are all based on response interference (Gawronski, Deutsch, LeBel, & Peters, 2008)—share a method variance that is not shared with the AMP. This possibility is examined in detail in a multitrait-multimethod structural analysis using these data (Bar-Anan, Spies, & Nosek, 2012).

Tables

Table 1

Number of Trials that Contributed to the Calculated Score

Measure	IAT	BIAT	GNAT	ST-IAT	SPF	EPT	AMP
N trials	120	128	160	192	120	180	48

Notes. The 48 trials of the AMP do not include 24 trials with the neutral prime; The 160 trials of the GNAT included 64 "No-go" trials that did not provide any latency data (but error-rate was combined to the score).

Table 2

A Summary of the Results

			Reliability				Correlations		Outliers exclusion		
			Internal Consistency		Test-retest		With implicit measures	With direct measures	% shared variance lost after dropping outliers (2 standard deviations)		
Overall			Average alpha		Average correlation		Average correlation	Average correlation	Internal consistency	Corr. with implicit measures	Corr. direct measures
IAT			**.88**		.45		.39	.35	*5.5*	2.7	2.2
BIAT			.83		**.63**		**.41**	**.38**	7.7	2.4	2.7
GNAT			.74		.42		.40	.33	10.0	1.3	2.0
ST-IAT			.77		.48		.36	.31	9.9	**2.8**	2.3
SPF			*.53*		.46		.31	.27	12.5	2.2	*0.8*
EPT			*.57*		*.33*		*.25*	*.24*	16.0	2.1	1.1
AMP			.69		.50		.26	.32	**29.2**	2.7	**4.7**
Race	Mean Effect-size	Known groups effect		95% CI	Correlation	95% CI					
IAT	0.75	**1.12**	**.86**[a]	.85-.87	.40[bc]	.22-.55	**.36**	.27	*6.7*	**3.5**	*1.2*
BIAT	0.73	0.77	.81[b]	.80-.82	**.63[a]**	.50-.73	.34	.27	7.9	2.9	*1.2*
GNAT	**0.83**	0.67	.70[d]	.69-.73	.30[bc]	.13-.45	.35	.27	11.8	1.1	2.3
ST-IAT	0.20	*0.33*	.74[c]	.72-.76	.34[bc]	.16-.49	.30	.24	11.2	2.2	2.2
SPF	0.24	0.76	*.52*[f]	.49-.55	.38[bc]	.21-.53	.24	.24	11.8	1.4	1.8
EPT	0.07	0.57	.54[f]	.51-.57	*.18*[c]	.00-.36	*.20*	*.19*	15.4	1.1	1.5
AMP	*-0.23*	0.39	.66[e]	.64-.68	.33[bc]	.18-.47	*.21*	**.31**	**34.0**	1.5	**7.4**
Politics											
IAT	0.49	**1.49**	**.93[a]**	.92-.93	.65[abc]	.54-.74	.58	.60	*1.9*	3.8	3.7
BIAT	**0.63**	1.40	.89[b]	.88-.90	**.78[a]**	.69-.85	**.60**	**.63**	5.3	3.3	4.2
GNAT	0.47	1.38	.84[c]	.83-.85	.72[ab]	.63-.79	.59	.59	4.9	2.0	3.4
ST-IAT	0.42	1.04	.84[c]	.83-.85	.54[c]	.40-.66	.55	.56	8.1	5.1	5.8
SPF	*0.21*	1.08	*.59*[f]	.57-.61	.58[bc]	.44-.70	.52	.48	13.7	4.0	*2.4*
EPT	0.27	*0.73*	.63[e]	.61-.65	*.51*[c]	.34-.63	.45	*.42*	18.5	5.0	4.0
AMP	0.31	0.88	.81[d]	.80-.82	.73[a]	.65-.79	*.43*	.48	**27.2**	**5.7**	**10.2**
Self											
IAT	**1.31**		.82[a]	.81-.83	*.26*[a]	.09-.41	.21	.14	*7.9*	0.8	0.8
BIAT	1.03		.76[b]	.74-.78	**.42**[a]	.25-.57	**.25**	**.18**	1.2	0.9	1.2
GNAT	1.23		.65[d]	.63-.67	.34[a]	.14-.51	.21	.08	13.3	0.9	*-0.1*
ST-IAT	0.57		.70[c]	.68-.72	.36[a]	.19-.51	.20	.09	1.6	1.0	0.4
SPF	0.96		*.48*[f]	.45-.51	.39[a]	.23-.53	.14	*.06*	12.2	**1.1**	0.1
EPT	0.43		.54[e]	.51-.56	.29[a]	.36-.43	*.07*	.08	14.0	0.1	0.1
AMP	*0.16*		.55[e]	.53-.57	.35[a]	.20-.48	.10	.16	**26.6**	0.9	**1.7**

Notes. The effect sizes were computed from the preference for White people, Democrats and the Self. Mean correlations and internal consistencies were averaged after applying Fisher‘s transformation and then transformed back to correlations; **Bold** font = best performance in the relevant criterion; *Italic* font = worst performance in the relevant criterion; The sample sizes for the test-retest analyses were between 83 and 158 (average = 116); The Cronbach alphas were compared using Feldt test (1969); The test-retest values are correlations between the first and the second tests.

Table 3

Correlations among the Implicit Measures

	IAT	BIAT	GNAT	ST-IAT	SPF	EPT	AMP
Avg. N	395	374	365	387	254	393	509
Range N	294-558	304-496	278-520	278-548	313-524	298-539	414-558
Average							
IAT		.51	.50	.41	**.38**	.24	**.30**
BIAT	**.51**		**.53**	.46	**.38**	**.29**	.28
GNAT	.50	**.53**		**.48**	.33	.27	.28
ST-IAT	.41	.46	.48		.31	*.23*	.26
SPF	.38	.38	.33	.31		.25	*.19*
EPT	*.24*	.29	*.27*	*.23*	.25		.23
AMP	.30	*.28*	.28	.26	*.19*	*.23*	
Race							
IAT		$.49^{a}$	$.48^{a}$	$.33^{bc}$	.28	$.29^{a}$	$.27^{a}$
BIAT	$.49^{a}$		$.42^{a}$	$.40^{ab}$	.28	$.22^{ab}$	$.23^{ab}$
GNAT	$.48^{a}$	$.42^{a}$		$.47^{a}$	.25	$.19^{ab}$	$.24^{ab}$
ST-IAT	$.33^{b}$	$.40^{ab}$	$.47^{a}$		.24	$.15^{b}$	$.16^{b}$
SPF	$.28^{b}$	$.28^{bc}$	$.25^{a}$	$.24^{cd}$		$.20^{ab}$	$.21^{ab}$
EPT	$.29^{b}$	$.22^{c}$	$.19^{b}$	$.15^{d}$	.20		$.16^{b}$
AMP	$.27^{b}$	$.23^{c}$	$.24^{ab}$	$.16^{d}$	.21	$.16^{b}$	
Politics							
IAT		$.65^{a}$	$.70^{a}$	$.62^{a}$	$.60^{a}$	.41	.45
BIAT	$.65^{ab}$		$.70^{a}$	$.63^{a}$	$.63^{a}$	.52	.43
GNAT	$.70^{a}$	$.70^{a}$		$.65^{a}$	$.53^{ab}$	.47	.43
ST-IAT	$.62^{ab}$	$.63^{ab}$	$.65^{a}$		$.48^{bc}$	.42	.47
SPF	$.60^{b}$	$.63^{ab}$	$.53^{ab}$	$.48^{b}$		.45	.38
EPT	$.41^{c}$	$.52^{b}$	$.47^{bc}$	$.42^{b}$	$.45^{bc}$		.43
AMP	$.45^{c}$	$.43^{c}$	$.43^{c}$	$.47^{b}$	$.38^{c}$	.43	
Self							
IAT		$.37^{a}$	$.29^{ab}$	$.21^{ab}$	$.22^{a}$	.00	$.14^{a}$
BIAT	$.37^{a}$		$.36^{a}$	$.32^{a}$	$.18^{ab}$	.10	$.16^{a}$
GNAT	$.29^{ab}$	$.36^{a}$		$.24^{ab}$	$.18^{ab}$	.03	$.16^{a}$
ST-IAT	$.21^{bc}$	$.32^{a}$	$.24^{ab}$		$.18^{ab}$	.11	$.12^{a}$
SPF	$.22^{bc}$	$.18^{b}$	$.18^{bc}$	$.18^{ab}$		.10	$-.03^{b}$
EPT	$.00^{d}$	$.10^{b}$	$.03^{c}$	$.11^{b}$	$.10^{ab}$		$.06^{a}$
AMP	$.14^{c}$	$.16^{b}$	$.16^{bc}$	$.12^{b}$	$-.03^{b}$	.06	

Notes. The average correlation was calculated after applying Fisher's transformation, and then was transformed back to a correlation coefficient; In the overall section: **Bold** = the strongest correlation of that column; *Italics* = the weakest correlation

of that column; In the by-topic sections: in each column, correlations that do not share superscript are significantly different than each other;

Table 4

Correlations of the Implicit Measures with Direct Attitude Measures and other Criterion Variables

Race	Preference	Thermometer	Items	MRS	Contact with Black people	Speeded Report
Effect-size	0.39	0.31	-0.79	--	--	-0.14
Range N	480-630	494-653	464-684	480-671	492-647	421-569
IAT	.32^{ab}	.32^{a}	.21^{b}	.29^{ab}	-.14^{ab}	.31^{ab}
BIAT	.29^{ab}	.32^{a}	.27^{b}	.29^{a}	-.13^{ab}	.28^{ab}
GNAT	.31^{ab}	.25^{ab}	*.20*^{b}	**.32**^{a}	-.18^{ab}	**.37**^{a}
ST-IAT	.23^{bc}	.28^{a}	.27^{b}	.24^{ab}	-.11^{ab}	.29^{ab}
SPF	.28^{ab}	.28^{a}	.21^{b}	*.18*^{b}	-**.20**^{a}	.27^{ab}
EPT	*.15*^{c}	*.17*^{b}	.26^{b}	.22^{ab}	-*.09*^{b}	*.24*^{b}
AMP	**.35**^{a}	**.33**^{a}	**.41**^{a}	.29^{ab}	-.13^{ab}	.33^{ab}

