A Comparative Investigation of Seven Indirect Attitude Measures

Yoav Bar-Anan

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

Brian A. Nosek

University of Virginia

Authors note: Correspondence should be addressed to: Yoav Bar-Anan, Psychology Department, Ben Gurion University of the Negev, Be'er Sheva, Israel. E-mail: baranany@ bgu.ac.il.

This project was supported by grants from the European Union (PIRG06-GA-2009-256467) and the Israeli Science Foundation (1012/10) to YB-A.

Abstract

We compared the psychometric qualities of seven indirect attitude measures across three attitude domains (race, politics and self-esteem) with a large sample (n = 23,413). We compared the measures on internal consistency, sensitivity to known effects, relationship with indirect and direct measures of the same topic, reliability and validity of single-category attitude measurement, ability to detect meaningful variance among people with non-extreme attitudes, and robustness to the exclusion of misbehaved or well-behaved participants. All seven indirect measures correlated with each other, and with direct measures of the same topic. These relations were always weak for self-esteem, moderate for race and strong for politics. This pattern suggests that some of the source of variation in reliability and predictive validity of indirect measures is a function of the concepts rather than the methods. The Implicit Association Test (IAT) and Brief IAT (BIAT) showed the best overall psychometric quality, followed by the Go-No go Association Task, Single-target IAT (ST-IAT), Affective Misattribution Procedure (AMP), Sorting Paired Features task, and Evaluative Priming. The AMP showed a steep decline in its psychometric qualities when people with extreme attitude scores were removed. Single-category attitude scores computed for the IAT and BIAT showed good relationship with other attitude measures, but no evidence of discriminant validity between paired categories. The other measures, especially the AMP and ST-IAT, showed better evidence for discriminant validity. The results inform on the validity of measures as attitude assessments, but do not speak to the implicitness of the measured constructs.

Keywords: Implicit Social Cognition; Indirect Measures; Implicit Attitudes; The Brief Implicit Association Test

A comparative investigation of seven indirect attitude measures

The emergence of implicit social cognition in the last three decades 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). These measures share a signature feature of assessing social cognitions indirectly wherein the behavioral response does not require the participant to report those cognitions directly. The cognition 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 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. Existing theory and evidence suggest that indirect attitude measures are more sensitive to automatic evaluation, whereas direct attitude measures are more sensitive to deliberate evaluation (Gawronski & De Houwer, in press; Gawronski & Payne, 2010).

A substantial research literature using these indirect 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 indirect 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 Gawronski & De Houwer, in press; Gawronski & Payne, 2010). Despite increasing diversity of indirect 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 indirect measures. There are relatively few studies that use multiple indirect measures in the same study and experimental setting thus creating a vacuum of comparative knowledge between indirect measures.

In this article, we report the results of a large investigation of a variety of psychometric properties of seven indirect measures of evaluation and self-concept. 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). For most of the measures individually, 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 indirect measures and with as large a participant sample enabling precise estimation. The present investigation allowed a direct comparison of the psychometric properties of the seven indirect measures with the same sample, same setting, and same criterion variables across three different content domains – race, politics, and self-esteem. The investigation also provides evidence regarding the variation of the relations among different indirect measures across different content domains.

Besides adding to the psychometric evaluation of individual measures and comparing between these psychometric qualities across measures, a key contribution of this article is inter-relations assessment. Little is known about the relations among indirect measures. The fact that two measures are indirect 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 indirect measures of 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 indirect measures have demonstrated stronger interrelations among the two or three measures investigated (Bar-Anan, Nosek, & Vianello, 2009; Karpinski & Steinman, 2006; Ranganath, Smith, & Nosek, 2008). In the present research we provide evidence whether (and which) indirect 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 indirect measures. Based on previous research and theory (Nosek, 2005; Bosson et al., 2000), we predicted that politics would elicit the strongest indirect-direct relations, and that self-esteem would show the weakest indirect-direct relations. However, there was no consistent pattern of results in previous research that allowed a strong prediction regarding variations in the relations among indirect measures. Nosek (2005, 2007) interpreted the effect of attitude domain on indirect-direct relations as revealing insights about the interplay between implicit and explicit cognition. However, if the same variation would be found for relations among indirect measures,

then the prior findings might apply to relations among attitude measures in general, rather than implicit-explicit relations.

One reason for the paucity of comparative research on indirect measures is likely practical – it requires substantial resources to conduct such studies. Each indirect measure requires a non-trivial amount of time to administer and is mentally taxing to complete. Further, large participant samples are necessary to obtain reliable estimates for comparing across many measures and topics. These constraints may be prohibitive for ordinary laboratory resources. We addressed these practical challenges via a public website that attracts a high volume of participants. To avoid over-taxing individual participants we settled on a planned incomplete design. More than 24,000 participants each completed a random subset of the available indirect measures. This allowed for comparison among all measures despite the fact that any given participant completed only a few of the possible measures.

Evaluation Criteria

Internal consistency. It is desirable for a measure to have little random error during task performance to elicit strong internal consistency. Without strong internal consistency, conclusions concerning individual scores are undermined. 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, if the measure also shows stronger validity, then it is more likely that the strong internal consistency is due to effective construct measurement rather than extraneous influences.

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. 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 indirect measure of the same topic was 116. 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. However, even the short time scale has some information value—it may reflect the stability of the measure in repeated administrations, and after having taken other measures in between the repeating measurements. Therefore, we report test-retest reliability, but do not include it as a primary evaluation criterion.

Sensitivity to group differences. All else being equal, better measures will be 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 when measured indirectly (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 indirectly measured evaluations between Democrats and Republicans for their political parties are observed (Nosek, Smyth, et al., 2007). Because the status of group differences with self-esteem is less clear, we included only racial and political attitudes for evaluation of known-group differences.

Correlation with other indirect 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 indirect measures. This criterion is straightforward, but with an important qualification. Two indirect measures could both be valid but assess different components or qualities of the construct (Olson & Fazio, 2003). 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). The AMP is the only indirect measure that does not use response latency as a dependent variable, the SPF is the only measure that requires responding to two stimuli simultaneously, the SPF and EPT are the only measures that most plausibly influenced by the association between the two stimuli in each trial (and not only associations between categories), 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. However, if two measures (e.g., SPF and EPT, or AMP and EPT) share unique methodological features, then they should relate more strongly with each other than they do with the other measures. If the unique methodological features of a measure are not shared with any of the other measures, then the measure might be inferior to the other measures on this criterion but not necessarily on the other criteria in this study.

Correlation with direct measures of same topic and other criterion variables. To the extent that there is a meaningful relationship between direct and indirect measures of the same topic, better measures will be more sensitive to detecting it. Evaluation models (e.g., Gawronski & Bodenhausen, 2006; Fazio, 2007) and existing psychometric evidence (e.g., Nosek & Smyth, 2007) suggests that indirect and direct (self-report) measures assess distinct, but related constructs. As such, 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 indirect and direct measures are exclusive of one another – that is entirely unrelated intra- and inter-individually.

Nonetheless, there is an important challenge with using direct measures as an evaluation criterion across multiple indirect measures. Each measure has a unique procedure and may engage distinct psychological processes. As a consequence, it is possible that the indirect measures vary in the extent to which they are influenced by deliberate evaluation. So, on its own, variation in correlations with direct measures is ambiguous as a criterion. However, if an indirect measure is actually a direct measure in disguise, then the stronger correlations with direct measures could be accompanied by weaker correlations with other indirect measures. If, on the other hand, the indirect measure is simply a more effective measure, then the stronger correlation with direct measures will likewise be accompanied by stronger correlations with the other indirect measures than they have amongst themselves.

Measurement of single-category evaluation. The present research focused on the indirect measurement of preferences between two categories, rather than evaluation of each category separately. However, some of the measures are designed to allow measurement of a single-category evaluation (the ST-IAT, AMP, and EPT). Additionally, it is possible to compute single category scores with the measures that are relative by design (IAT, BIAT, GNAT, and SPF; though this computational strategy does not guarantee that the assessment is valid, Nosek, Greenwald, & Banaji, 2005). A measure that can validly discriminate evaluations between distinct social categories is useful for measurement flexibility because it extends the potential application of the measure. We compared the reliability, convergent validity and discriminate validity of the single-category evaluation scores of each measure.

Sensitivity to non-extreme attitudes. It is generally easier for measures to detect large differences than small differences. However, a more sensitive measure can detect meaningful differences across the range of possible scores. For example, a measure could be effective at distinguishing extreme political partisans, but fail to distinguish between people that lean to the political left or right. In this case, the measure's psychometric performance will be reliant on the presence of extreme scores and fail when those are removed. Following this rationale, we tested how well the measures retained their psychometric qualities even without extreme scores.

Effects of data exclusion. Respondents must follow task instructions or else interpretation of the assessment may be compromised. We expect the psychometric qualities of to improve when removing participants 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) biasing the sample if exclusion is more likely among some participants more than others (e.g., high versus low intelligence or conscientiousness).

Method

Participants

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. It was possible to be randomly assigned to this study more than once (up to 32 times).

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 indirect 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).

Indirect Measures

All the procedures of the indirect measures were tested prior to the study with the stimuli that were selected for this 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). Table 1 summarizes key features of the measures and the particular procedures used. The supplemental materials provide full details. All tasks– exactly as they were

Table 1

Summary of procedural features of the indirect measure tasks

Measure	# critical trials	Contrast Categories	Latency based	Response Deadline	Categories Labeled	Task on Evaluative Stimuli
IAT	120	+	+	-	+	Categorize
BIAT	128	+	+	-	+	Categorize
GNAT	160	+	+	+	+	Categorize
ST-IAT	192	-	+	-	+	Categorize
SPF	120	+	+	-	+	Categorize
EPT	180	+	+	+	-	Memorize
AMP	48	+	-	-	-	Ignore

Notes. The 48 trials of the AMP do not include the additional 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 into the score).
administered in the study - can be experienced at:

http://opensciencefframework.org/project/Qf9jX/node/YJQiq/.

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 practice block, participants categorized items representing the two attitude objects (e.g., Democrats vs. Republicans). In the second block, participants categorized good and bad words. The third block 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 was the same as the third block. Block 5 was like block 1, but the attitude objects switched sides (i.e., the object that was categorized with the left key in block 1-3 was now categorized with the right key). Blocks 6 and 7 combined blocks 2 and 5.

Brief Implicit Association Test. The BIAT procedure followed the one described in Sriram and Greenwald (2009), but with a different block sequence. Each block in the BIAT is like a combined block in the IAT, but instead of four categories, only the two categories that would appear on the right side of the IAT screen appear on screen. Participants sort items that belong to these categories with the right key, and hit the left key for any item that does not belong to these categories (these items always belong to the two non-focal categories).

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 is like the BIAT, but when the target item belongs to the categories on the screen, participant must

hit the space key before a response deadline. For other items that belong to the other categories, participants must wait without hitting any keys.

Single-Target Implicit Association Test. The ST-IAT is similar to the IAT, but instead of two attitude object categories, only one attitude object is presented. That category shares a key with Good words in one block, and with Bad words in the next block. Participants completed four blocks with one attitude object (e.g., Democrats), and then four blocks with the other object (e.g., Republicans).

Sorting Paired Features. The SPF procedure followed the one described in Bar-Anan et al. (2009). In each trial, participant sort item pairs into category pairs appearing in the four screen corners. The category pairs are 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).

Evaluative Priming Task. The procedure followed the one described by Fazio et al. (1995). In the first block, participants categorized words as "Good" or "Bad." In the next three blocks, participants continued with the same sorting, but a prime item appeared before each word. The prime items were from the attitude object categories. Participants were instructed to memorize the prime items for a memory test, and categories the words.

Affective Misattribution Procedure. The procedure followed the one described by Payne et al (2005). In each trial, a prime item was presented briefly, followed by the target, a Chinese letter, and then a mask. Participants were instructed to rate the target as more pleasant than the average Chinese symbol, or more unpleasant. They were instructed not to let the prime item influence their evaluation of the target stimulus.

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 indirect 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. The full details are provided in the online supplemental materials.

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) with all scale points labeled. 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 included three novel measures of self-esteem 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 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. Otherwise, there was no constraint on repetition of topic or method in the same session. For example, all four measures could measure race; or two measures could measure self-esteem, the third race attitudes, and the fourth political attitudes. 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,

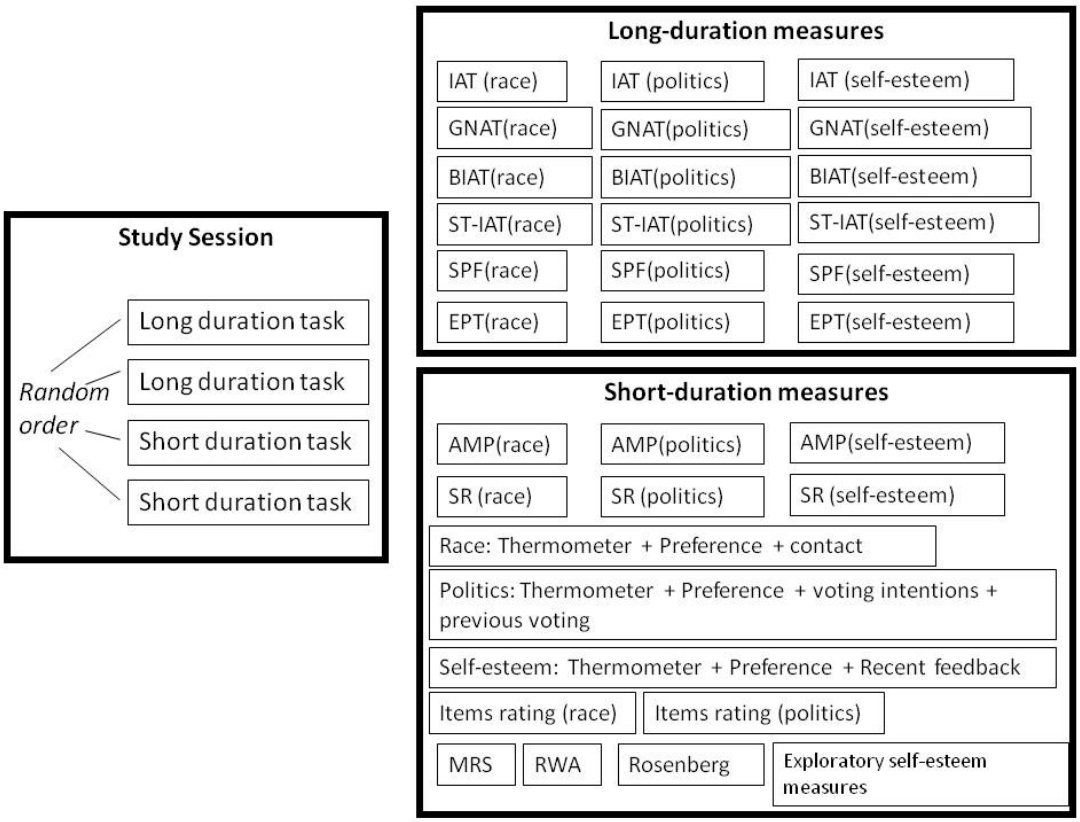


Figure 1. Illustration of the study procedure. Each study session included 2 long and short tasks, selected and ordered randomly. The tasks are listed on the right. Measures that share a rectangle were presented together in the same questionnaire.

the purpose of the study was explained to the participants, and they received result feedback for the indirect measures they performed.