Politics	Preference	Thermometer	Items	RWA	Voted	Voting intentions	Speeded Report
Effect	0.62	0.60	0.46	--	--	--	0.47
Avg. N	554	561	431	559	284	523	547
Range N	468-600	516-607	396-453	472-593	234-316	444-572	459-589
IAT	.64^{ab}	.60^{a}	**.69**^{a}	-.43^{bc}	**.66**^{a}	.51^{a}	**.64**^{a}
BIAT	**.66**^{a}	**.65**^{a}	**.69**^{a}	-**.57**^{a}	.65^{a}	**.54**^{a}	.61^{a}
GNAT	.65^{ab}	.60^{a}	**.69**^{a}	-.49^{ab}	.65^{a}	.46^{abc}	.58^{a}
ST-IAT	.58^{b}	.59^{ab}	.56^{bc}	-.44^{bc}	.64^{a}	.49^{ab}	.61^{a}
SPF	.48^{c}	.46^{c}	.58^{b}	-.39^{cd}	.51^{b}	.41^{bc}	*.47*^{c}
EPT	*.43*^{c}	.43^{c}	*.46*^{c}	-*.33*^{d}	*.43*^{b}	*.36*^{c}	.49^{bc}
AMP	.48^{c}	.51^{bc}	.59^{b}	-.36^{cd}	.49^{b}	.35^{c}	.52^{bc}

Self	Preference	Thermometer	Rosenberg	Speeded Report
Effect	0.44	0.37	--	0.44
Avg. N	601	604	591	534
Range N	459-691	462-693	494-667	450-579
IAT	.11^{ab}	.13	.17^{a}	.14^{ab}
BIAT	.14^{a}	.16	**.18**^{a}	**.24**^{a}
GNAT	*.00*^{b}	.11	.06^{abc}	.13^{ab}
ST-IAT	.05^{ab}	.12	.11^{ab}	*.07*^{b}
SPF	.10^{ab}	*.06*	.00^{bc}	.09^{b}
EPT	.13^{a}	*.06*	-*.04*^{c}	.16^{ab}
AMP	**.16**^{a}	**.17**	.07^{abc}	**.24**^{a}

Notes. In each column of each topic section, correlations that do not share superscript are significantly different from each other; The thermometer and the items columns refer to a difference score.

Table 5

The Influence of Excluding Outlier Scores on the Psychometric Qualities of the Measure

	Internal consistency (alpha Cronbach)					**Average correlation with implicit measures**					**Average correlation with direct measures**				
		Exclusion method					Exclusion method					Exclusion method			
	All cases	By STD		By percent		All cases	By STD		By percent		All cases	By STD		By percent	
	Alpha	Alpha	% loss	Alpha	% loss	R	R	% loss	R	% loss	R	R	% loss	R	% loss
Overall															
IAT	**.88**	**.85**	*5.2*	**.81**	*11.9*	.39	.35	2.7	**.34**	3.8	.35	.32	2.2	.30	3.0
BIAT	.83	.78	8.0	.73	15.6	**.41**	.37	2.4	**.34**	4.4	**.38**	**.34**	2.7	**.34**	2.8
GNAT	.77	.67	9.9	.59	19.5	.40	**.38**	1.3	**.34**	3.5	.33	.31	2.0	.29	3.2
ST-IAT	.74	.70	8.8	.62	19.5	.36	.31	**2.8**	.26	**5.4**	.31	.26	2.4	.23	4.7
SPF	*.53*	*.39*	12.9	.26	21.3	.31	.27	2.2	.23	3.9	.27	.24	*0.8*	.22	2.8
EPT	.57	.41	16.0	*.25*	25.0	*.25*	.20	2.1	.18	*3.4*	*.23*	*.20*	1.1	*.18*	*2.4*
AMP	.69	.39	**32.4**	.21	**35.6**	.26	*.19*	2.7	*.16*	3.8	.32	*.20*	**5.0**	*.18*	**7.8**
Race															
IAT	**.86**	**.82**	*6.7*	**.78**	*13.5*	**.36**	.30	**3.5**	.29	4.5	.27	.24	*1.2*	.22	2.4
BIAT	.81	.76	7.9	.70	17.3	.34	.29	2.9	.26	**4.8**	.27	**.25**	*1.2*	**.24**	*1.6*
GNAT	.71	.61	13.2	.53	21.6	.35	**.33**	1.1	**.30**	2.8	.27	.23	2.2	.22	2.6
ST-IAT	.74	.66	11.2	.58	21.3	.30	.25	2.2	.21	3.6	.24	.18	2.3	.15	3.2
SPF	*.52*	.39	11.8	.26	23.0	.24	.21	1.4	.17	2.6	.24	.20	1.8	.18	2.6
EPT	.54	*.37*	15.0	.21	24.6	*.20*	*.16*	1.1	.15	*1.6*	*.19*	.15	1.5	*.13*	2.0
AMP	.66	.31	**34.0**	*.10*	**42.7**	*.21*	.17	1.5	*.14*	2.2	**.31**	*.12*	**8.4**	.13	**8.2**
Politics															
IAT	**.93**	**.92**	*1.9*	**.90**	*6.0*	.58	.54	3.8	.52	*5.5*	.60	.57	3.7	.56	5.2
BIAT	.89	.86	5.3	.83	9.9	**.60**	**.57**	3.3	**.53**	6.5	**.63**	**.59**	4.3	**.58**	5.6
GNAT	.84	.81	4.9	.76	13.1	.59	**.57**	2	**.53**	*5.5*	.59	.56	3.4	.54	6.6
ST-IAT	.84	.79	6.5	.74	15.6	.55	.49	5.1	.42	**11.2**	.56	.51	5.7	.46	*1.5*
SPF	.59	*.46*	13.7	.32	24.5	.52	*.47*	4.0	.42	7.7	.48	.45	*2.4*	.42	5.6
EPT	.63	*.46*	18.5	*.34*	28.0	.45	.38	5.0	.33	8.2	*.42*	.37	3.9	.34	5.9
AMP	.81	.62	**27.2**	.56	**34.8**	*.43*	*.35*	**5.7**	*.31*	8.1	.48	*.35*	**10.3**	*.31*	**13.2**
Self															
IAT	**.82**	**.77**	*7.9*	**.72**	*16.1*	.21	.19	0.8	.17	1.3	.14	.10	0.8	.06	1.5
BIAT	.76	.69	1.2	.62	19.7	**.25**	**.23**	0.9	**.21**	1.8	**.18**	**.14**	1.2	**.14**	1.2
GNAT	.65	.55	13.3	.43	23.7	.21	.19	0.9	.16	**2.2**	.08	.09	*-0.1*	.06	0.4
ST-IAT	.65	.62	1.6	.52	21.6	.20	.18	1	.15	1.5	.09	.06	0.4	*.05*	0.5
SPF	*.48*	.33	12.2	.19	19	.14	.10	**1.1**	.08	1.4	*.06*	*.05*	0.1	*.05*	*0.1*
EPT	.54	.39	14.0	.26	22.5	*.07*	*.05*	0.1	.05	*0.3*	.08	.07	0.1	.08	*0.1*
AMP	.55	*.19*	**26.6**	*-.09*	**29.3**	.10	*.05*	0.9	*.03*	1.1	.16	.11	**1.6**	.09	**1.9**

Notes. The removal by standard deviation removed scores that were 2 standard deviations away from the mean; The removal by percentiles removed the 10% most extreme scores; The average % loss is the average loss of shared variance.

Figures

Figure 1

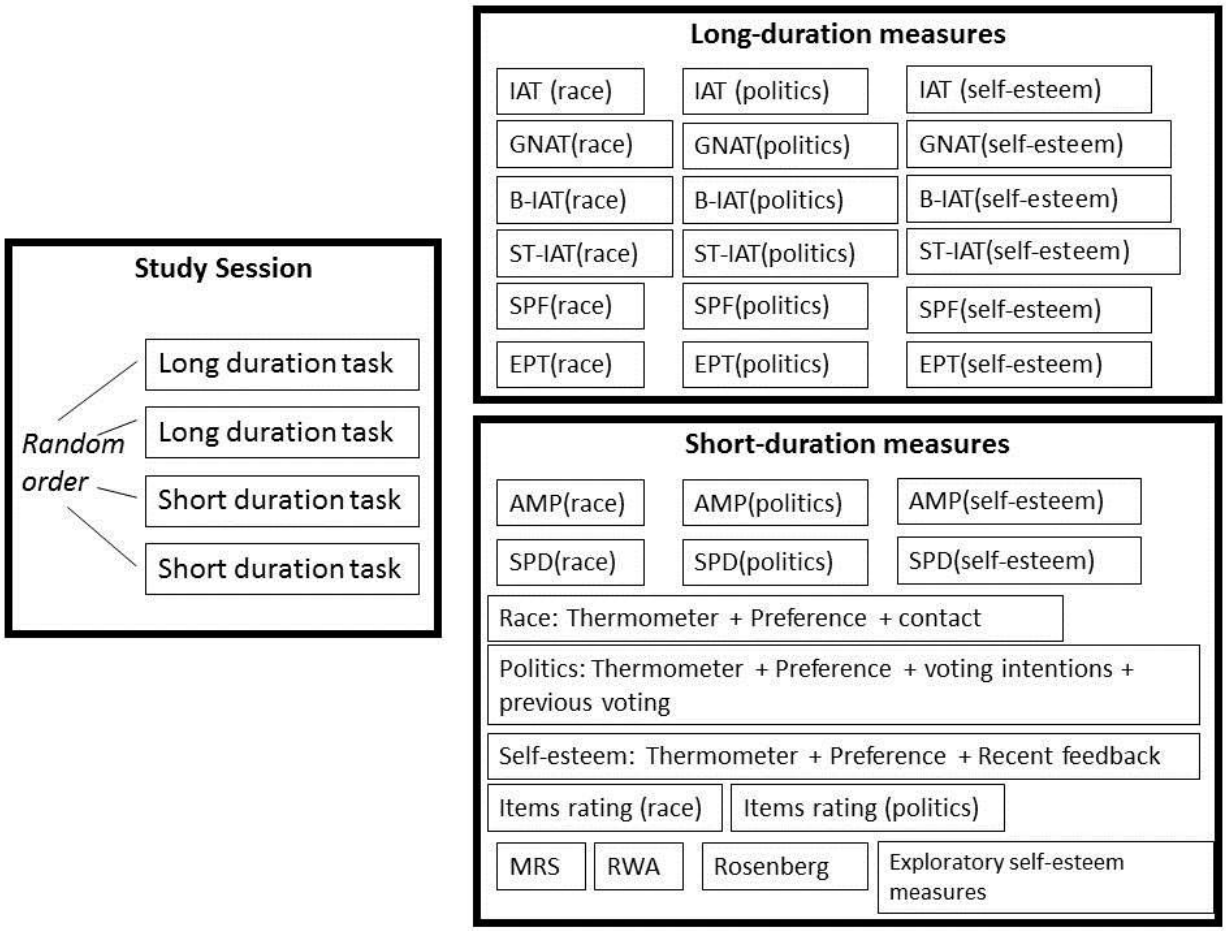
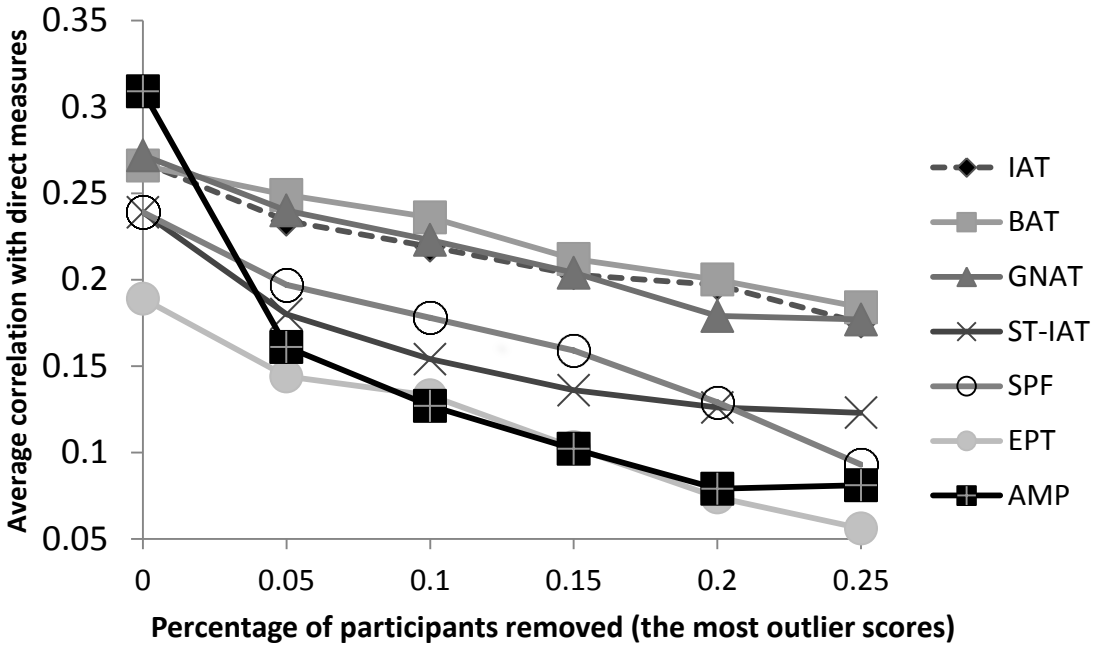


Figure caption: Illustration of the study procedure. Each study session included 2 long and short tasks, selected and ordered randomly. The tasks are listed on the right. Measures that share a rectangle were presented together in the same questionnaire.

Figure 2

A


0.35
0.3
0.25
0.2
0.15
0.1
0.05
Average correlation with direct measures
0
0.05
0.1
0.15
0.2
0.25
Percentage of participants removed (the most outlier scores)
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B

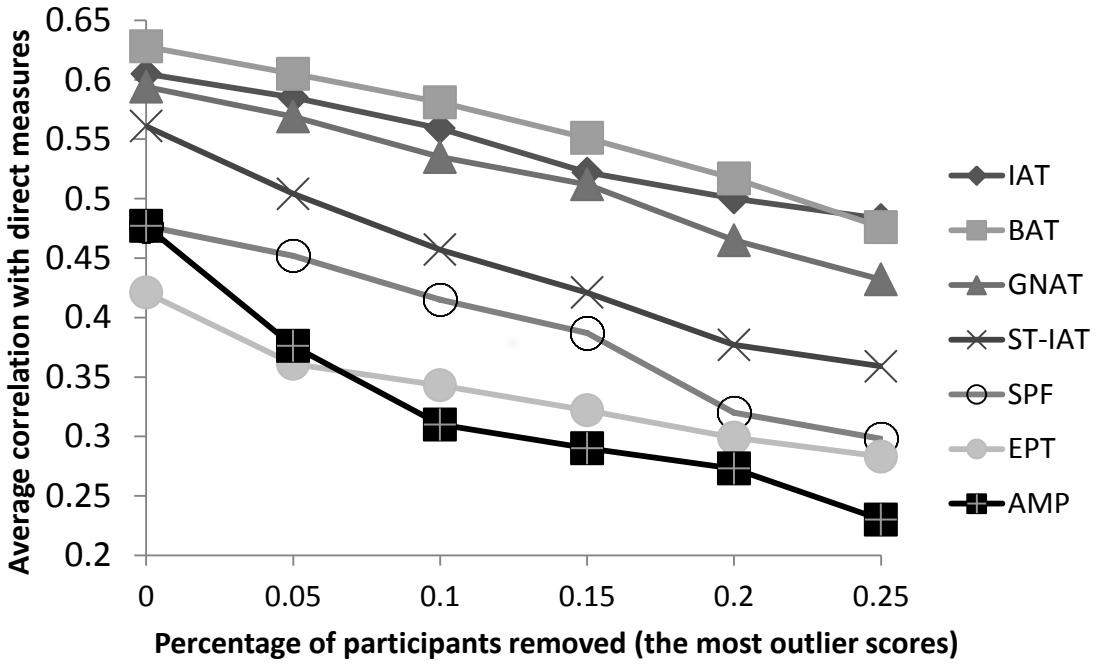

0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
Average correlation with direct measures
0
0.05
0.1
0.15
0.2
0.25
Percentage of participants removed (the most outlier scores)
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

Figure caption: The effect of removing outlier scores (by percentage) from each implicit measure on its average correlation with the direct measures and other criterion variables. Panel A: Race measures; Panel B: politics measures.

Figure 3

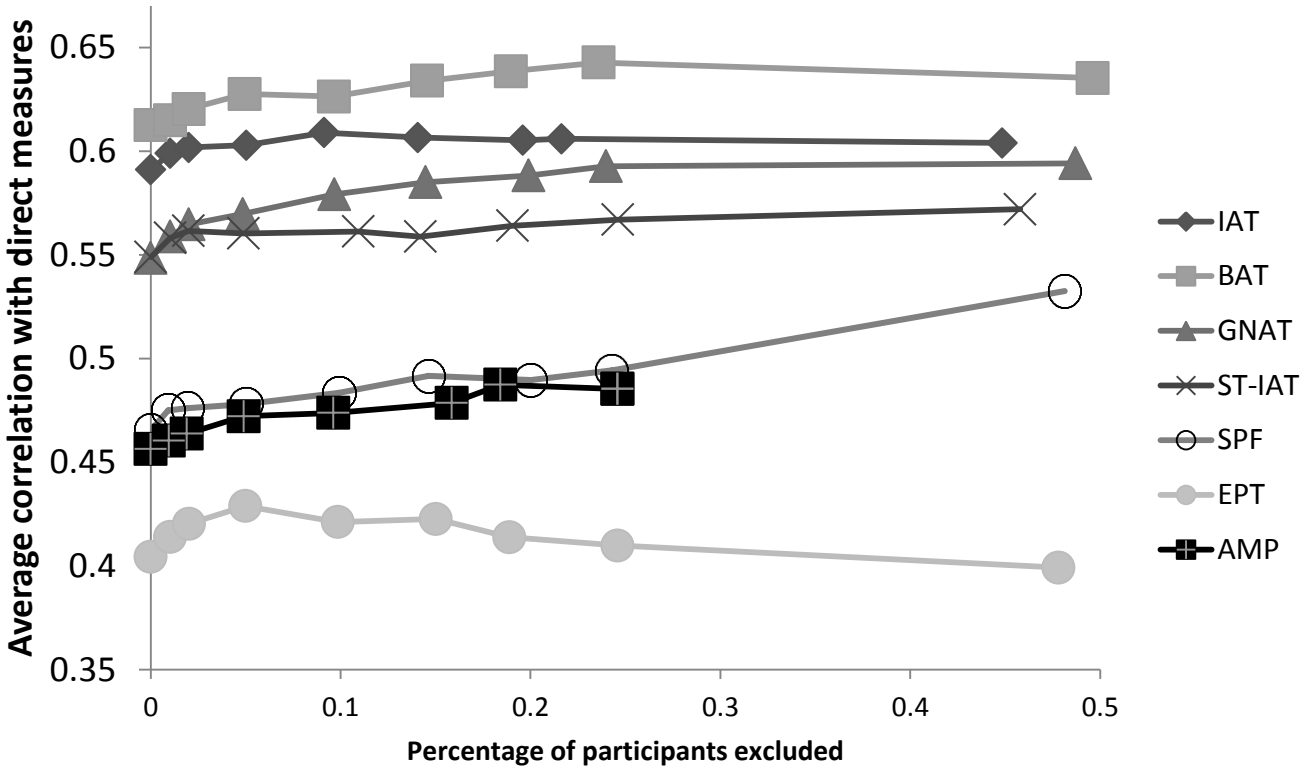


Figure Caption: The effect of removing misbehaving participants on the average relationship of the implicit measures with direct measures (politics). About 75% of the participants who performed the AMP had no fast or slow trials, so we could not remove more than 25%.

Figure 4

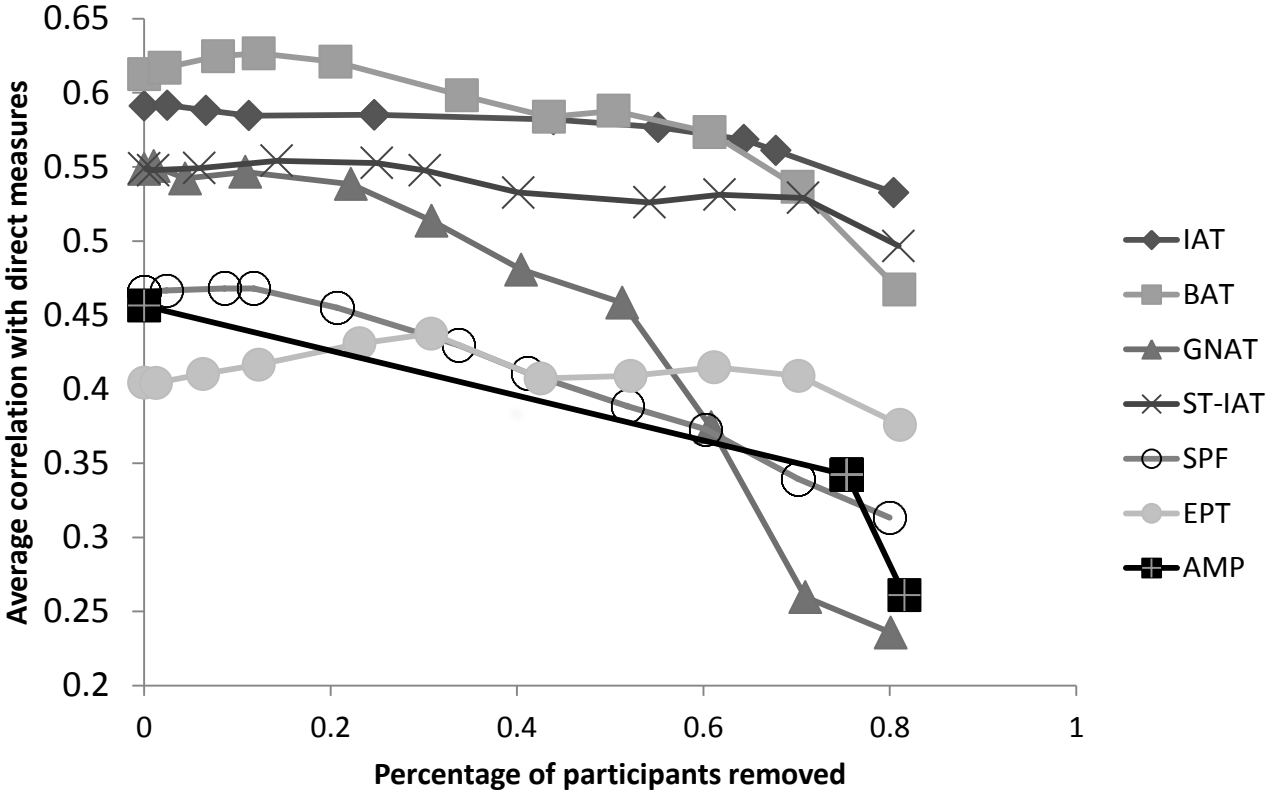


Figure caption. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with direct measures (politics). The AMP has less data points because about 75% of the participants who performed the AMP showed perfect behavior.