Results

Given the large samples, the emphasis in this report is on effect size rather than significance testing.

Data Processing

We detail the data processing procedures and the scoring of each task in the supplemental materials. 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.

Mean preference

All the indirect measures except the AMP (d = -0.23) indicated a preference for White people compared to Black people (Table 2 presents effect sizes, and Table C1 in the online supplemental materials adds 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, with race attitudes, perhaps 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.

Table 2
A Summary of the Results

	Reliability		**Correlations**		**Extreme Scores Exclusion**		
	Internal Consistency	Test-retest	With indirect measures	With direct measures	% shared variance lost after dropping the 10% most extreme attitude scores		
	Average alpha	Average correlation	Average correlation	Average correlation	Internal consistency	Corr. with indirect measures	Corr. direct measures
Overall							
IAT	**.88**	.45	.39	.35	**11.9**	3.8	3.0
BIAT	.83	**.63**	**.41**	**.38**	15.6	4.4	2.8
GNAT	.74	.42	.40	.33	19.5	3.5	3.2
ST-IAT	.77	.48	.36	.31	19.5	*5.4*	4.7
SPF	*.53*	.46	.31	.27	21.3	3.9	2.8
EPT	*.57*	*.33*	*.25*	*.24*	25.0	**3.4**	**2.4**
AMP	.69	.50	.26	.32	*35.6*	3.8	*7.8*

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	**13.5**	4.5	2.4
BIAT	0.73	0.77	.81[b]	.80-.82	**.63[a]**	.50-.73	.34	.27	17.3	*4.8*	1.6
GNAT	0.83	0.67	.70[d]	.69-.73	.30[bc]	.13-.45	.35	.27	21.6	2.8	2.6
ST-IAT	0.20	*0.33*	.74[c]	.72-.76	.34[bc]	.16-.49	.30	.24	21.3	3.6	3.2
SPF	0.24	0.76	*.52*[f]	.49-.55	.38[bc]	.21-.53	.24	.24	23.0	2.6	2.6
EPT	0.07	0.57	.54[f]	.51-.57	*.18*[c]	.00-.36	*.20*	*.19*	24.6	**1.6**	2.0
AMP	-0.23	0.39	.66[e]	.64-.68	.33[bc]	.18-.47	*.21*	**.31**	*42.7*	2.2	*8.2*
Politics											
IAT	0.49	**1.49**	**.93[a]**	.92-.93	.65[abc]	.54-.74	.58	.60	**6.0**	*5.5*	5.2
BIAT	0.63	1.40	.89[b]	.88-.90	**.78[a]**	.69-.85	**.60**	**.63**	9.9	6.5	5.6
GNAT	0.47	1.38	.84[c]	.83-.85	.72[ab]	.63-.79	.59	.59	13.1	**5.5**	6.6
ST-IAT	0.42	1.04	.84[c]	.83-.85	.54[c]	.40-.66	.55	.56	15.6	*11.2*	**1.5**
SPF	0.21	1.08	*.59*[f]	.57-.61	.58[bc]	.44-.70	.52	.48	24.5	7.7	5.6
EPT	0.27	*0.73*	.63[e]	.61-.65	*.51*[c]	.34-.63	.45	*.42*	28.0	8.2	5.9
AMP	0.31	0.88	.81[d]	.80-.82	.73[a]	.65-.79	*.43*	.48	*34.8*	8.1	*13.2*
Self											
IAT	1.31		.82[a]	.81-.83	*.26*[a]	.09-.41	.21	.14	**16.1**	1.3	1.5
BIAT	1.03		.76[b]	.74-.78	**.42**[a]	.25-.57	**.25**	**.18**	19.7	1.8	1.2
GNAT	1.23		.65[d]	.63-.67	.34[a]	.14-.51	.21	.08	23.7	*2.2*	0.4
ST-IAT	0.57		.70[c]	.68-.72	.36[a]	.19-.51	.20	.09	21.6	1.5	0.5
SPF	0.96		*.48*[f]	.45-.51	.39[a]	.23-.53	.14	*.06*	19	1.4	**0.1**
EPT	0.43		.54[e]	.51-.56	.29[a]	.36-.43	*.07*	.08	22.5	**0.3**	**0.1**
AMP	0.16		.55[e]	.53-.57	.35[a]	.20-.48	.10	.16	*29.3*	1.1	*1.9*

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; *Underlined Italic* font = worst performance in the relevant criterion; The sample sizes for the test-retest analyses were between 83 and 158 (average = 116); Within each topic, in the internal consistency and test-retest correlations criteria, identical superscripts indicate no significant difference. The Cronbach alphas were compared using Feldt test (1969); The test-retest values are correlations between the first and the second tests.

All the direct (mean $d = 0.54$) and indirect (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. All the direct (mean $d = 0.41$) and indirect (mean $d = 0.64$) measures indicated a preference for the self over others.

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

Known-Group Differences

All else being equal, better measures will be more sensitive to detecting known group differences. Table 2 summarizes the comparison of Black and White participants for all racial attitude measures (See Table C2 in the supplemental materials 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 at least half the size as the strongest ones (Cohen's d = 0.39 and 0.33, respectively).

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). 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).

In summary, the IAT and the BIAT showed the best sensitivity to detect expected effect of participants' social identity. The GNAT and the SPF were the next most sensitive measures. The ST-IAT, AMP and EPT showed the weakest sensitivity.

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.

Relationships with Other Indirect Measures

Assuming that the indirect attitude measures are valid to some degree, all else being equal, more valid measures will be more strongly related to other indirect measures than less valid measures will be. Of course, this general statement is qualified by the possibilities that (a) subsets of indirect measures assess different components of implicit cognition constructs – each valid but distinct, and (b) subsets of indirect measures share a confounding influence that create spuriously strong relations that are not relevant to the construct. Concern (a) can be addressed by examining the possibility of distinct covariation with other criterion measures, as we pursue in the next section. Concern (b) can be addressed by examining whether there are clusters of strong covariation. An in-depth examination of the structural relations among indirect measures goes beyond the scope of this article, but is taken up in detail with these data by Bar-Anan, Shahar, and Nosek (2013).

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 C3 in the supplemental materials).

Table 3
Correlations among the Indirect Measures

	IAT	BIAT	GNAT	ST-IAT	SPF	EPT	AMP
Avg. N	395	374	365	387	254	393	509
Range	294-	304-	278-	278-	313-	298-	414-

N	558	496	520	548	524	539	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; *Underlined 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 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 indirect measures – IAT and ostensibly-related derivatives as one cluster and the AMP and the EPT as a separate cluster. If so, then the AMP and EPT would be related to each other more strongly than they relate 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). Therefore, the most likely explanation for this pattern, 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 indirect 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 indirect measures.

Correlations with Direct Measures and Other Criterion Variables

Table 4 presents the average relationship of each indirect measure with the direct measures of the same topic and the other criterion variables. For each criterion measure, the correlations of the indirect measures were compared statistically, and also ranked. The aggregated correlations are presented in Table 2 (see Table C3 for the average rankings).

Table 4

Correlations of the Indirect 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[a]	--	--	-0.14
Avg. N	593	612	623	622	608	541
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-size	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-size	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; **Bold** = the strongest correlation of that column; *Underlined Italics* = the weakest correlation of that column; [a] The effect-size of the race items-rating score is negative to indicate preference for black people over white people (opposite to the effect of the other direct measures).

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. The main difference between the performance of the indirect 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. This may suggest that deliberate evaluation influences the AMP more than it influences other measures. Cameron et al. (2012) reviewed evidence against that possibility. Another possibility is that the distinct construct measured by the AMP is related to deliberate evaluation more than to the constructs measured by the all the other indirect measures.

Single-Category Measurement

For the ST-IAT, SPF, AMP and EPT, the computation of the single-category evaluation scores was a part of the preference scores calculation. For the IAT, BIAT and the GNAT we computed the single-category evaluation score by including only trials that required a response with the key that was associated with the category. The online supplemental materials provide more details about the computations.

There is little research about the quality of indirect measures in separate measurement of two attitude objects. Prior research found poor discriminant validity for single attitude measurement with the IAT (Nosek, Greenwald, & Banaji, 2005), good discriminant validity for

the ST-IAT (Karpinski & Steinman, 2006), and possible threats to non-relative single attitude measurement in the EPT and AMP, when multiple attitude objects are included in the same task (Scherer & Lambert, 2009). One of the unique contributions of the present study is that it provides direct test between the measures that are constrained by relative measurement and the measures that, at least theoretically, seem to provide separate measurement for each attitude object.

We tested the single category scores for known group effects, internal consistency, test-retest correlation, and relationship to other indirect measures and direct measures of the same category. In addition, we looked at relationship of the evaluation of Self with the Rosenberg self-esteem scale and of the evaluation of Black people with the MRS, and the evaluation of Republicans with the RWA. According to Karpinski and Steinman (2006), the evaluation of the category Self is more strongly related to self-esteem measures than the Self-Other preference score because the direct measures of self-esteem (including the Rosenberg scale) do not compare the evaluation of the self to the evaluation of others. The same rationale can also be applied to the MRS which focuses on Black people, but not in comparison to White people. The RWA is more focused on conservative evaluations relevant to Republicans than to liberal evaluations relevant to Democrats. In support of these assumptions, we found that the direct measures of the categories Self, Black people and Republicans were more strongly related to the Rosenberg, MRS and RWA, than the direct measures of Other, White people and Democrats, respectively (Table C4 in the online supplement materials).

Finally, and perhaps most important, we looked at the difference between the absolute average correlation of each category score (e.g., indirectly measured White attitude) with the direct measures of the same category (White attitude) and absolute average correlation of that category with the direct measures of the other category (Black attitude). That difference provided

an estimation of discriminant validity: how much the single evaluation score is related to the same category more than to the other category measured in the same topic. That is, does the single category score measures attitudes toward the single category or does it remain constrained to relative assessment between the categories (Nosek et al., 2005)? The former was true for direct measures: Table C4 shows that the thermometer rating of each separate category was related more strongly to the direct rating of the items (or speeded rating in the case of the self-esteem pair) of the same category than to the direct rating of the items of the other category.

Table 5 presents the summary of the single-category measurement criteria (see Table C5 in the online supplemental materials for more details). We found that the AMP showed the best reliability and discriminant validity, whereas the IAT and the BIAT showed the best convergent validity. The IAT and the BIAT were also the only measures that showed no sign of discriminant validity. The ST-IAT showed reliability that was not far behind the AMP and the IAT, convergent validity that was better than the AMP and often not far behind the IAT, and discriminant validity that was much better than the IAT, and only slightly worse than the AMP. The GNAT showed internal consistency weaker than the ST-IAT's, but its convergent validity was always better than the ST-IAT's, and its discriminant validity was only slightly weaker than the ST-IAT's. The SPF and EPT were weak on most criteria.

Table 5
Summary of Single-category Measurement Criteria

	Reliability		**Convergent Validity**							**Discriminant Validity**
			Known-groups		Correlation with other measures					
Overall	*Alpha Cronbach*	*Test-Retest*	*Race*	*Politics*	*With indirect*	*With direct*	*With Rosenberg*	*With MRS*	With RWA	
IAT	.77	.41	**1.0**	**1.37**	**.29**	**.26**	**.18**	-.25	.43	*0*

			3							
BIAT			0.6				.17	**-.27**		
	.67	.53	5	1.27	**.29**	**.26**			**.53**	*0*
GNAT			0.4				.15	-.26		
	.64	*.29*	4	1.06	.27	.24			.41	.06
ST-IAT			0.2				.14	-.25		
	.76	.30	5	0.77	.23	.22			.31	.07
SPF			0.5				*0*	-.21		
	.44	.34	4	0.82	.20	.17			.29	.04
EPT			0.3				*0*	-.21		
	.63	.32	6	*0.50*	.16	*.15*			.27	.05
AMP			*0.1*				.13	*-.18*		
	.82	**.59**	*8*	*0.50*	*.12*	.21			*.25*	**.09**

Notes. For the convergent validity, the correlation was with measures of the same category. The correlation with Rosenberg's scale was the correlation of the evaluation of "Self." The correlation with MRS was the correlation of the evaluation of "Black People." The correlation with RWA was the correlation of the evaluation of "Republicans." The discriminant validity is the average difference between the absolute correlation with each direct measure of the same category and the absolute correlation with the same direct measure of the opposite category; **Bold** = the strongest correlation of that column; *Underlined Italics* = the weakest correlation of that column;

We did not find advantage for the AMP and the ST-IAT in predicting scales that are related more strongly to one of the categories in each topic than the other. For instance, Rosenberg was not related to the evaluation score of Self as measured by the AMP (r = .13) or the ST-IAT (r = .14) more than to the measurement of the Self category by the IAT (r = .18) or the BIAT (r = .17). So, while the AMP and ST-IAT show stronger discriminant validity in providing separable assessments of Blacks and Whites (and politics and self-esteem), their weaker overall psychometric performance resulted in them still showing less convergent validity than the IAT and BIAT in predicting single attitude criterion variables.

In summary, the AMP, ST-IAT, and GNAT showed good signs of single evaluation measurement qualities with a superior discriminant validity and fair reliability and convergent validity. The IAT and the BIAT's good convergent validity suggest that the superior discriminant validity of the other measures does not guarantee an advantage in convergent validity. An

anonymous reviewer suggested that in the IAT and the BIAT, each category provided a context to interpret the meaning of the other category (e.g., *White people* provides context for the category *Black people*). Similarly, the direct evaluations of each single category in our study might have been influenced by the context created when rating the two topic categories in temporal proximity (e.g., rating Black people right after White people). In that case, measures that induce a similar context by contrasting the two categories would be more strongly related to the separate direct evaluation than measures that do not induce that context (Perugini, Richetin, & Zogmaister, 2010). At the same time, if some features of each attitude object—unrelated to the contrast context (e.g., liking the word *other* because it sounds nice)—had even a small effect on the evaluation of a target category, then indirect measures that do not emphasize the contrastive context might show better discriminant validity.

Sensitivity to Non-extreme Attitudes

Participants with extreme attitudes may contribute to the psychometric qualities of measures more than participants with moderate attitudes because most of the psychometric qualities depend on variability. But meaningful individual differences are not only in the extremes. As such, detecting differences between people with moderate attitudes is a positive psychometric quality.

Table 6

The Influence of Excluding Extreme Scores on the Psychometric Qualities of the Measures

	Internal consistency			**Average correlation with indirect measures**			**Average correlation with direct measures**		
	All cases	*The middle 90%*		*All cases*	*The middle 90%*		*All cases*	*The middle 90%*	
	Alpha	Alpha	% loss	R	R	% loss	R	R	% loss
Over all									
IAT	**.88**	**.81**	**11.9**	.39	**.34**	3.8	.35	.30	3.0
BIAT	.83	.73	15.6	**.41**	**.34**	4.4	**.38**	**.34**	2.8
GNA	.77	.59	19.5	.40	**.34**	3.5	.33	.29	3.2

T									
ST-IAT	.74	.62	19.5	.36	.26	*5.4*	.31	.23	4.7
SPF	*.53*	.26	21.3	.31	.23	3.9	.27	.22	2.8
EPT	.57	.25	25.0	*.25*	.18	**3.4**	*.23*	*.18*	**2.4**
AMP	.69	*.21*	*35.6*	.26	*.16*	3.8	.32	*.18*	*7.8*
Race									
IAT	**.86**	**.78**	**13.5**	**.36**	.29	4.5	.27	.22	2.4
BIAT	.81	.70	17.3	.34	.26	*4.8*	.27	**.24**	**1.6**
GNAT	.71	.53	21.6	.35	**.30**	2.8	.27	.22	2.6
ST-IAT	.74	.58	21.3	.30	.21	3.6	.24	.15	3.2
SPF	*.52*	.26	23.0	.24	.17	2.6	.24	.18	2.6
EPT	.54	.21	24.6	*.20*	.15	**1.6**	*.19*	*.13*	2.0
AMP	.66	*.10*	*42.7*	*.21*	*.14*	2.2	**.31**	.13	*8.2*
Politics									
IAT	**.93**	**.90**	**6.0**	.58	.52	**5.5**	.60	.56	5.2
BIAT	.89	.83	9.9	**.60**	**.53**	6.5	**.63**	**.58**	5.6
GNAT	.84	.76	13.1	.59	**.53**	**5.5**	.59	.54	6.6
ST-IAT	.84	.74	15.6	.55	.42	*11.2*	.56	.46	**1.5**
SPF	*.59*	*.32*	24.5	.52	.42	7.7	.48	.42	5.6
EPT	.63	.34	28.0	.45	.33	8.2	*.42*	.34	5.9
AMP	.81	.56	*34.8*	*.43*	*.31*	8.1	.48	*.31*	*13.2*
Self									
IAT	**.82**	**.72**	**16.1**	.21	.17	1.3	.14	.06	1.5
BIAT	.76	.62	19.7	**.25**	**.21**	1.8	**.18**	**.14**	1.2
GNAT	.65	.43	23.7	.21	.16	*2.2*	.08	.06	0.4
ST-IAT	.65	.52	21.6	.20	.15	1.5	.09	*.05*	0.5
SPF	*.48*	.19	19	.14	.08	1.4	*.06*	*.05*	**0.1**
EPT	.54	.26	22.5	*.07*	.05	**0.3**	.08	.08	**0.1**
AMP	.55	*-.09*	*29.3*	.10	*.03*	1.1	.16	.09	*1.9*

Notes. The average % loss is the average loss of shared variance; **Bold** = the strongest correlation of that column; *Underlined Italics* = the weakest correlation of that column;

To examine this psychometric quality, we removed the 10% most extreme scores (regardless of whether the score was above or below the average score). As detailed in Table 6, the AMP suffered the most from trimming the extremes. After trimming, the AMP dropped to the last place in all three main criteria: internal consistency, relationship with indirect measures, and relationship with direct measures. The race AMP in isolation illustrates this effect. Without the 10% most extreme cases, the internal consistency of the race AMP decreased from $\alpha = .66$ to $\alpha = .10$, and the average correlation between the AMP and the direct race measures declined from $r = .31$ to $r = .13$. Compare that with the race IAT's psychometric resistance to trimming the extreme scores: a small decrease from $\alpha = .86$ to $\alpha = .78$ in internal consistency, and from $r = .27$ to $r = .22$ in average correlation with direct measures. The supplemental materials display plot figures that illustrate the deterioration in specific psychometric qualities for each of the measures as a function of the percentage of extreme score trimming. The plots show that the results presented here are a general trend for each measure, and not specific for a 10% cut-off.

After the AMP, the ST-IAT was most sensitive to the loss of extreme scores. The IAT was most resistant to sample trimming, followed by the BIAT, GNAT and SPF. The EPT usually showed small loss, but even that small loss was usually enough to keep its place as the worst, or the second-worst measure on each criterion.

Sensitivity to Data Exclusion due to Unusual Behavior

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. The supplemental materials detail our analyses of the effect of removing participants who showed evidence of misbehavior on the psychometric qualities of each measure. In short, 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. 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.

General Discussion

The present research compared the psychometric qualities of seven indirect attitude 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 indirect measures of the same topic, relations with direct measures of the same topic, relations with other criterion variables, psychometric qualities of single category measurement, ability to detect meaningful variance among people with non-extreme attitudes, and robustness to the exclusion of misbehaved or well-behaved participants. The data provide evidence about the psychometric qualities of individual indirect measures, comparative knowledge of psychometric qualities, practical information for the selection of measures for research application, and general knowledge about indirect measurement.

The Validity of Indirect Measures

The present study provides support for existing claims about indirect measurement that previously have been based on evidence from just one or two indirect measures. All seven indirect 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 indirect measures of the same topic, (c) related to direct, explicit measures of the same topic, and (d) predicted criterion variables related to the topic. The

evidence leaves no doubt that indirect measures are valid assessments of social cognition, affirming their usefulness for research applications.

Most attitude research that has made use of indirect attitude measurement—either to increase predictive validity of attitudes (complementing direct measures), or as a separate measure to assess non-explicit evaluation—employed only one indirect measure. Standard practice is to interpret the results of any indirect measure as assessment of the same latent construct (e.g., implicit attitudes). One threat to this practice is the evidence that indirect measures sometimes have weak or no relationship amongst themselves (e.g., Bosson et al., 2000; Olson & Fazio, 2003; Payne et al., 2008). The present study found moderate to strong relationships among seven indirect measures in at least two topics (politics and race) and poor relationships between the measures in one topic (self-esteem). Given the rarity of research that has examined inter-relations between indirect measures, the strength and breadth of the present findings provide confidence that indirect attitude measures are inter-related, but that this relation varies across attitude domains. This finding reduces the concern raised by studies that failed to find inter-relations.

Variations across Attitude Domains

The present study found that the variation in indirect-*indirect* relations was concordant with the variation in indirect-*direct* relations (Table 2). Relations among indirect measures were strongest for political attitudes and weakest for self-esteem, just as they were between indirect and direct measures of those topics. 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 direct measures relations with one another (Table C6 in the online supplemental materials), and indirect measures relations with themselves (internal consistency; Table 2). Because reliability limits validity, it is

possible that the effect of topic on internal consistency is the reason for the same pattern found with measures inter-relations. Until further evidence, we can only speculate that the concept *Self* is more multi-faceted and less clear than race concepts, and that politics is the clearest. However, an exact definition of this *concept clarity* variable and further evidence to support this speculation would require further research. This presents an opportunity for theoretical generativity.

Do Indirect Measures Measure Implicit Social Cognition?

The similar effect of attitude topic on interrelations among direct measures, among indirect measures, and between direct and indirect measures casts doubt on the perspective that indirect and direct measures tap distinct constructs (implicit versus explicit social cognition). For instance, a central assumption in contemporary attitude research is that self-presentation motivation influences the relation between direct and indirect measures (e.g., Fazio, 2007; Nosek, 2005). That seems a likely account why people show stronger direct-indirect relations regarding politics than race. However, in the present study, relations between measures were weaker for race than for politics even among indirect measures. Another finding from the present research that may not fit well with the common view that indirect measures of social cognition tap different constructs than direct measures is that indirect-indirect relations were not substantially stronger than indirect-direct relations. Although inter-relations among direct measures were stronger than their interrelations with indirect measures, this may be attributed to the lower reliability of indirect measures, and not to sensitivity to different constructs or processes. Therefore, the present results do not provide any support to the assumption that indirect and direct measures of social cognition are sensitive to different theoretical constructs or different psychological processes.

Because much previous research supported the assumption that indirect measures (more than direct measures) tap into implicit cognitions (e.g., Cameron et al., 2012; Greenwald et al., 2009), and because the correlations between indirect measures and other measures in this study were usually only moderate—we hesitate to treat our results as strong evidence that direct and indirect measures tap a single construct. Rather, the inter-relations correlations might reflect the lower reliability of most indirect measures (in comparison to direct measures) that prevent strong correlation among indirect measures. Alternatively, the present results might reflect variability in methodological or theoretical sources of variance that influence indirect measures. We address this issue further in a separate investigation (Bar-Anan, Shahar, & Nosek, 2013) that examines the mapping of the measurement outcomes of the various direct and indirect measures into a small number of theoretical constructs (latent variables).

Comparative Conclusions

Of the seven indirect 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 preference measurement, for each of the three topics (total of 23 because self-esteem did not have a known-groups difference criterion). The IAT was the best measure on ten of these criteria and the BIAT was the best on eight of these criteria. One of these two measures was the second best on 11 of the criteria. The average ranking of the BIAT in the 23 criteria was 2.35, and the average ranking of the IAT was 2.39. Next were GNAT (3.74), ST-IAT (4.26), SPF (4.39), AMP (5.04) and EPT (5.30).

On the other end, EPT had the worst psychometric qualities, and the SPF and AMP were not much better. Of these, the AMP's relatively weak psychometric qualities were the most surprising. In particular, removing the 10% most extreme scores 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.

On one of the psychometric criteria—relationship with other indirect measures—the present design might have put the IAT, BIAT, GNAT, and ST-IAT in advantage over the other measures because these measure may share procedures that seem similar. However, the procedures of the AMP and the EPT seem more similar to each other than to the IAT, BIAT, GNAT and ST-IAT—and yet the average correlation of the IAT, BIAT, and GNAT with the EPT and with the AMP was stronger than the average correlation between EPT and AMP. In addition, these three measures were often superior to the AMP and EPT in other criteria.

Another possibility is that the relatively poorer performance of the AMP and EPT was caused by the specific stimuli chosen for the present study. We have not tested whether the stimuli in our study were representative examples of their categories, nor did we try to balance them on any objective criteria (e.g., facial expression). It is possible that poor selection of stimuli could cause more damage to measures that are more sensitive to the items than the categories (i.e., the AMP and EPT). Therefore, a follow-up study (Bar-Anan & Nosek, 2013) used stimuli selected especially for the AMP and compared them experimentally with the stimuli from the main study. The follow-up study also added trials to the AMP, using 120 instead of 48 trials.

When we computed the AMP's preference score with the first 48 trials (like in the main study), we found that the stimuli set influenced the average preference score of the AMP and the EPT, but had no significant effect on the psychometric qualities of any of the four measures. These results suggest that the stimuli set has no impact on the most important psychometric evaluation criteria for the indirect measures.

Importantly, when we computed the AMP's score with 120 trials, the AMP's psychometric qualities improved substantially to be similar to the best performing measures.

Like in the present study, however, the exclusion of the 10% most extreme scores in the follow-up study damaged the AMP's psychometric qualities (average decline 20.4%) more than it damaged the other measures (average decline 6.7%). Nevertheless, in the follow-up study the AMP's psychometric qualities were still acceptable, even without the 10% most extreme scores. This indicates that most of the AMP's poor results in the present study can be improved by adding trials beyond the numbers used in most existing applications of the AMP.

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 SPF and the AMP) often showed acceptable psychometric qualities, relative to what has been found with other indirect 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 extreme scores, or to the exclusion of well-behaved or misbehaved participants.

The present research also tested some psychometric qualities of single category measurement. One known disadvantage of the IAT is that it measures preference and not a single category evaluation. Other measures, especially the ST-IAT, that present only one attitude object in each block, seem more suitable for single category evaluation. However, the single category evaluation scores computed from the IAT were not inferior to any other measures in predicting single category evaluation, or in predicting scales that were supposed to relate to one category evaluation stronger than to the other. The IAT (and the BIAT) proved inferior only when we looked at the difference between the relationship of each single category evaluation score and the direct evaluation measurement the same versus the other category (for each attitude domain). Evaluation scores computed from the IAT for one category (e.g., Black people) were not related to direct evaluation of that category measures more than to the direct evaluation of the other

category (e.g., White people). The BIAT showed the same poor discriminant validity. The other measures, especially the AMP and the ST-IAT showed some discriminant validity, suggesting that these measures might be better in discriminating between the evaluations of different categories, while simultaneously showing less convergent validity overall.

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 indirect 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. The present research affirmed its strong psychometric qualities, and also its lack of sensitivity to assessing separate scores for single attitude objects.