Appendix: Additional Tables

Table A1

Main Effects for All the Attitude Measures

Measure	Race				Politics				Self			
	N	Mean	SD	d	N	Mean	SD	d	N	Mean	SD	d
IAT	3043	0.30	0.40	0.75	2955	0.27	0.55	0.49	2943	0.46	0.35	**1.31**
BIAT	2783	0.24	0.33	0.73	2635	0.27	0.43	**0.63**	2639	0.31	0.30	1.03
GNAT	2987	0.67	0.81	**0.83**	2774	0.42	0.89	0.47	2205	1.01	0.82	1.23
ST-IAT	2955	0.10	0.50	0.20	2774	0.25	0.59	0.42	2863	0.25	0.44	0.57
SPF	2950	0.12	0.49	0.24	2739	0.12	0.56	*0.21*	2902	0.46	0.48	0.96
EPT	2947	0.03	0.46	0.07	2836	0.14	0.52	0.27	3077	0.20	0.46	0.43
AMP	3091	-0.05	0.22	*-0.23*	3191	0.09	0.29	0.31	3177	0.03	0.19	*0.16*
Preference	3659	0.40	1.03	0.39	3511	1.14	1.82	0.63	3856	0.59	1.35	0.63
Thermometer	3779	0.55	1.78	0.31	3568	2.41	4.00	0.60	3869	0.76	2.08	0.44
Items rating	3843	-0.83	1.05	-0.79	3142	1.25	2.73	0.46				
Speeded	3284	-0.09	0.64	-0.14	3400	0.54	1.14	0.47	3335	0.31	0.71	0.37

Notes. Cohen's d indicates the magnitude of the effect compared to 0 (no preference between Blacks and Whites for race, liberals and conservatives for politics, and self and others for self); In the columns of measure names, *preference* was the self-reported preference, *thermometer* was the difference between thermometer rating of the two categories, *items rating* was the difference between the rating of the individual items used for each category, and *speeded* was the difference score in the speeded rating measure.

Table A2

Known Groups Differences for All the Attitude Measures

	Race							**Politics**						
	White participants			Black participants				Liberals			Conservatives			
	N	M	SD	N	M	SD	d	N	M	SD	N	M	SD	d
IAT	2290	0.33	0.39	165	-0.07	0.32	**1.12**	1602	0.51	0.41	631	-0.20	0.54	**1.49**
BIAT	2072	0.27	0.32	156	0.02	0.33	0.77	1473	0.45	0.34	554	-0.09	0.43	1.40
GNAT	2243	0.71	0.77	160	0.15	0.89	0.67	1538	0.79	0.70	576	-0.31	0.88	1.38
ST-IAT	2195	0.12	0.50	164	-0.05	0.53	*0.33*	1491	0.46	0.53	591	-0.13	0.60	1.04
SPF	2198	0.16	0.49	167	-0.21	0.48	0.76	1511	0.31	0.48	570	-0.25	0.56	1.08
EPT	2184	0.06	0.45	153	-0.19	0.42	0.57	1538	0.27	0.52	608	-0.10	0.49	*0.73*
AMP	2315	-0.05	0.22	181	-0.14	0.24	0.39	1716	0.18	0.28	706	-0.08	0.31	0.88
Preference	2714	4.52	0.90	215	2.99	1.32	1.38	1910	6.17	1.09	730	3.18	1.88	2.01
Thermometer	2791	0.76	1.66	226	-1.45	1.95	1.22	1934	4.51	2.93	745	-1.52	3.99	1.74
Items Rating	2827	-0.79	1.06	204	-1.26	1.11	0.43	1627	2.71	1.95	682	-1.38	2.75	1.74
Speeded	2419	-0.03	0.59	180	-0.63	0.81	0.86	1802	1.10	0.95	704	-0.46	1.10	1.52
Scale	2750	2.04	0.91	236	1.56	0.62	0.63	1882	2.28	0.72	780	3.44	0.81	-1.52

Notes. Cohen's d values indicate the magnitude of the difference between Whites and Blacks for race, and between liberals and conservatives for politics; In the columns of measure names, *preference* was the self-reported preference, *thermometer* was the difference between thermometer rating of the two categories, *items rating* was the difference between the rating of the individual items used for each category, and *speeded* was the difference score in the speeded rating measure.

Table A3

Correlations of Each Measure with Other Measures, Including the Average Ranking of Each Measure's Relatedness to Each of the Other Measures

	Average correlations			
Measure	With indirect measures		With direct measures	
Overall	Average correlation	Average rank	Average correlation	Average rank
IAT	.39	2.1	.35	2.8
BIAT	**.41**	**1.9**	**.38**	**2.3**
GNAT	.40	2.2	.33	3.8
ST-IAT	.36	3.4	.31	4.5
SPF	.31	4.0	.27	5.5
EPT	*.25*	*5.4*	*.23*	*5.9*
AMP	.26	5.2	.32	3.1
Race				
IAT	**.36**	**1.5**	.27	3.0
BIAT	.34	2.0	.27	3.7
GNAT	.35	2.2	.27	3.3
ST-IAT	.30	4.0	.24	4.7
SPF	.24	4.0	.24	4.8
EPT	*.20*	*5.5*	*.18*	*6.3*
AMP	*.21*	5.2	**.31**	**2.2**
Politics				
IAT	.58	2.8	.60	2
BIAT	**.60**	2.0	**.63**	**1.6**
GNAT	.59	**1.8**	.59	2.7
ST-IAT	.55	3.3	.56	4.0
SPF	.52	4.0	.48	5.6
EPT	.45	5.2	*.42*	*6.7*
AMP	*.43*	*5.3*	.48	5.4
Self				
IAT	.21	2.7	.14	3.3
BIAT	**.25**	**1.8**	**.18**	**1.8**
GNAT	.21	**2.3**	.08	5.5
ST-IAT	.20	3.0	.09	5.0
SPF	.14	4.0	*.06*	*6.0*
EPT	*.07*	5.7	.08	4.8
AMP	.10	*5.0*	.16	1.8

Table A4

Number of Outlier Scores (2 standard deviations from the average score)

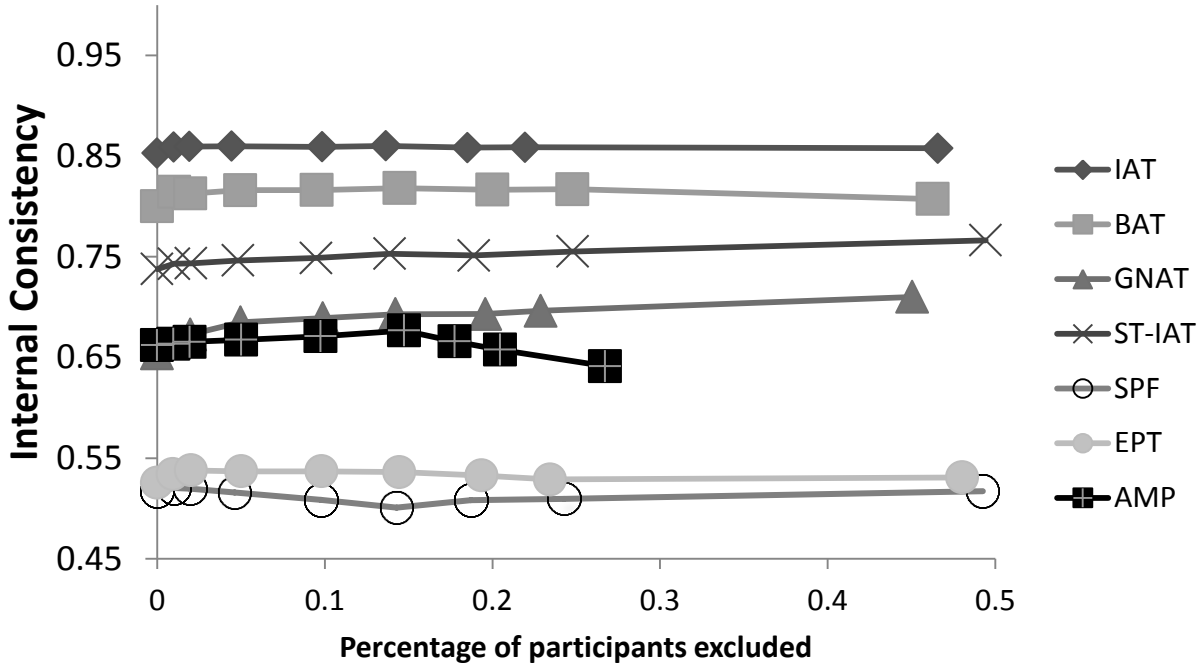
	Race **Total**	Race **Outliers**	Politics **Total**	Politics **Outliers**	Self **Total**	Self **Outliers**
IAT	3043	139 (5%)	2956	99 (3%)	2943	130 (4%)
BIAT	2269	124 (5%)	2635	91 (3%)	2639	113 (4%)
ST-IAT	2955	135 (5%)	2744	128 (5%)	2862	117 (4%)
GNAT	2987	120 (4%)	2774	114 (4%)	2203	90 (4%)
SPF	2950	133 (5%)	2739	129 (5%)	2902	139 (5%)
EPT	2947	139 (5%)	2837	148 (5%)	3077	151 (5%)
AMP	3091	165 (5%)	3191	284 (9%)	3177	133 (4%)