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., 2013; 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). However, the lack of discriminant validity in the present study suggests that the BIAT is similar to the IAT in being constrained to relative assessment. Nonetheless, the present research provides strong and broad evidence that the BIAT

has excellent psychometric qualities. Overall, the BIAT was 16% shorter than the IAT in this study and elicited similar psychometric qualities.

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. 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.

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, and as a measure of single-category evaluation. 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 indirect 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. In our comparison of single category measurement quality, the ST-IAT showed evidence for discriminant and convergent validity that was better than most other measures.

Because it is a relatively easy task, the ST-IAT may seem more vulnerable than other measures to non-automatic processes (Stieger, Göritz, Hergovich, & Voracek, 2011). Indeed the ST-IAT had the lowest error-rate of all the indirect measures. 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, when measuring race attitudes and self-esteem, the ST-IAT was related to indirect 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 indirect 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 indirect measure to replicate effects found with another indirect 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 over other measures). These are important features that differentiate EPT from most other indirect measures. So, despite the fact that EPT showed the worst internal consistency, the weakest relationship to other indirect 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 measure. 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 indirect 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 (Cameron et al., 2012), 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 (however, there is still no research on the effect of the number of categories on the psychometric qualities of the AMP).

The present research provides support that AMP is superior to EPT in many psychometric qualities – internal consistency, and relationship with other direct measures. In addition, in the present research the AMP showed some promising qualities in measuring single category evaluation. Mainly, it showed the best discriminant validity. The AMP was very

sensitive to the removal of extreme scores. Extreme scores contributed most of the AMP's positive psychometric qualities. In another line of research Bar-Anan and Nosek (2012) 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 indirect measures.

In the follow-up study mentioned earlier (Bar-Anan & Nosek, 2013), again, the AMP was more sensitive than other measures to the removal of extreme scores. However, the AMP's psychometric qualities were still acceptable even after removing extreme scores, probably because of the increased number of trials. Therefore, research applications that use the AMP should include a larger number of trials than was used in most past AMP research, and should also examine whether the results are dependent on the extreme scores rather than reflective of the entire samples. 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.

Study limitations

It is important to explicitly list a number of weaknesses of the present study. First, the study did not include behavioral measures that are known as sensitive to automatic more than deliberate evaluation (e.g., impression formation toward a black man, Fazio et al., 1995). Establishing the extent to which the measures are influenced by automatic evaluation and distinct from explicit evaluation requires evidence separate from what is provided here (e.g., Cameron et al., 2012; Greenwald et al., 2009)

To compare seven indirect measures without exhausting our participants, we used a web platform and an incomplete data design. These features bring several limitations. First, the website that we used is known to measure attitudes, and some of the sessions were conducted by participants who already completed an earlier session of the study, or other studies that measured attitudes. Additionally, even in one session, the completion of two or three indirect measures, sometimes very similar, might have caused various carry over effects, including fatigue, loss of interest, improvement in performance, and improved understanding of the study's general purpose (attitude measurement). All these may limit the generalization from the present results. For instance, perhaps the accessibility of the evaluative context increased the effect of attitudes on measurement, and was partly responsible for the general good psychometric qualities often observed in the present study.

Summary

The present study compared seven indirect measures on a variety of psychometric qualities. We found strong evidence for inter-relations among all indirect measures. We also found that the attitude domain moderated these relations similarly to its moderation of internal consistency, and of the relationship between each of the seven indirect measures and direct attitude measures. We also found much evidence to support the argument that each of the seven indirect measures is an attitude measure. The results provide comparative information regarding the strengths and weaknesses of each measure relatively to the other measures. We believe that further multi-measure research could help understand the strengths and the weaknesses of the various indirect measures and could also shed more light on evaluative processes, including the popular distinction between the construct measured by indirect measures versus direct measures.

References

Altemeyer, B. (1981). *Right-wing authoritarianism*. Winnipeg: University of Manitoba Press.

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

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38,* 1193-1207.

Bar-Anan, Y., & Nosek, B. A. (2013). The Effect of Number of Trials and Stimulus Set on the Psychometric Qualities of the Affective Misattribution Procedure. *Open Science Framework,* bHNd2. http://opensciencefframework.org/project/bHNd2/

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., Shahar, G., & Nosek, B. A. (2013). Unpublished data.

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.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*, 330-350.

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.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25,* 603-637.

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). Indirect 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., & Bodenhausen, G. V. (2006). Associative and prepositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., & De Houwer, J. (in press). Indirect 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., & 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., 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.

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.

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. (2013). Understanding and using the brief Implicit Association Test: I. Recommended scoring procedures. Unpublished manuscript.

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

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

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

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

Payne, B. K., Cheng, C.M., Govorun, O., & Stewart, B.D. (2005). An inkblot for attitudes: Affect misattribution as indirect 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.

Perugini, Richetin, and Zogmaister, 2010, Prediction of behavior. In B.Gawronski & B.K. Payne (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications, Chapter 14, pp. 255-277. New York: Guilford Press.

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.

Scherer, L. D. & Lambert, A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology, 97*, 383-403.

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

Stieger, R. S., Göritz, A. S., Hergovich, A., & Voracek, M. (2011). Intentional faking of the single category implicit association test and the implicit association test. *Psychological Reports, 109*, 219-230.

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

Supplements to "A Comparative Investigation of Seven Indirect Attitude Measures",

Yoav Bar-Anan and Brian A. Nosek

Appendix A: Methodological Details about the Indirect Measures and their Scoring

As mentioned in the main manuscript, all the procedures of the indirect **measures** were tested prior to the study with the stimuli that were selected for this 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). Table 1 in the main manuscript summarizes key features of the measures and the particular procedures used. All tasks– exactly as they were administered in the study - can be experienced at: http://opensciencefframework.org/project/Qf9jX/node/YJQiq/.

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 (Nosek, Greenwald, & Banaji, 2007). 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.

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.

Speeded Self-report (SR). Although this was a direct measure, we describe it here because it had a more complex procedure than the other direct measures. 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.

Data Processing

The data produced by the indirect 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, Nosek, & Banaji, 2003), and the logic of the *D*-measure (Equation 1) 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).

$$D = M_{condition1} - M_{condition2}/SD \quad (1)$$

D is the standardization of an average latency difference score. It is computed by dividing the difference between average latency of one condition ($M_{condition1}$) and the average latency of the other condition ($M_{condition2}$) by the standard deviation of the participant's response latencies across both critical performance conditions.

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). In all the measures that used D, the score was an average of two or three D scores, each computed from a separate parcel of the task (the parcels were different blocks). For other indirect 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[1].

[1] 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.

As detailed below and in Table A1, the trial and session exclusion rules were usually based on Greenwald et al.'s (2003) recommendation regarding the IAT, with small variations based on the idiosyncratic features of each procedure.

Table A1: Trial and session exclusion before scoring, for each measure

	Lower tail deletion	**Upper tail deletion**	**Error trials treatment**	**Other trial deletion**	**Session deletion**
IAT	< 400ms	> 10000ms	Use latency of correct response	No	> 10% trials faster than 300ms
BIAT	< 400ms	> 10000ms	Use latency of correct response	first 4 of each block	> 10% trials faster than 300ms
GNAT	< 400ms	> 1200ms (deadline)	Exclude trial	No	> 10% trials faster than 300ms
ST-IAT	< 400ms	> 10000ms	Use latency of correct response	No	> 10% trials faster than 300ms
SPF	< 400ms	> 10000ms	Use latency of correct response	No	> 10% trials faster than 300ms
EPT	< -2 std	> 2 std	Exclude trial	No	> 40% error trials
AMP	---No trial exclusion or treatment---				> 95% same response

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 (following 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).

Single category attitude scores. We described the computation of the single-category evaluation scores for the ST-IAT, SPF, AMP and EPT, as a part of the preference scores calculation (which was always the difference between the two single-category scores). For the IAT, BIAT and the GNAT we computed the single-category evaluation score by including only trials that required a response with the key that was associated with the category. For instance, to compute the evaluation score of Republicans in the IAT, we subtracted the average response latency to trials that required responding with the key that was mapped to Republicans when the key was shared with the category Good words from the average response latency to trials that

required responding with the key that was mapped to Republicans when the key was shared with the category Bad words. The rationale was that people who like Republicans would find it easier to respond with a key that was shared by the categories Republicans and Good words than with a key that shared with the categories Republicans and Bad words. We followed the same logic for computing the single category evaluation scores in the BIAT and the GNAT. To date, they have not been evaluated for single-category assessment potential by this analytic strategy. Prior research that tested this method with the IAT found poor discriminant validity for single attitude measurement (Nosek et al., 2005). One of the unique contributions of the present study is that it provides direct test between the measures that are constrained by relative measurement and the measures that, at least theoretically, seem to provide separate measurement for each attitude object.

References for Appendix A

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

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.

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

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.

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.

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.

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

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.

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 indirect 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.

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.

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

Appendix B: Sensitivity to Data Exclusion due to Unusual Behavior

The last comparison criterion in our study tested the sensitivity of each measure to session exclusion due to unusual behavior of the participant in the task. The main manuscript provided a summary of that test, and this appendix provides more details.

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 showed evidence of misbehavior 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 instructed. 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).

For each measure, we examined the internal consistency, average correlation with indirect measures and average correlation with direct measures as a function of removing a certain percentage of participants based on errant behavior. 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.

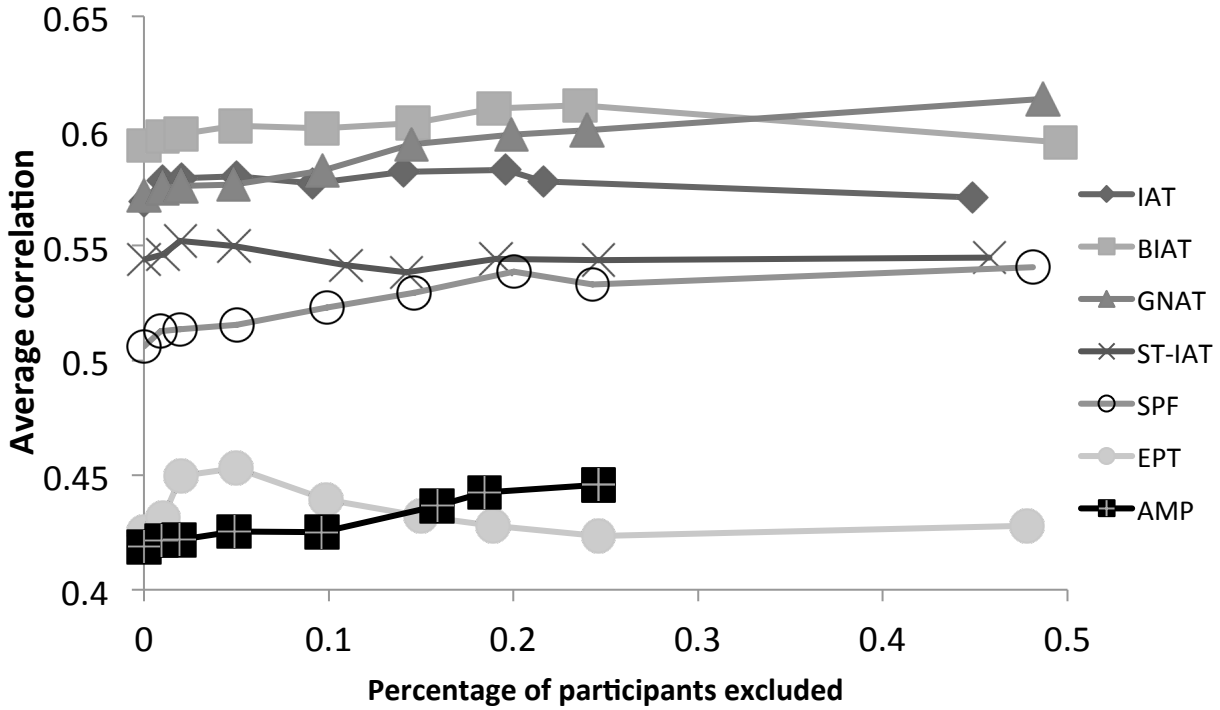


Figure B1. The effect of removing misbehaving participants on the average relationship of the indirect measures with direct measures (politics). About 75% of the participants who performed the AMP had no fast or slow trials, so we could not remove more than 25%.

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.

The psychometric qualities of most of the measures hardly changed when suspect participants were removed (For all the graphs see the Appendix D). 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 indirect and direct measures, regardless of how many suspect participants were removed (For an

example, see Figure B1). Notable exceptions were the SPF politics measure that, after removing the 50% most suspect participants, improved in almost .05 in its average correlation with indirect 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 indirect and direct 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 B2 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).

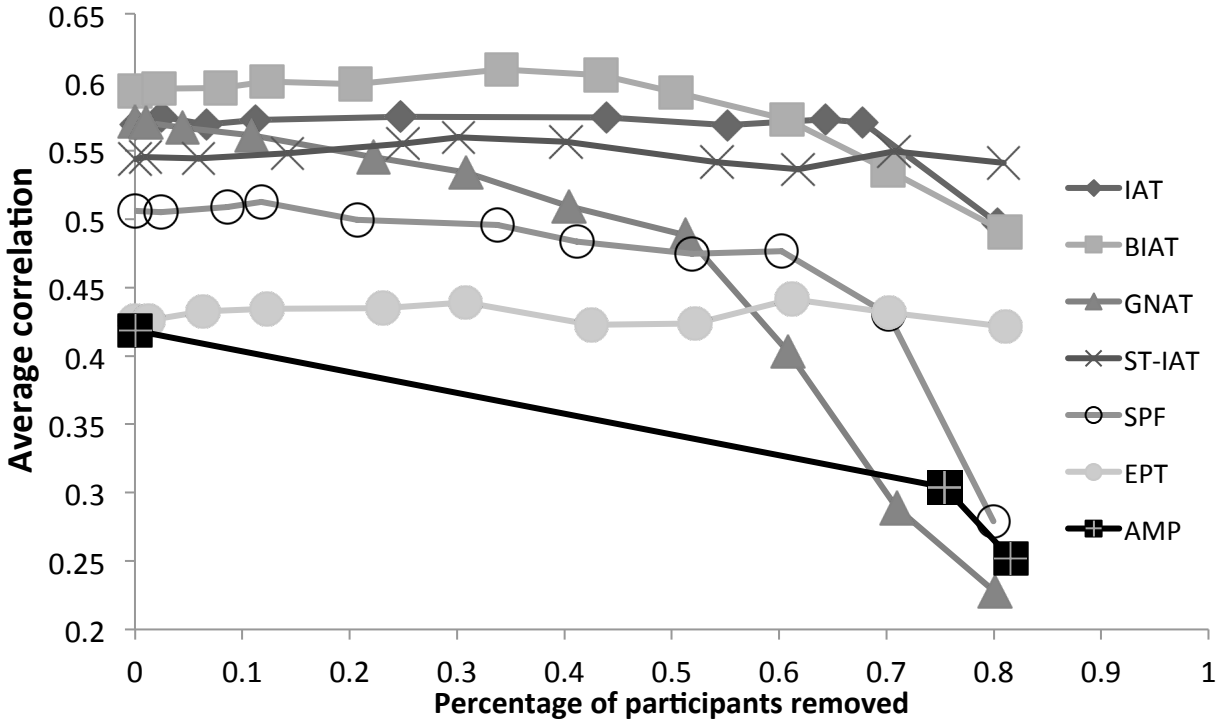


Figure B2. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with direct measures (politics). The AMP has less data points because about 75% of the participants who performed the AMP showed perfect behavior.

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% best behaved participants were removed. The GNAT politics also showed a drop of about .09 in its average correlation with indirect 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.

In summary, 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.

Appendix C: Additional Tables

Table C1

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; **Bold** = the strongest correlation of that column; *Underlined Italics* = the weakest correlation of that column.

Table C2

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; **Bold** = the strongest correlation of that column; *Underlined Italics* = the weakest correlation of that column.

Table C3

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 C4

Single-Category, discriminant validity of direct measures

	Thermometer category ratings	
	White people	Black people
White people item ratings	.31	.20
Black people item ratings	.20	.47
MRS	.01	-.28
	Democrats	Republicans
Democrats item ratings	.71	-.56
Republicans item ratings	-.47	.73
RWA	-.38	.60
	Self	Other
Self speeded self-report ratings	.28	.16
Other speeded self-report ratings	.09	.29
Rosenberg	.57	.27

Table C5

Single-category criteria, full summary

	Known groups effect	Internal Consistency	95% CI	Test-Retest Correlation	Mean correlation indirect	Mean correlation direct	Discriminant validity
Black people							
IAT	**1.01**	.73	.72-.75	.30	.28	**.20**	*-.16*
BIAT	0.72	.64	.62-.66	**.53**	.25	.17	-.09
GNAT	0.50	.61	.59-.63	*.11*	.24	.18	.02
ST-IAT	0.24	.73	.71-.75	.22	.21	.19	-.04
SPF	0.44	*.43*	.40-.46	.29	.16	.17	-.03
EPT	0.44	.66	.64-.68	.28	.14	*.14*	-.04
AMP	*0.24*	**.85**	.84-.85	.48	*.12*	*.14*	.12
White people							
IAT	**1.05**	.73	.72-.75	.35	**.23**	*.04*	*-.16*
BIAT	0.57	.65	.62-.67	.50	.22	.07	-.09
GNAT	0.38	.60	.57-.62	.21	.20	.08	.02
ST-IAT	0.26	.72	.71-.74	.31	.17	.07	-.04
SPF	0.63	*.45*	.42-.47	*.15*	.13	.06	-.03
EPT	0.27	.58	.55-.60	.31	.10	*.04*	-.05
AMP	*0.11*	**.79**	.78-.81	**.69**	*.07*	**.14**	**.12**
Democrats							
IAT	**1.36**	**.86**	.85-.87	.58	.48	**.51**	*-.06*
BIAT	1.26	.75	.77-.80	.61	.44	.47	*-.06*
GNAT	1.09	.73	.71-.75	.54	.44	.44	-.01
ST-IAT	0.78	.82	.81-.83	.28	.39	.44	.08
SPF	0.85	*.47*	.44-.50	.39	.38	.33	-.03
EPT	0.49	.66	.64-.68	*.34*	.26	*.28*	.03
AMP	*0.38*	.84	.83-.85	**.67**	*.21*	.32	.11
Republicans							
IAT	**1.38**	**.85**	.84-.86	.60	.42	.53	.06
BIAT	1.27	.78	.77-.80	**.74**	.46	**.54**	.08
GNAT	1.02	.71	.69-.73	.46	.41	.44	*.03*
ST-IAT	0.75	.81	.80-.82	*.41*	.32	.38	.08
SPF	0.79	*.49*	.46-.51	.51	.33	.36	.11

EPT	*0.51*	.66	.64-.68	.54	.32	*.30*	.10
AMP	0.61	**.85**	.84-.86	.72	*.25*	*.30*	.11
Self							
IAT		.68	.66-.70	.18	.17	.14	.06
BIAT		.59	.56-.61	.37	.18	.11	*-.02*
GNAT		.53	.51-.56	.16	.15	.16	.14
ST-IAT		.75	.73-.76	.35	.18	.13	.09
SPF		*.48*	.45-.51	.37	.11	*.03*	.02
EPT		.64	.62-.66	*.30*	.08	.07	.07
AMP		**.79**	.77-.80	**.47**	*.05*	**.17**	.10
Other							
IAT		.69	.68-.71	.34	.12	*.04*	*-.06*
BIAT		.57	.55-.60	.37	.15	.10	-.01
GNAT		.62	.59-.64	.19	.16	.09	**.06**
ST-IAT		.71	.31-.38	.25	.11	.07	.04
SPF		*.35*	.31-.38	.27	.05	*.04*	.04
EPT		.59	.56-.61	*.08*	.06	.08	**.06**
AMP		**.79**	.78-.81	**.46**	*.02*	**.18**	.03

Table C6

Inter-relations between the direct measures

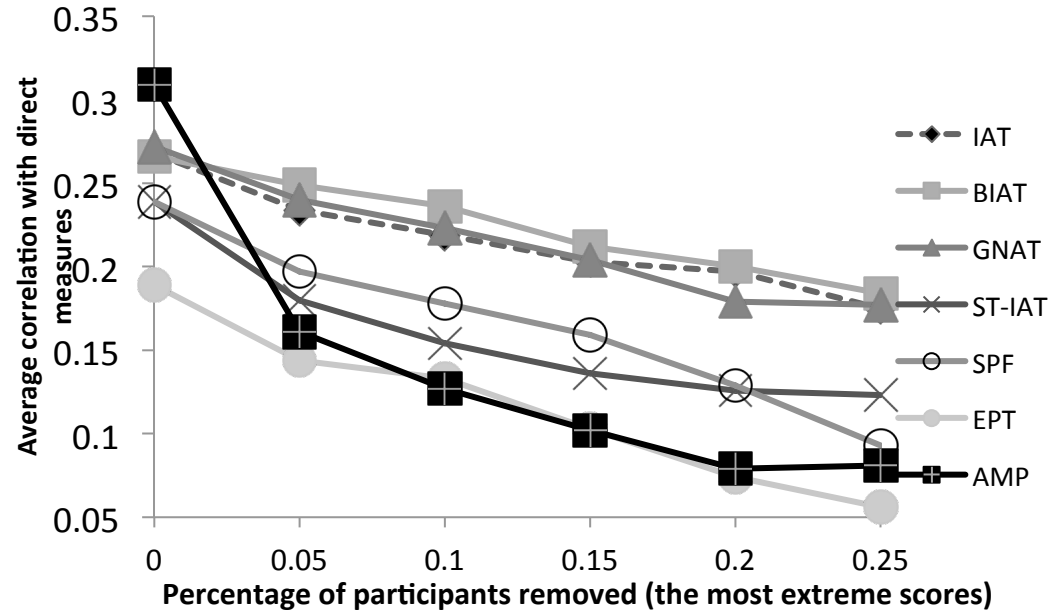
Race	Preference	Thermometer	Items	MRS	Contact (reversed)	Speeded Report
Avg. N	1796	1829	596	607	1823	531
Range N	516-3645	528-3743	555-645	532-645	524-3743	516-555
Preference		.75	.43	.35	.31	.37
Therm.	.75		.48	.35	.27	.45
Items	.43	.48		.43	.11	.50
MRS	.35	.35	.43		.08	.42
Contact	.31	.27	.11	.08		.15
Speeded	.37	.45	.50	.42	-.15	
Avg. Corr.	.46	.48	.40	.33	.13	.38

Politics	Pref.	Thermometer	Items	RWA	Voted	Voting intentions	Speeded Report
Avg. N	1655	1670	395	476	971	1597	456
Range N	439-3494	441-3295	232-442	273-557	232-1722	406-3295	249-556
Pref.		.87	.81	-.63	.79	.68	.70
Therm.	.87		.80	-.59	.73	.62	.70
Items	.81	.80		-.58	.81	.71	.81
RWA	-.63	-.59	-.58		-.62	-.50	-.52
Voted	.79	.73	.81	-.62		.67	.67
Intentions	.68	.62	.71	-.50	.67		.53
Speeded	.70	.70	.81	-.52	.67	.53	
Avg. Corr.	.64	.61	.65	-.58	.58	.50	.55

Self	Preference	Thermometer	Rosenberg	Speeded Report
Avg. N	1687	1688	614	562
Range N	568-3847	569-3847	548-648	548-569
Pref.		.56	.18	.17
Therm.	.56		.31	.24
Rosenberg	.18	.31		.16
Speeded	.17	.24	.16	
Avg. Corr.	.32	.38	.22	.19

Appendix D: Additional figures

A



B

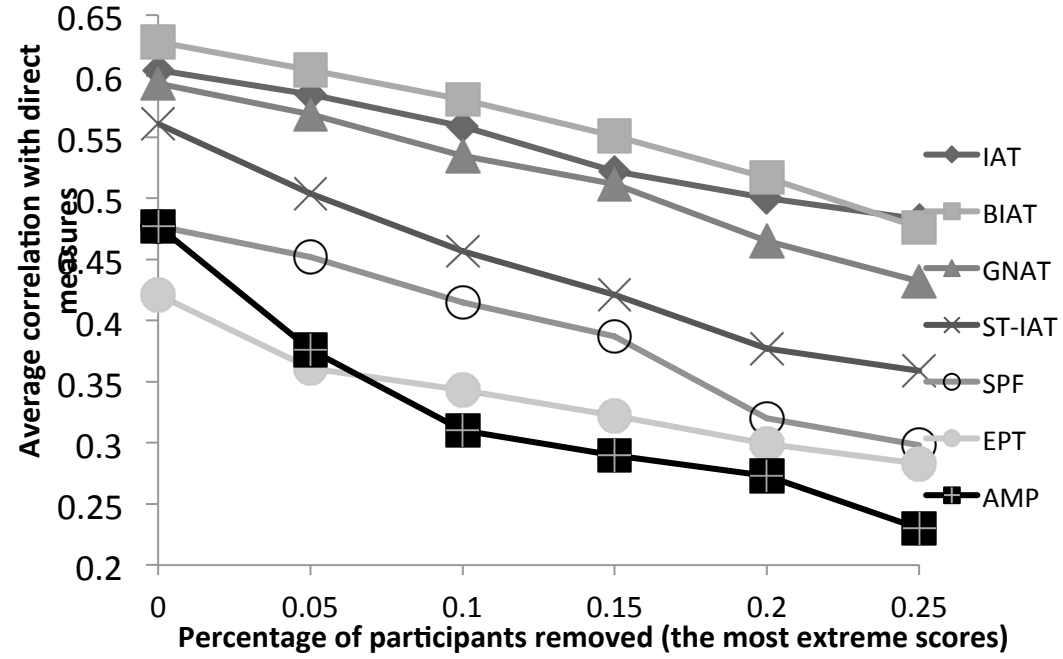
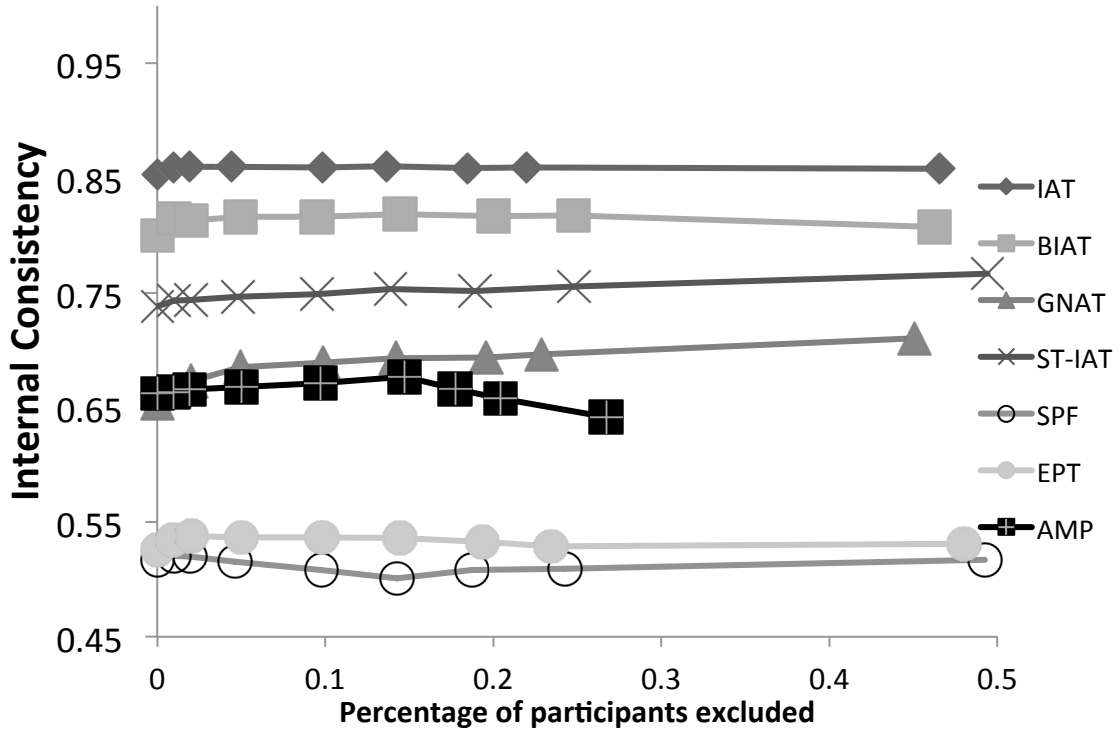
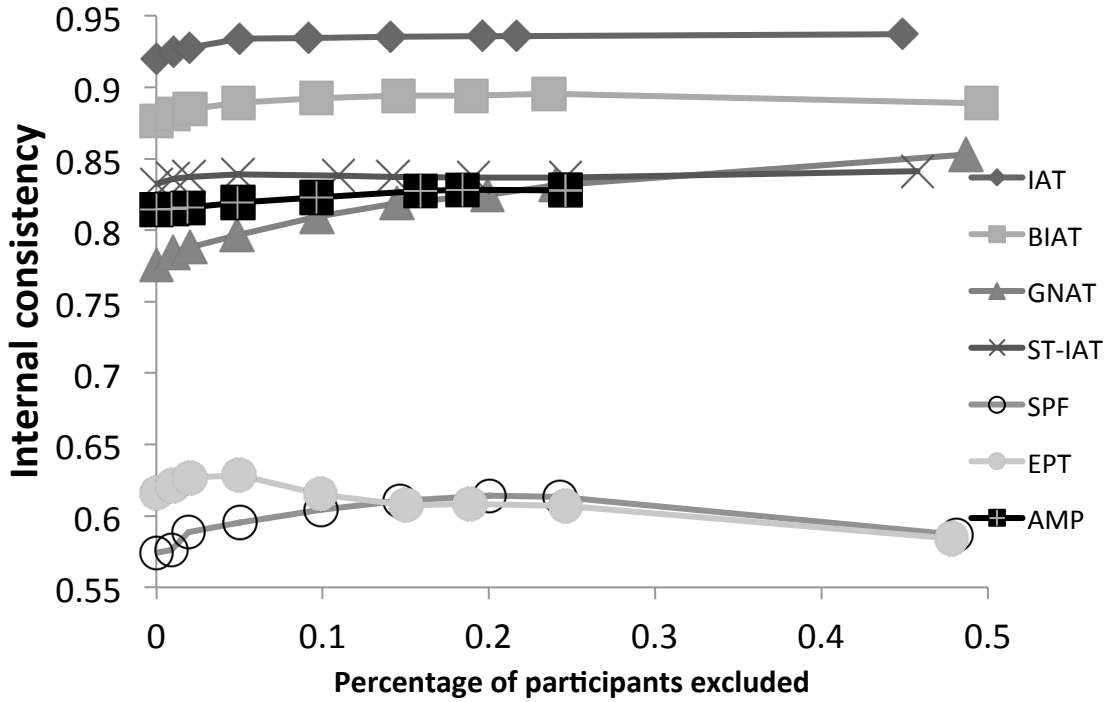