Figures for Web Supplement

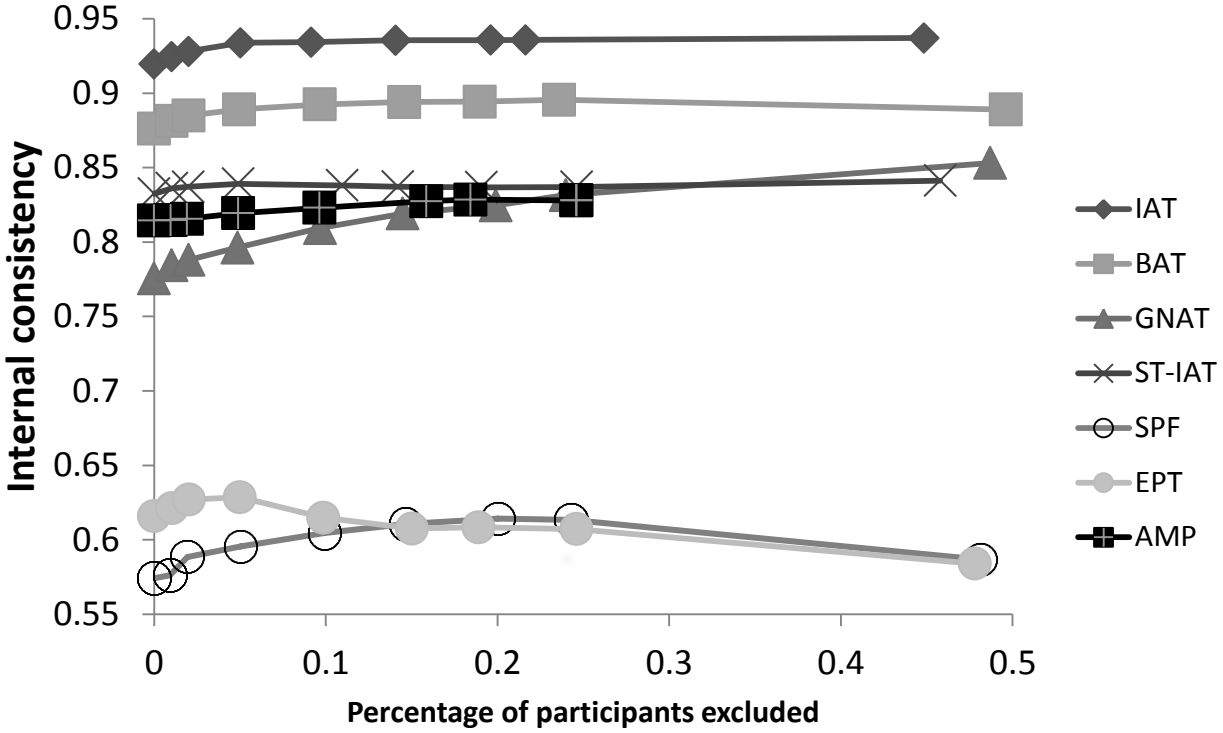
The figures present the full details of the data exclusion analyses reported in the main manuscript.

Figure W1

A



B


0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
Internal consistency
0
0.1
0.2
0.3
0.4
0.5
Percentage of participants excluded
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

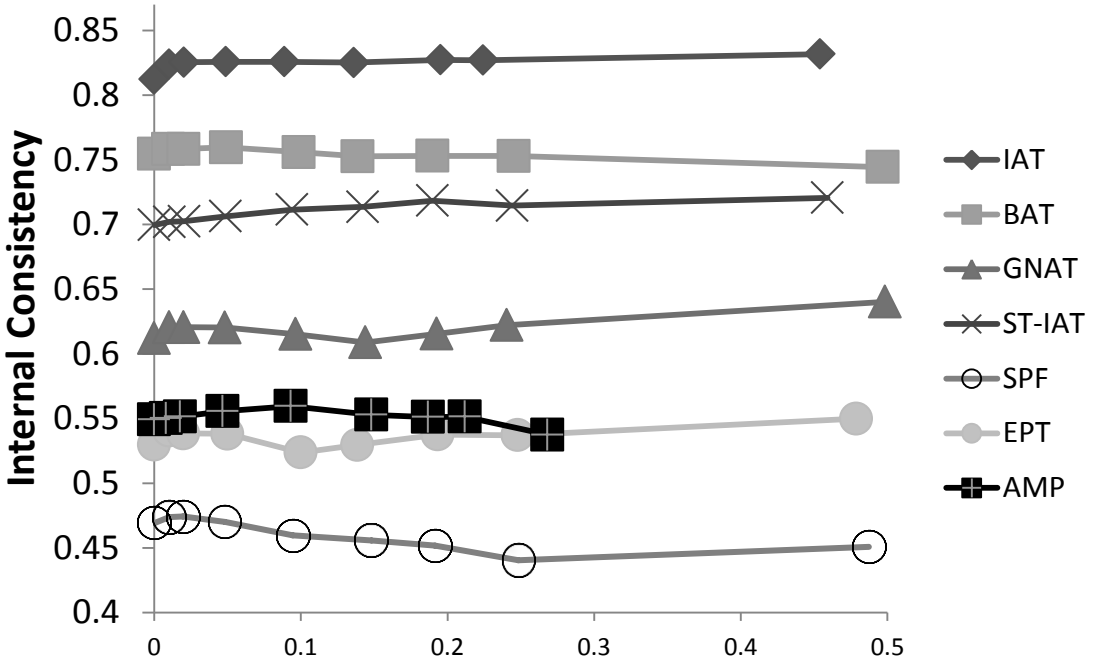
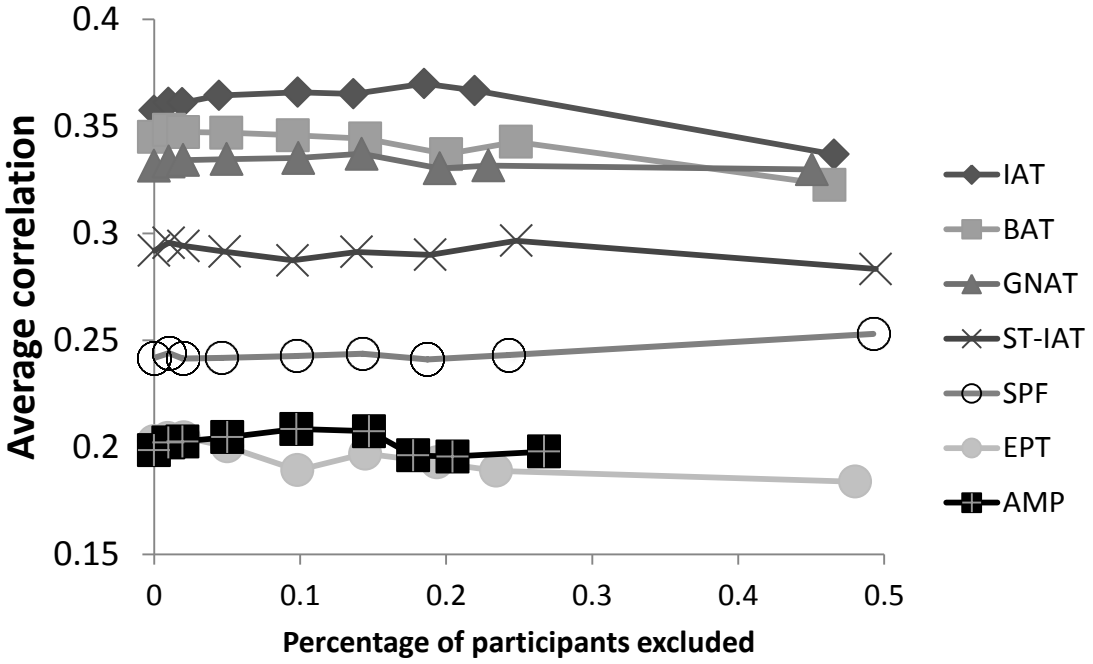

0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Internal Consistency
0
0.1
0.2
0.3
0.4
0.5
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

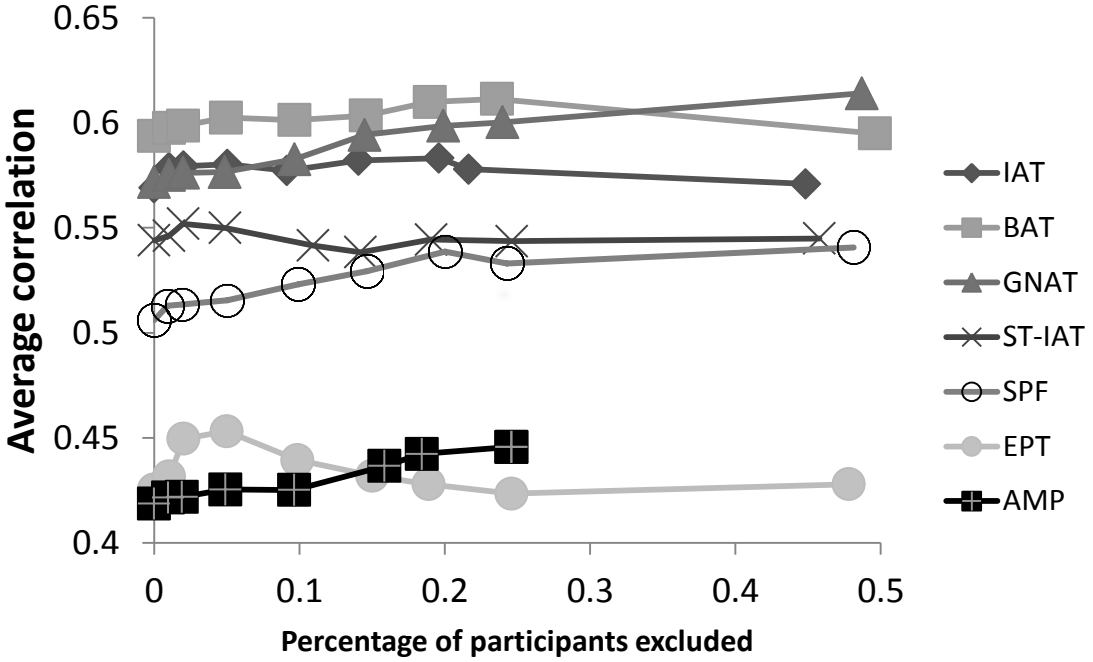
Figure caption: The effect of removing misbehaving participants on the internal consistency of the implicit measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

Figure W2

A


0.4
0.35
0.3
0.25
0.2
0.15
Average correlation
0
0.1
0.2
0.3
0.4
0.5
Percentage of participants excluded
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B


0.65
0.6
0.55
0.5
0.45
0.4
Average correlation
0
0.1
0.2
0.3
0.4
0.5
Percentage of participants excluded
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

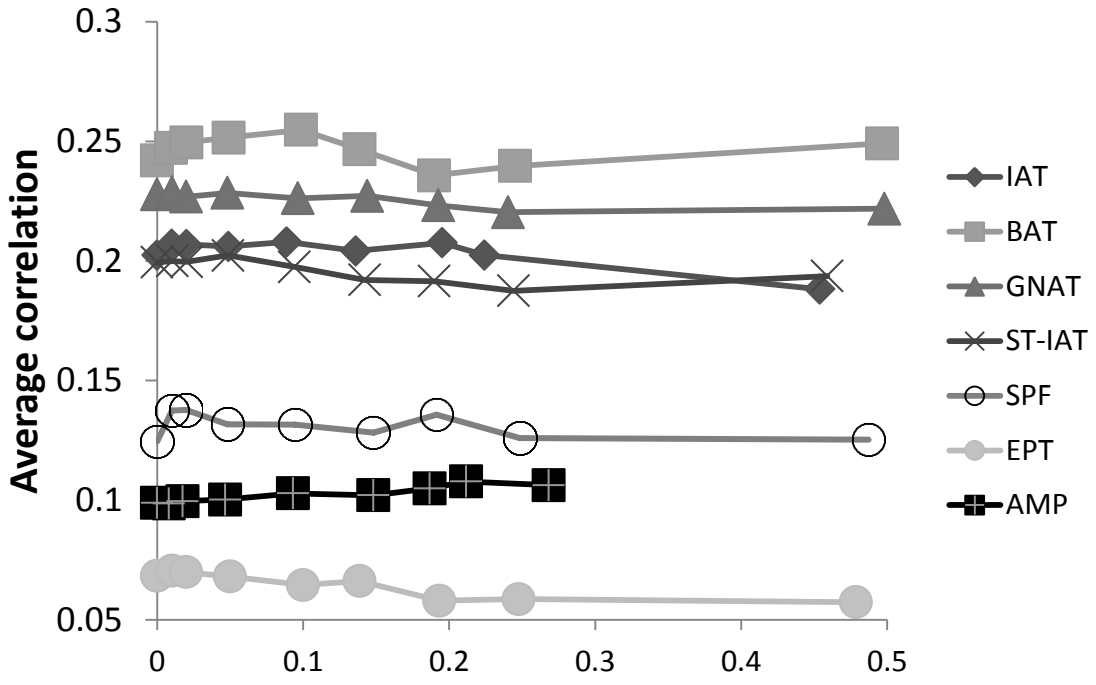
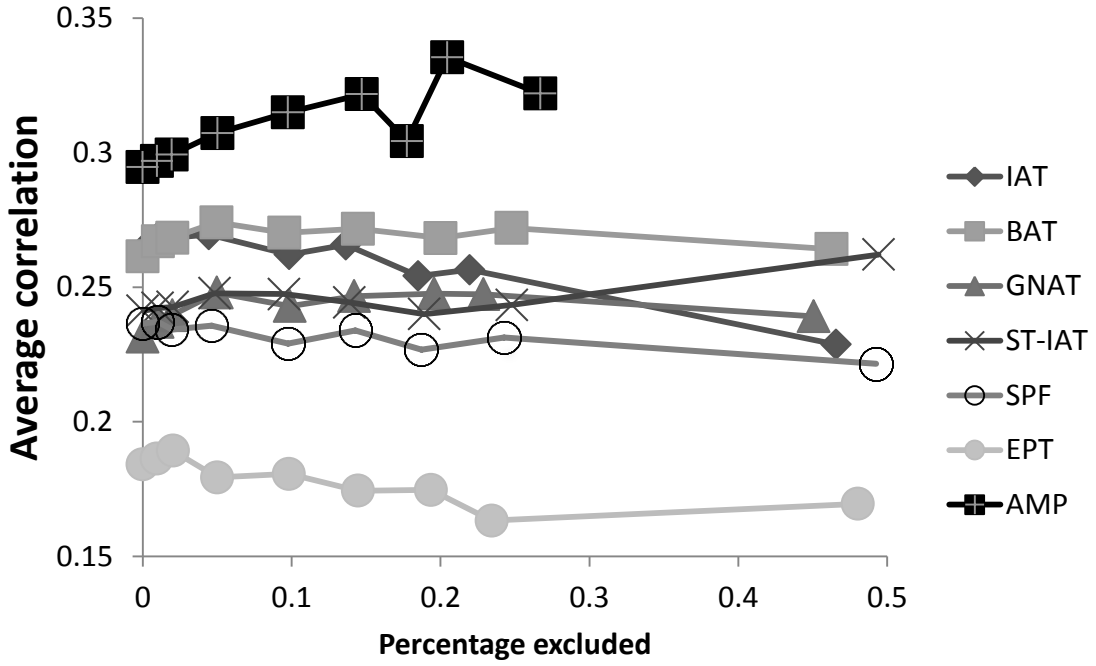


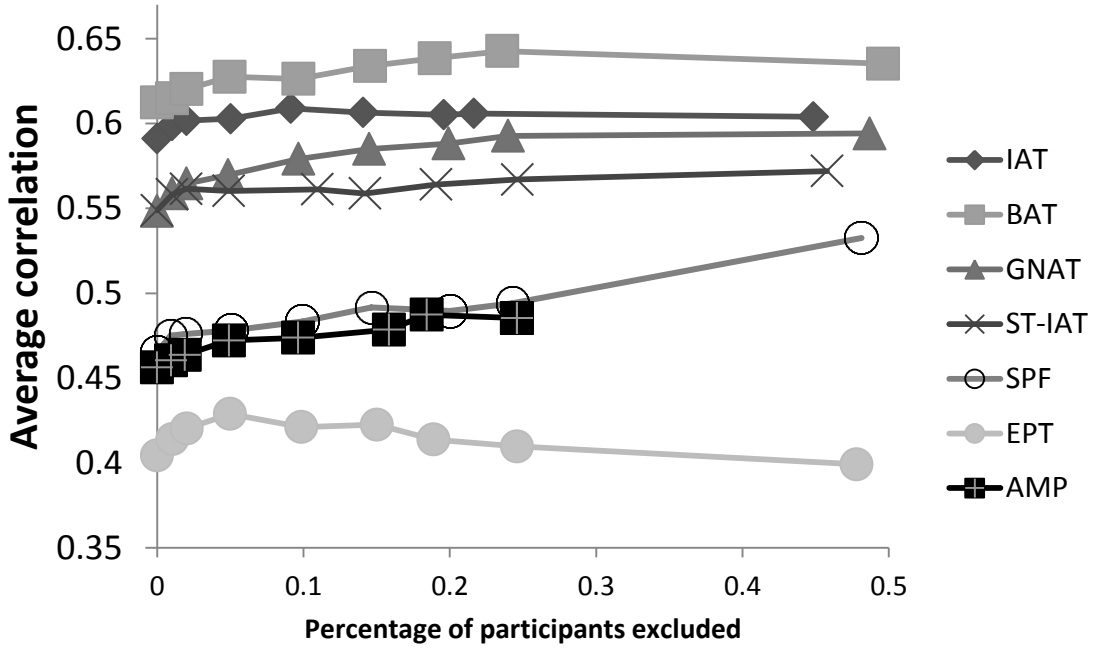
Figure caption: The effect of removing misbehaving participants on the average relationship of the implicit measures with other implicit measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

Figure W3

A


0.35
0.3
0.25
0.2
0.15
Average correlation
0
0.1
0.2
0.3
0.4
0.5
Percentage excluded
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B


0.65
0.6
0.55
0.5
0.45
0.4
0.35
Average correlation
0
0.1
0.2
0.3
0.4
0.5
Percentage of participants excluded
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

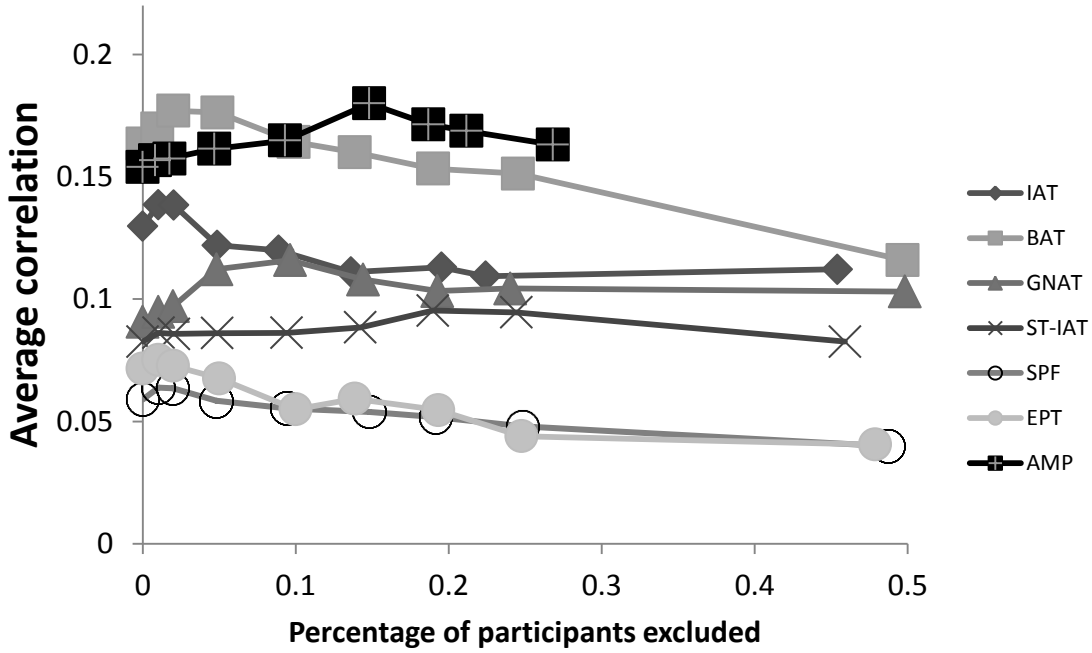
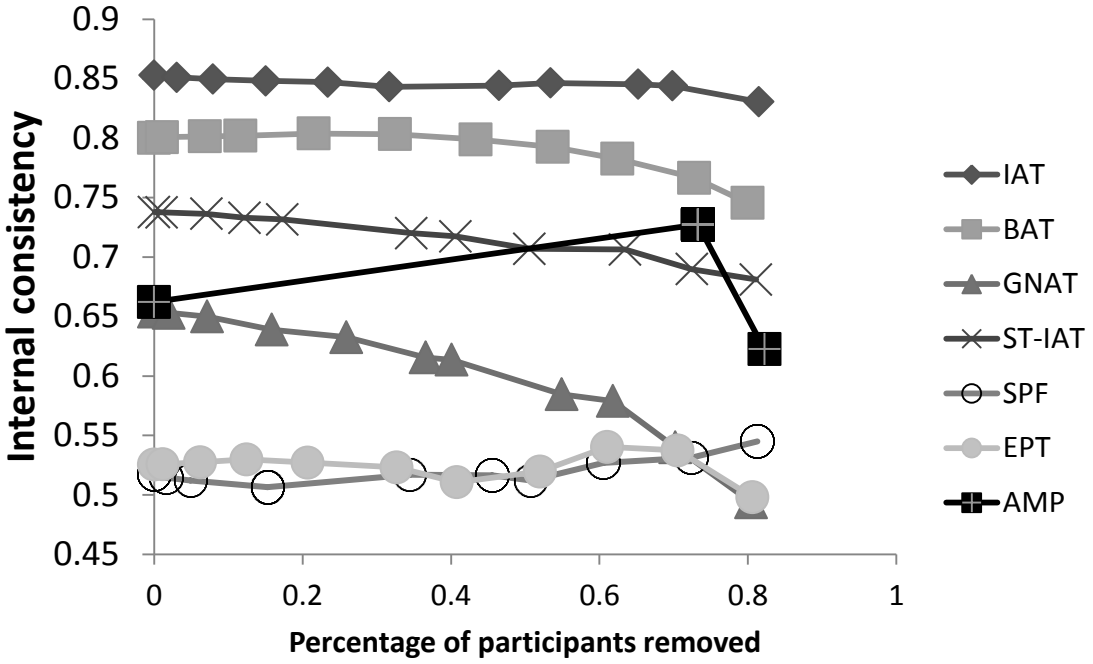