Figure D1. The effect of removing extreme scores (by percentage) from each indirect measure on its average correlation with the direct measures and other criterion variables. Panel A: Race measures; Panel B: politics measures.

A


Internal Consistency
0.95
0.85
0.75
0.65
0.55
0.45
0
0.1
0.2
0.3
0.4
0.5
Percentage of participants excluded
IAT
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

B


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

C

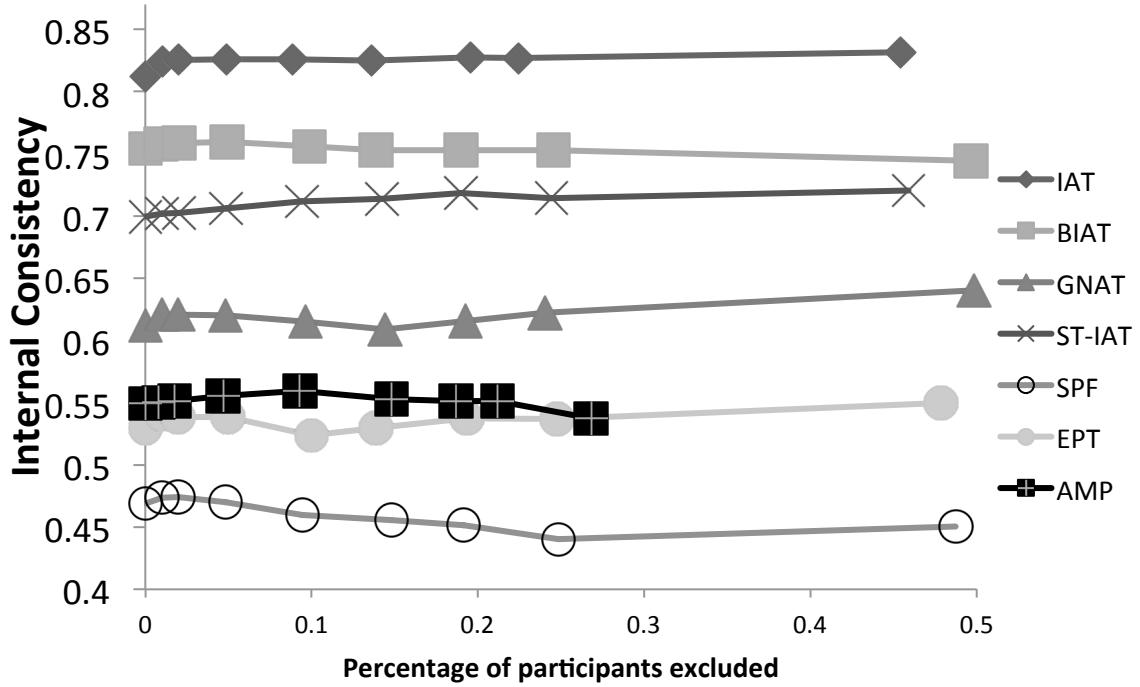
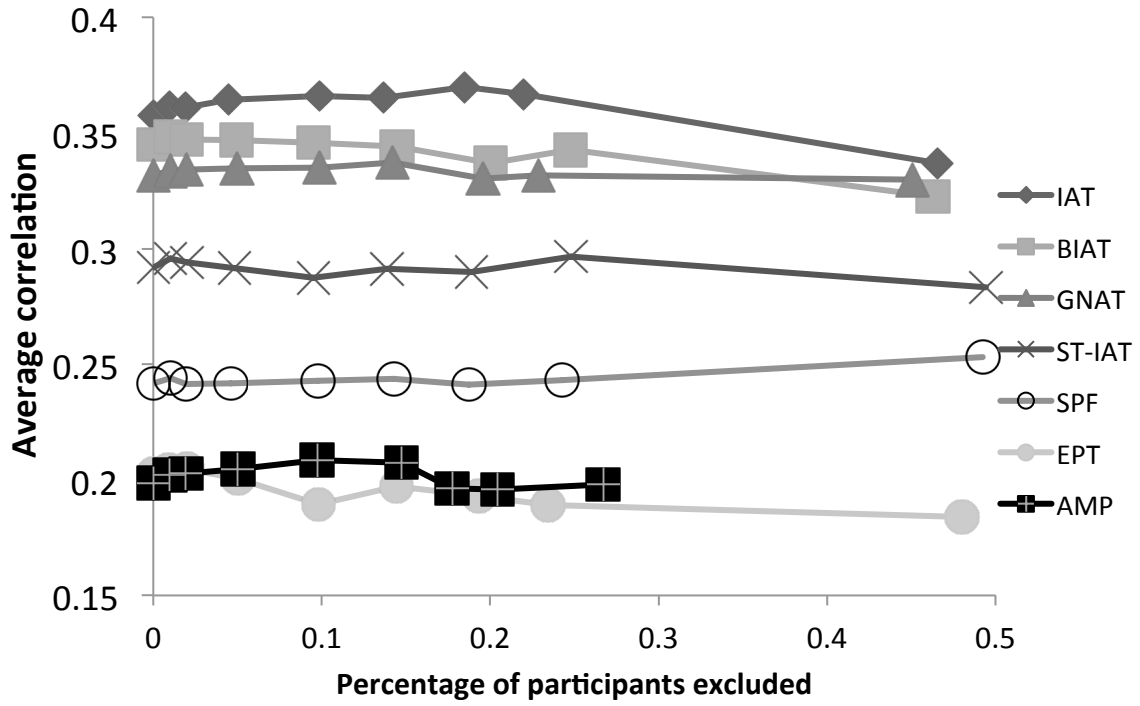
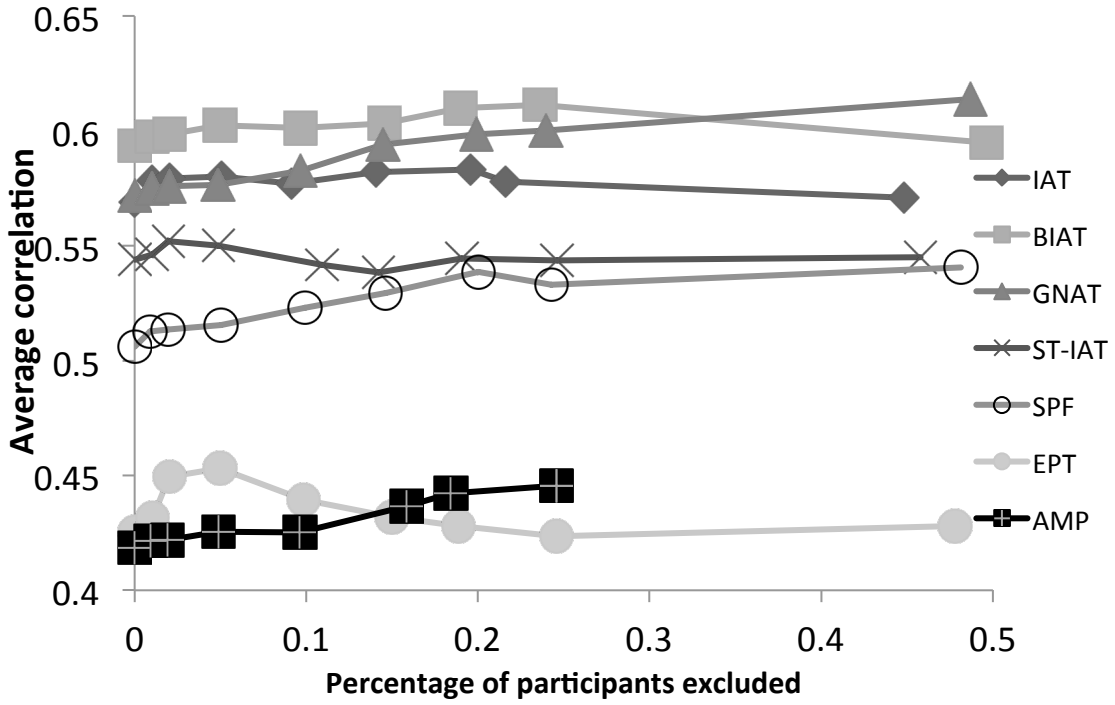


Figure D2: The effect of removing misbehaving participants on the internal consistency of the indirect measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

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
BIAT
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
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