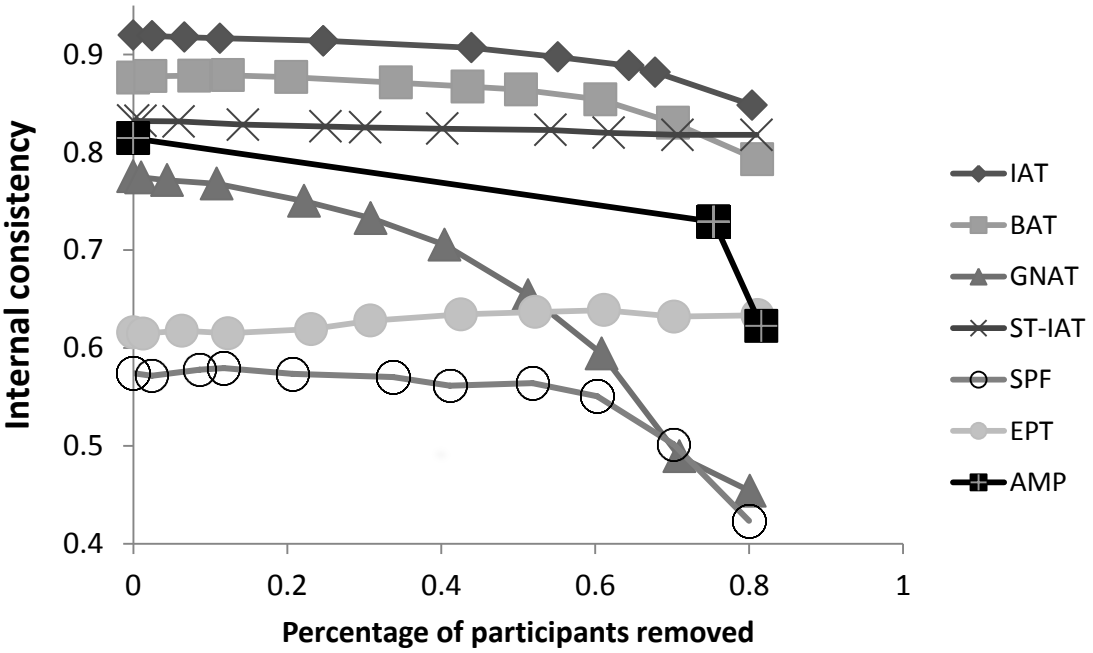
Figure caption: The effect of removing misbehaving participants on the average relationship of the implicit measures with direct measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

Figure W4

A


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
Internal consistency
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B


0.9
0.8
0.7
0.6
0.5
0.4
Internal consistency
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

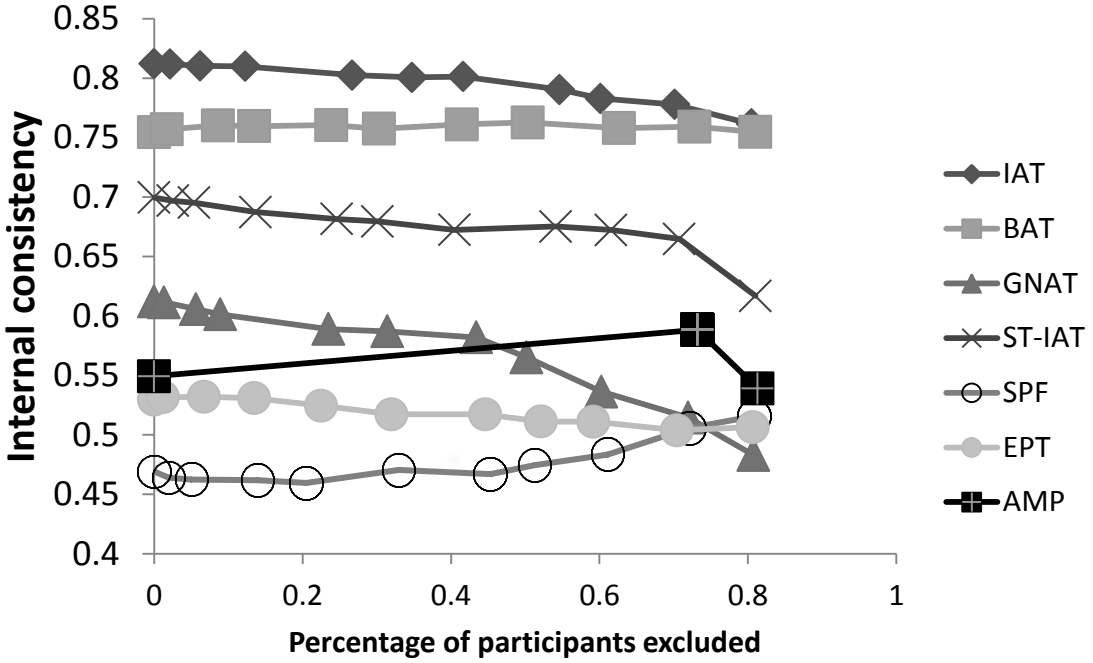
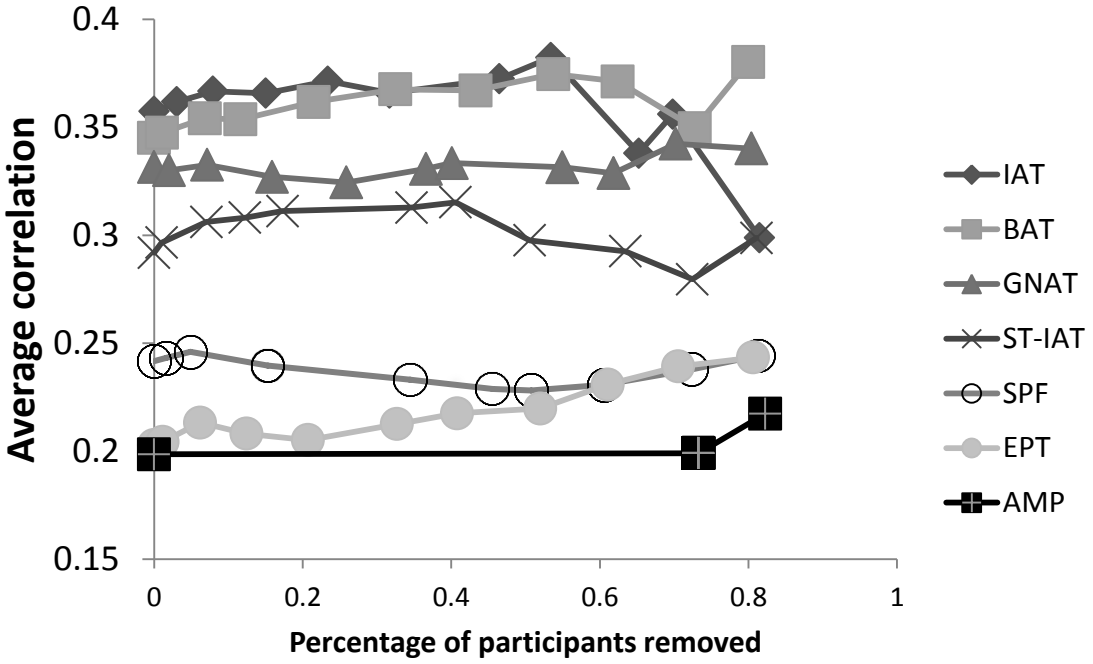


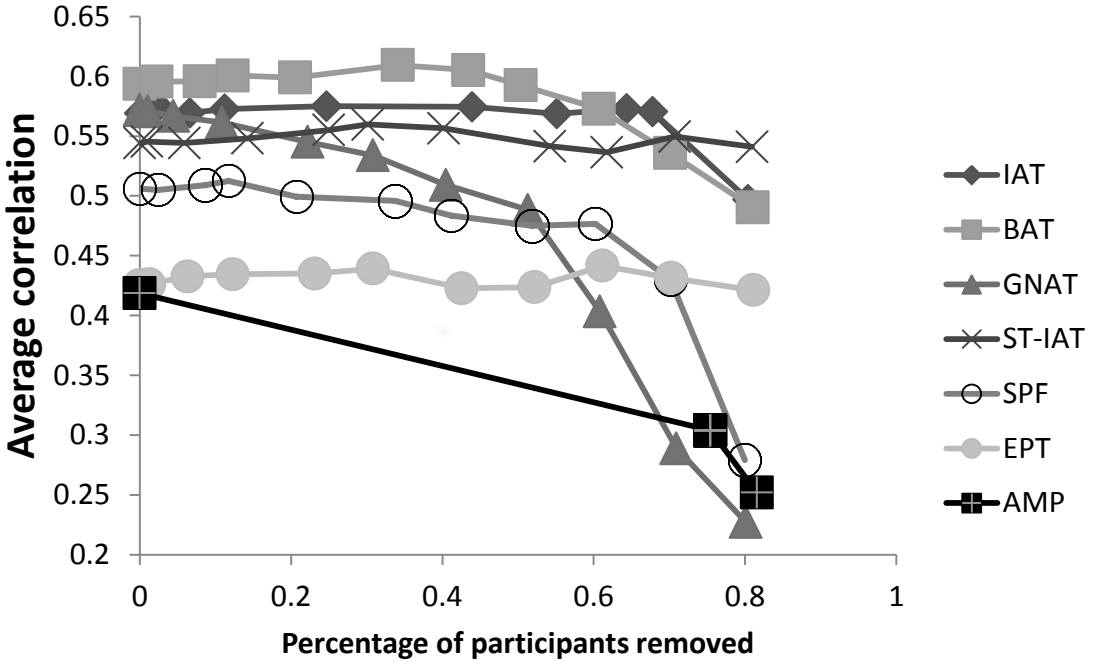
Figure caption. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the internal consistency. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