C

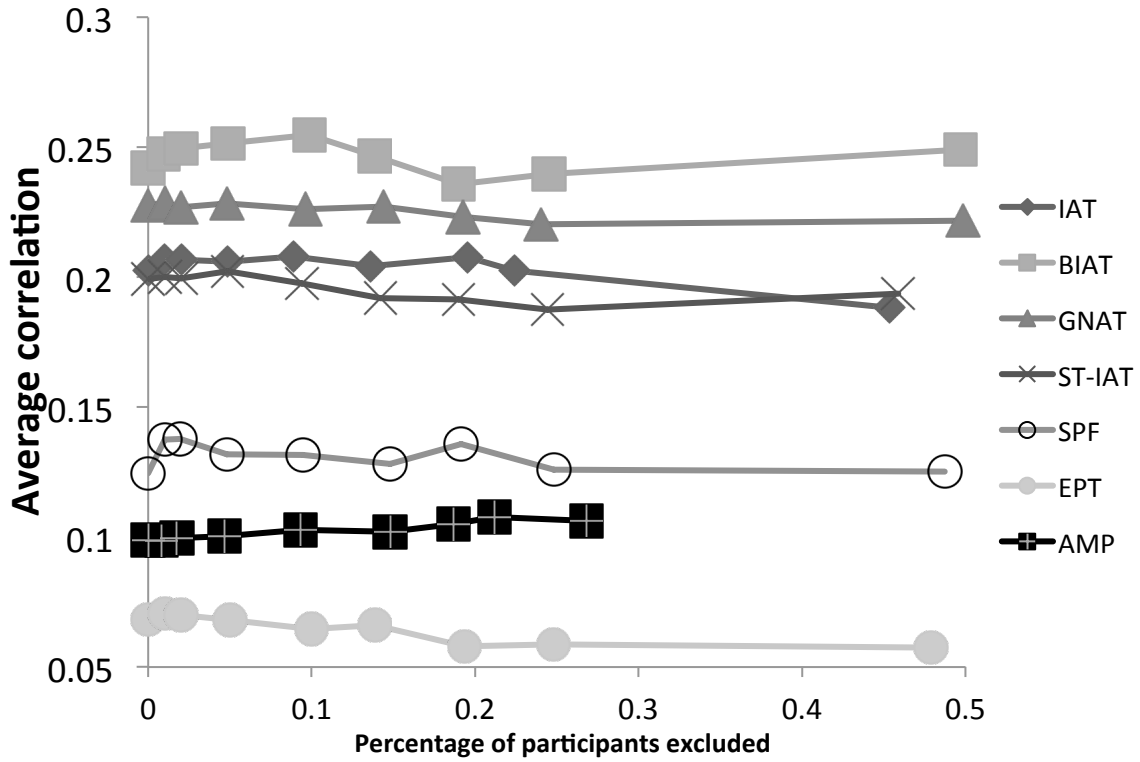
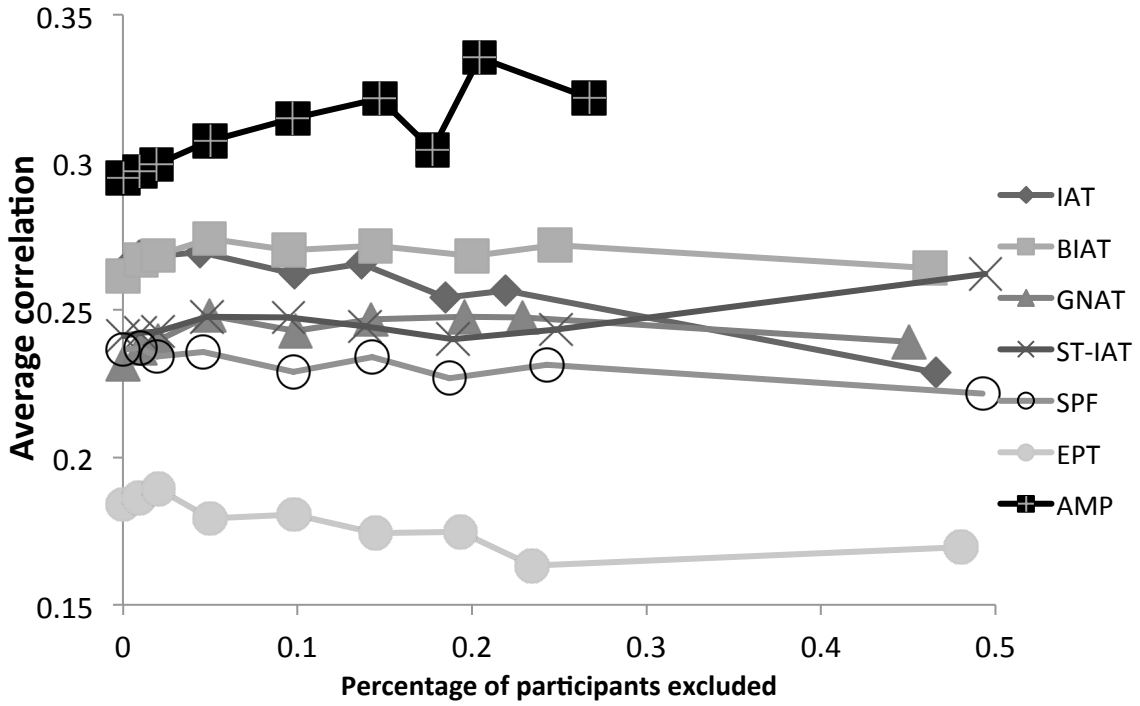
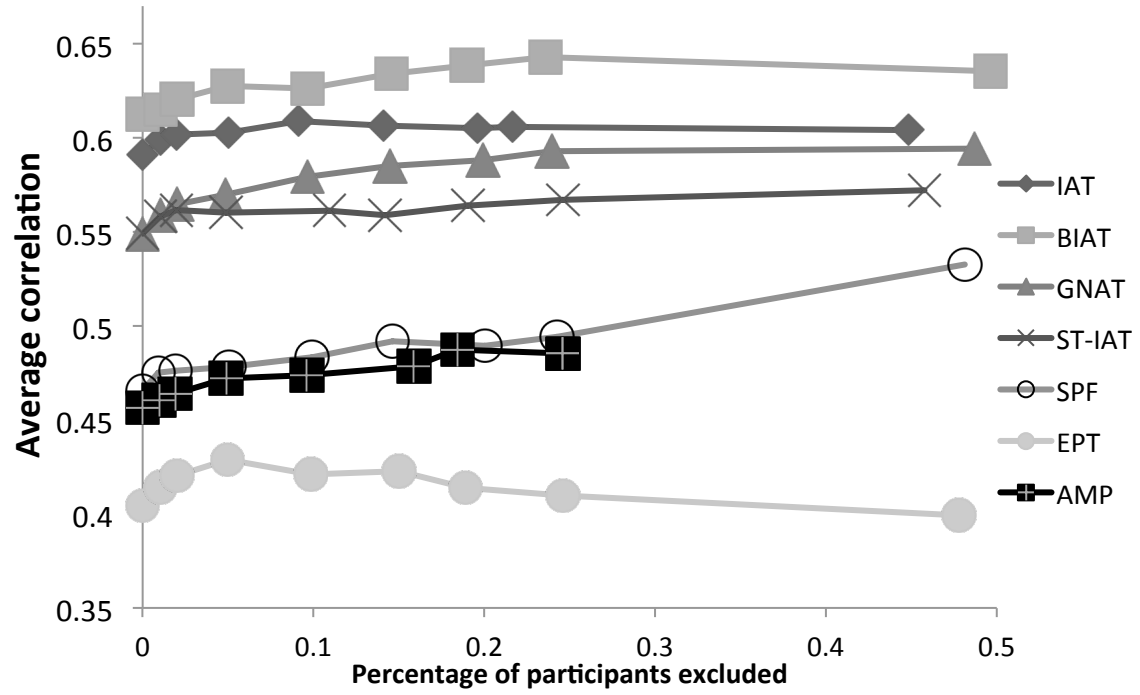


Figure D3: The effect of removing misbehaving participants on the average relationship of the indirect measures with other indirect measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

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 of participants excluded
IAT
BIAT
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
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

C

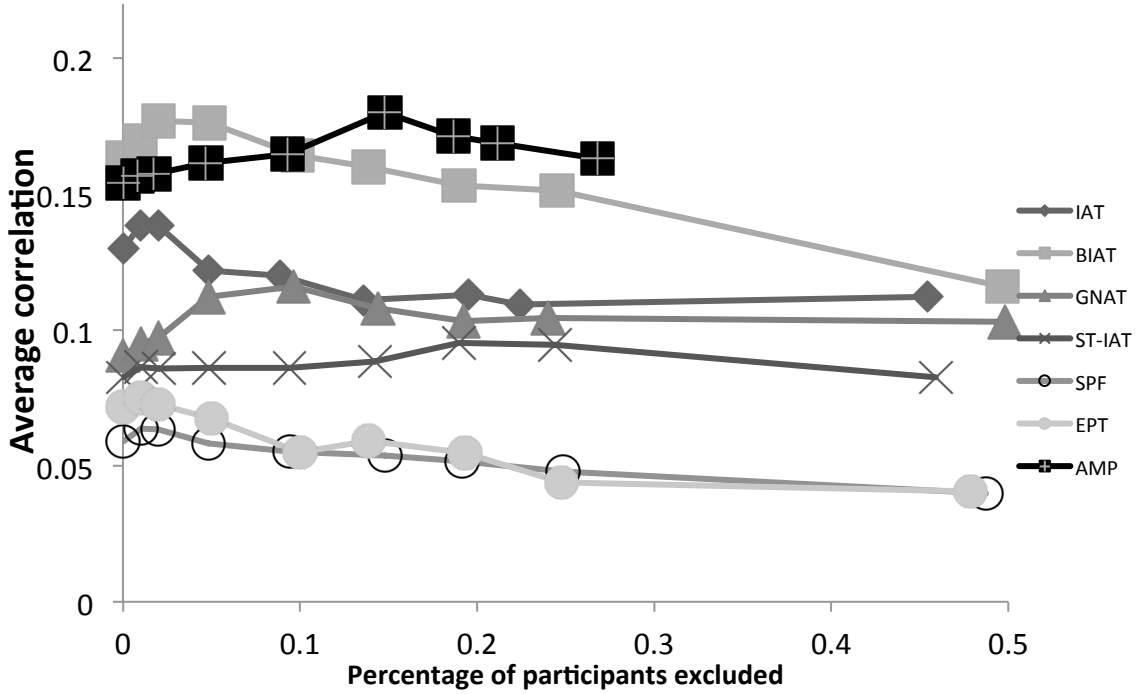
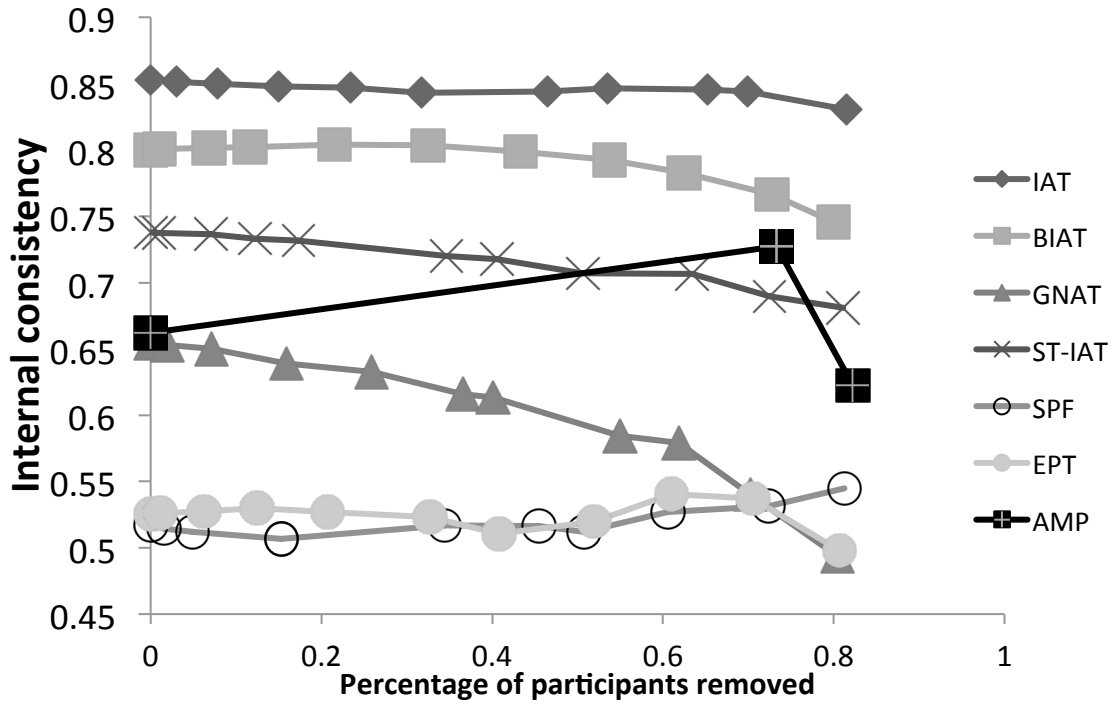
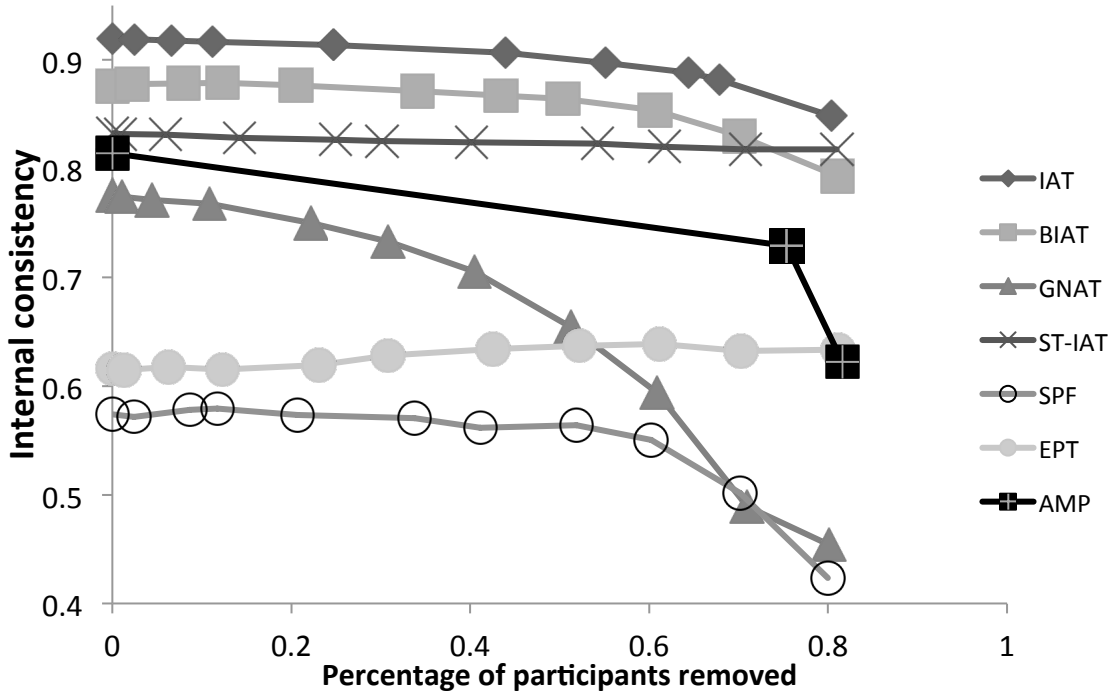


Figure D4: The effect of removing misbehaving participants on the average relationship of the indirect measures with direct measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

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
BIAT
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
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