Figure W5

A


0.4
0.35
0.3
0.25
0.2
0.15
Average correlation
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B


0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
Average correlation
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

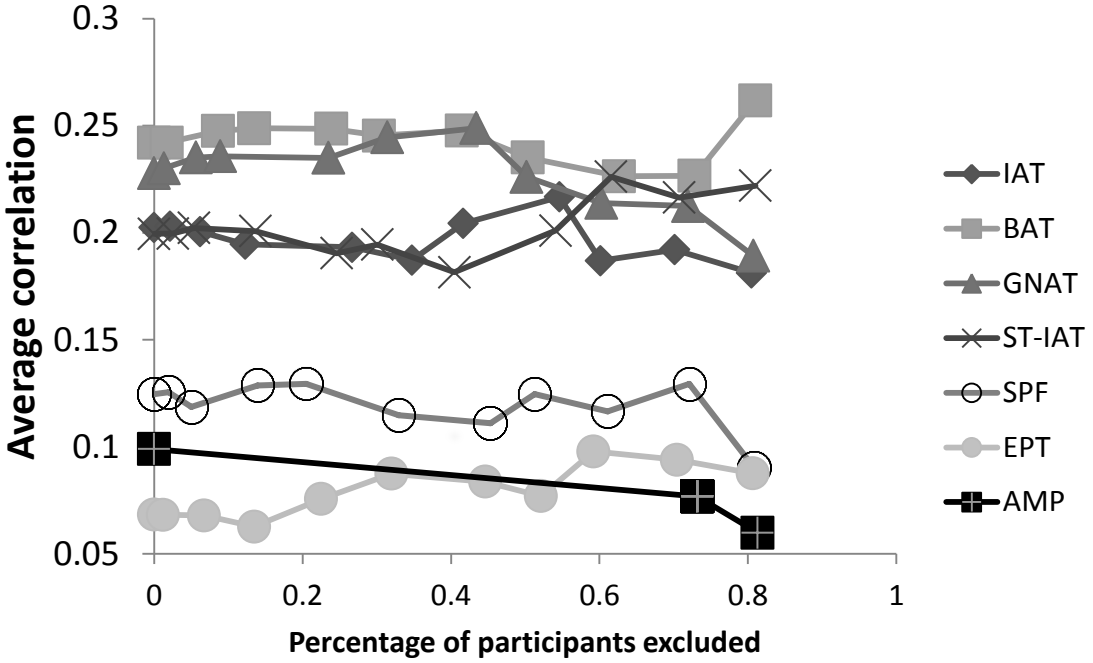
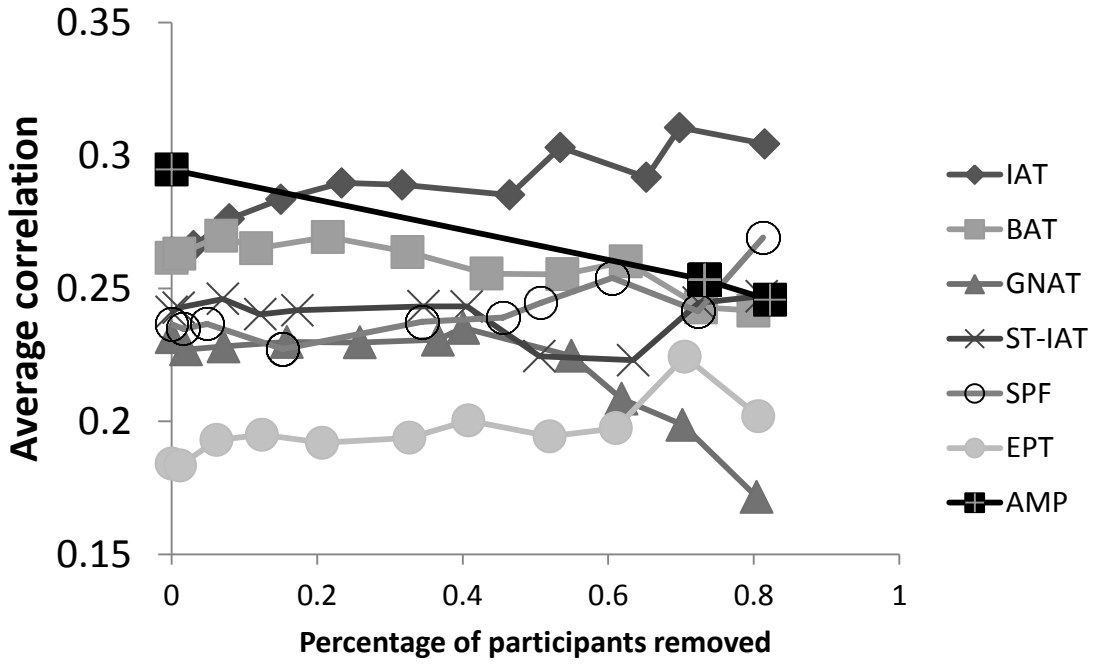


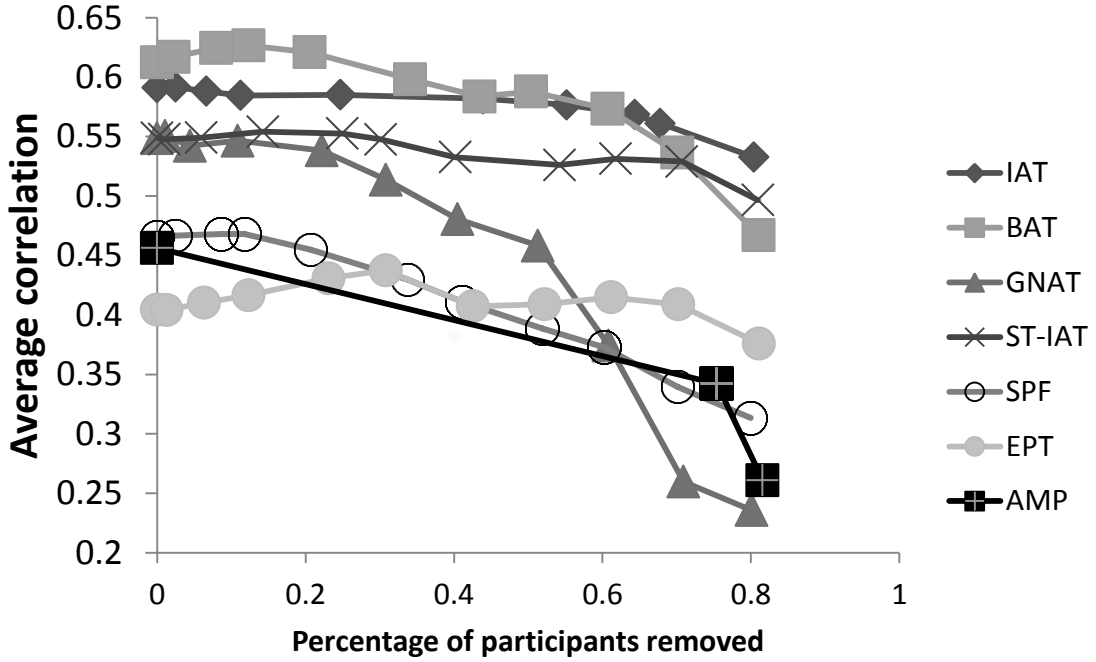
Figure caption. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with implicit measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

Figure W6

A


0.35
0.3
0.25
0.2
0.15
Average correlation
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

B


0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
Average correlation
0
0.2
0.4
0.6
0.8
1
Percentage of participants removed
IAT
BAT
GNAT
ST-IAT
SPF
EPT
AMP

C

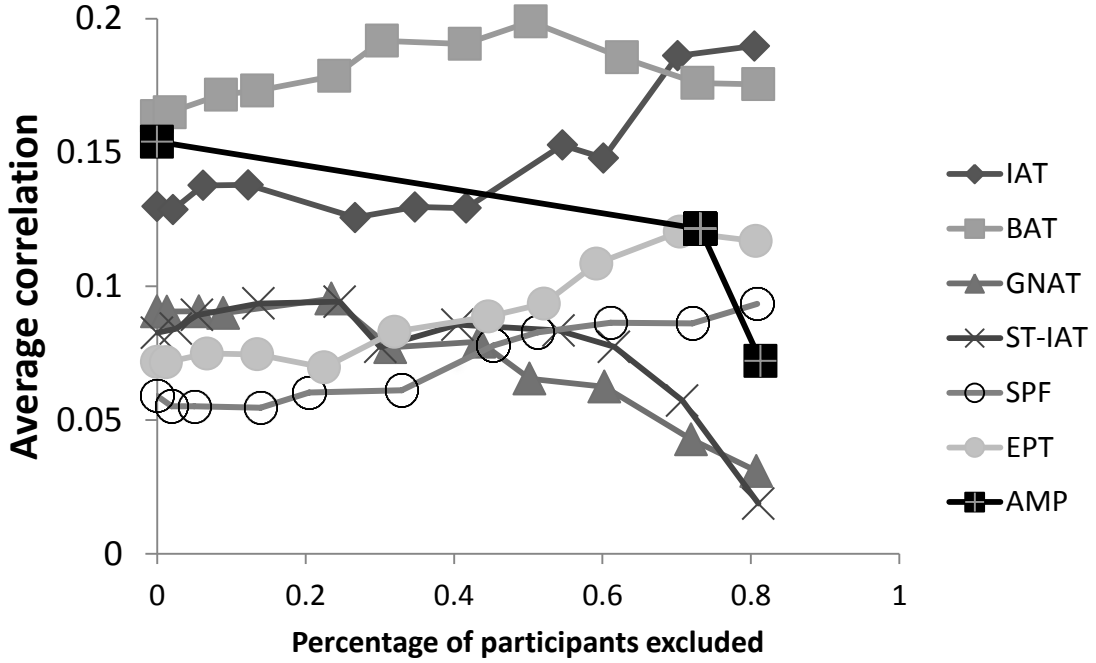


Figure caption. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with direct measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.