C

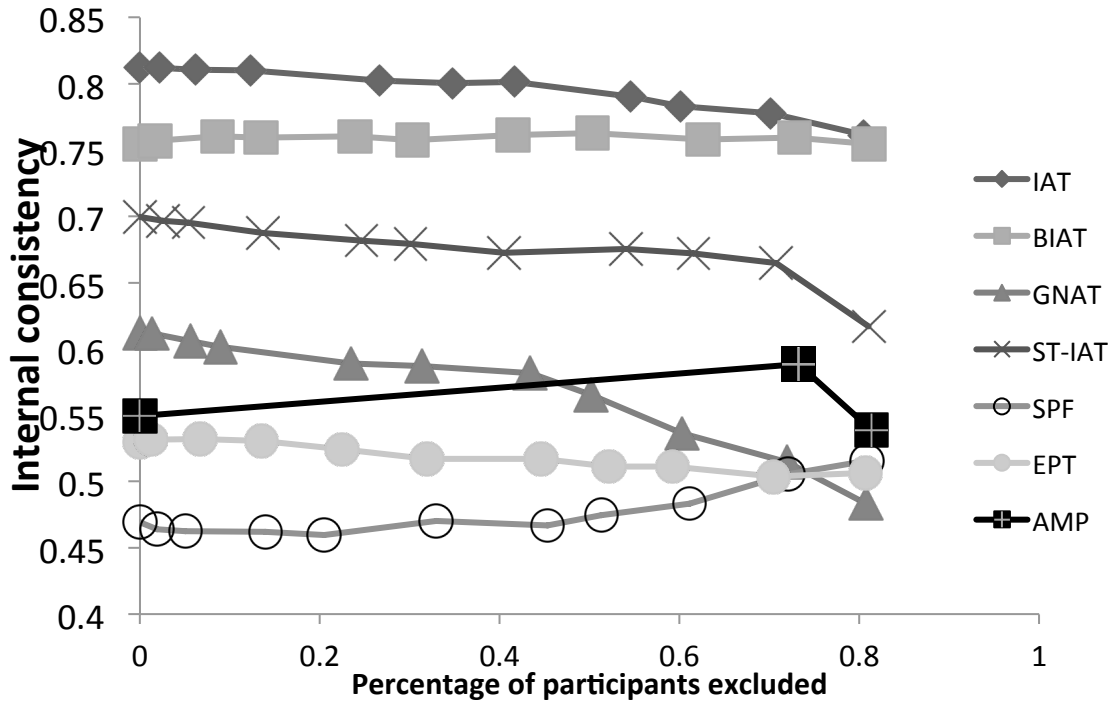
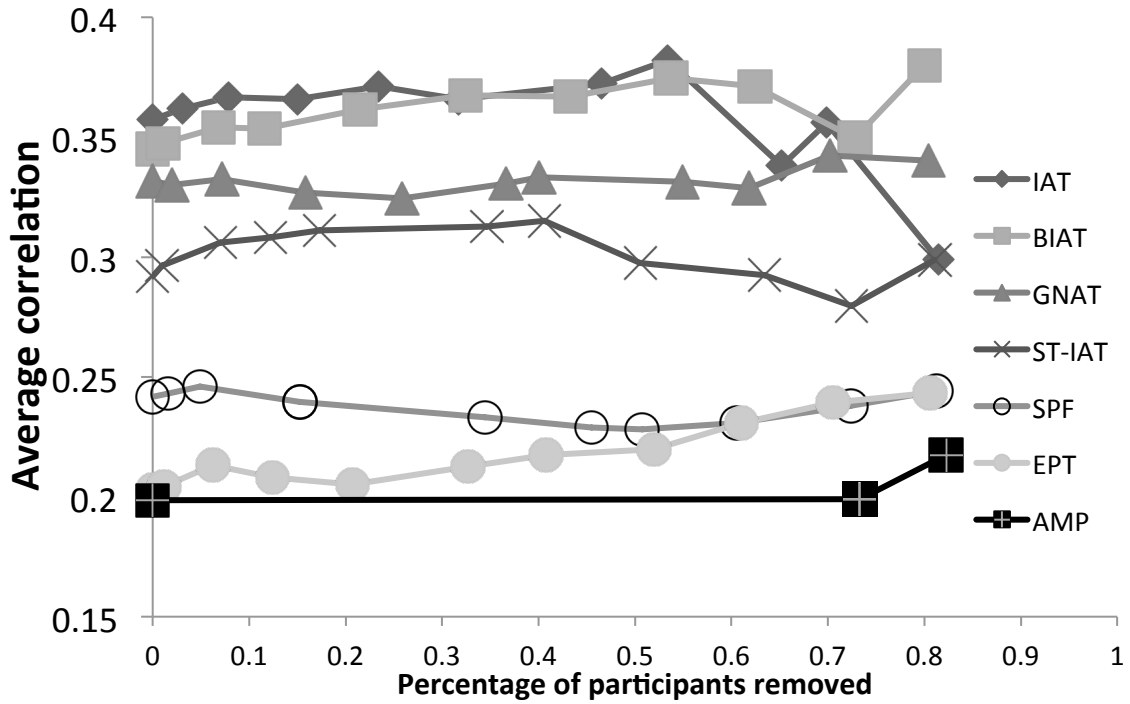
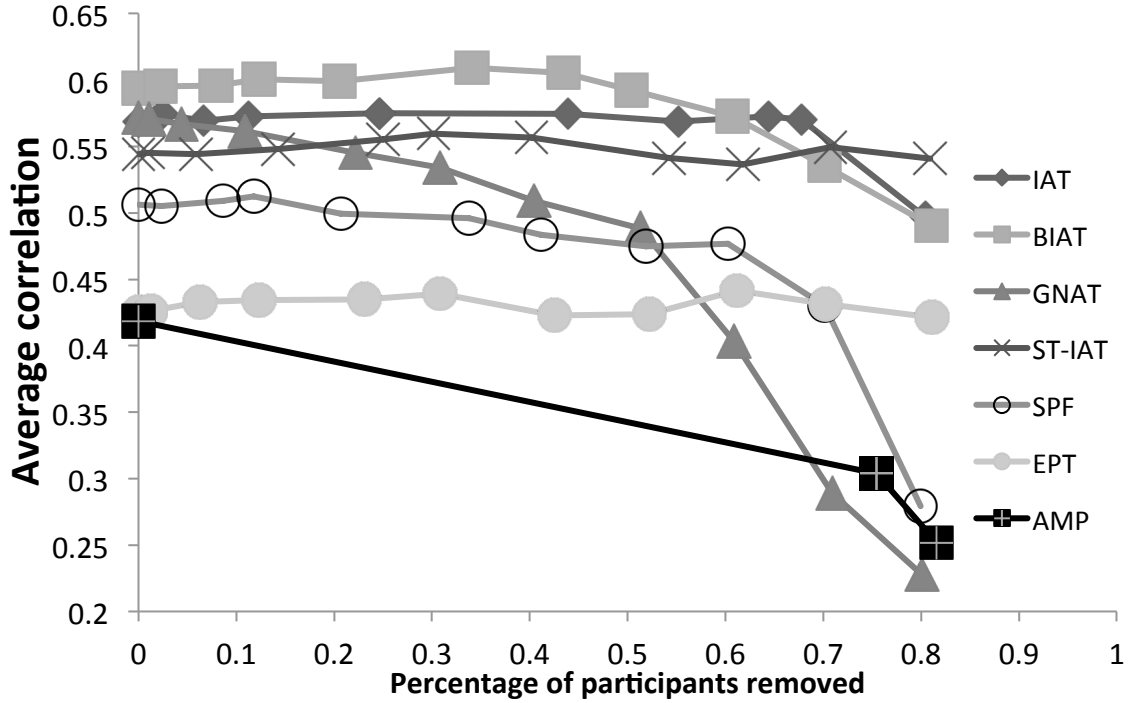


Figure D5. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the internal consistency. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

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
0.6
0.7
0.8
0.9
1
Percentage of participants removed
IAT
BIAT
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
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Percentage of participants removed
IAT
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

C

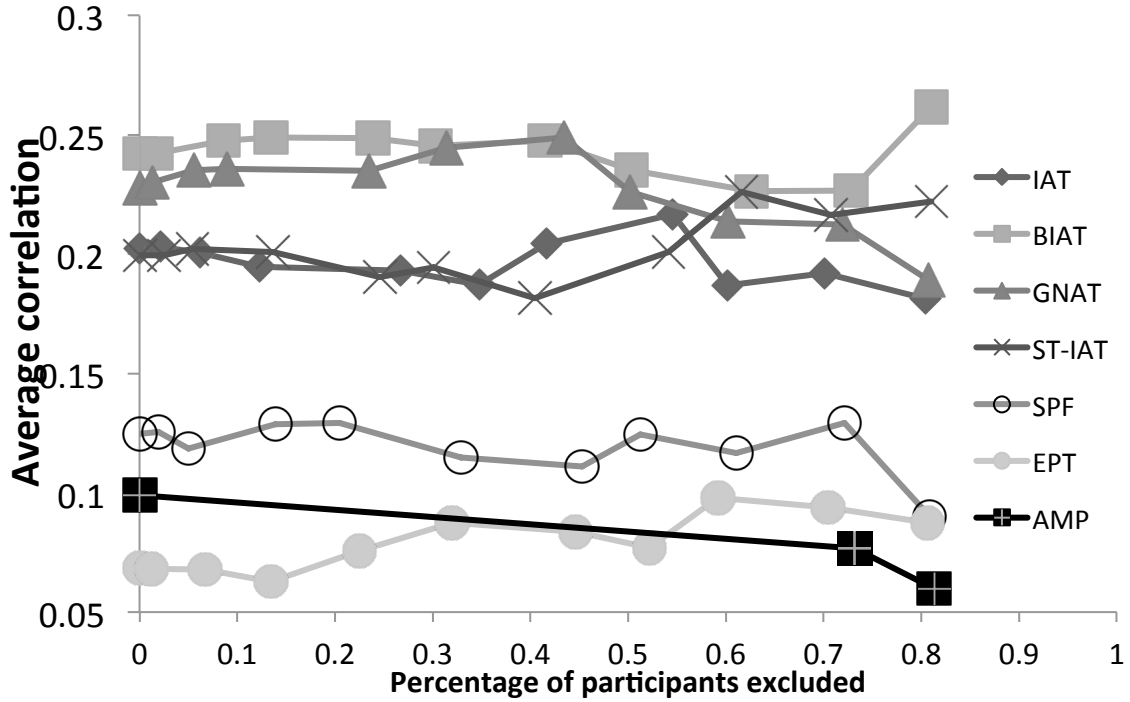
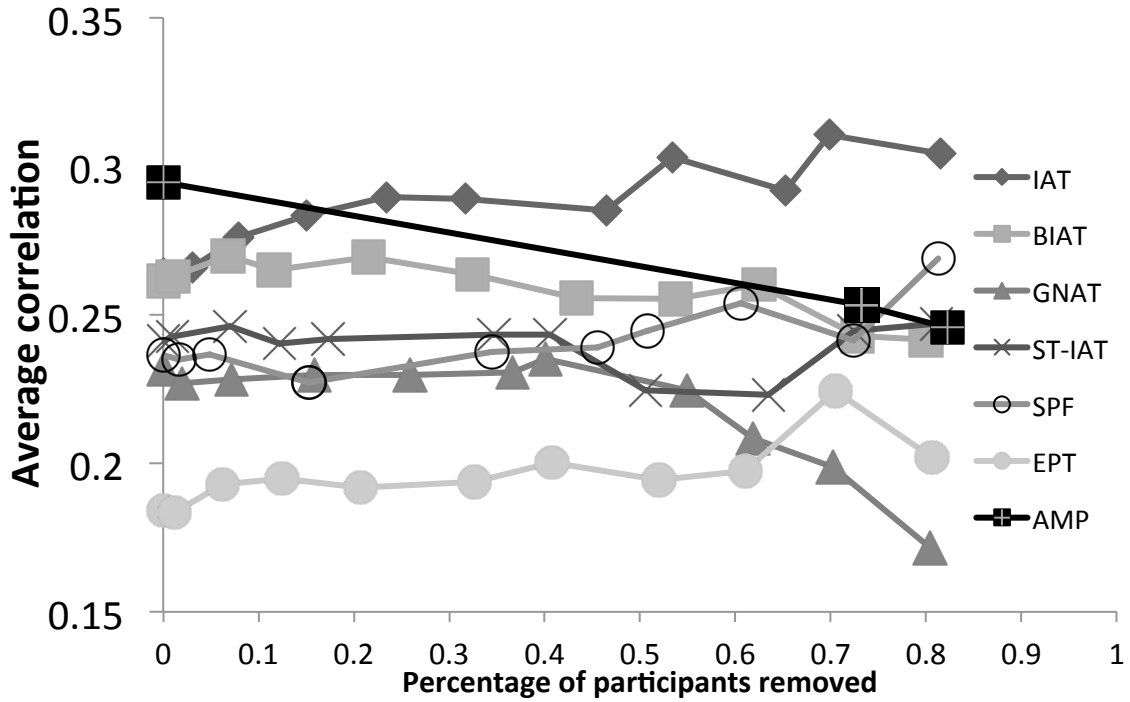
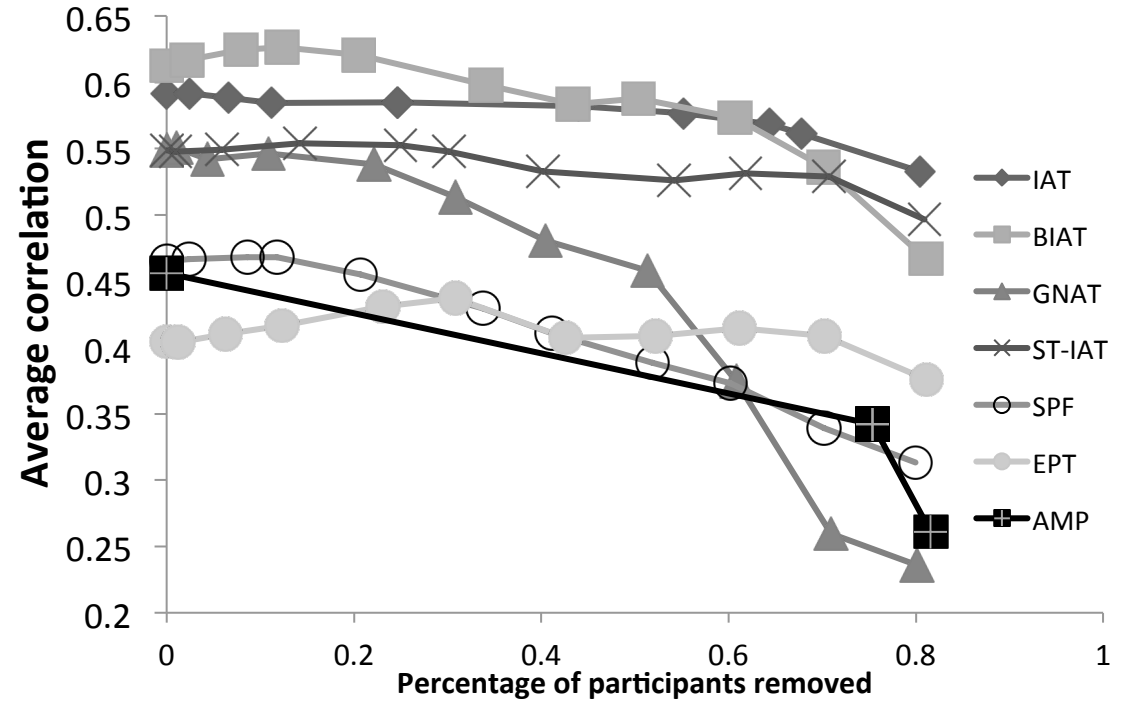


Figure D6. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with indirect measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.

A


0.35
0.3
0.25
0.2
0.15
Average correlation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Percentage of participants removed
IAT
BIAT
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
BIAT
GNAT
ST-IAT
SPF
EPT
AMP

C

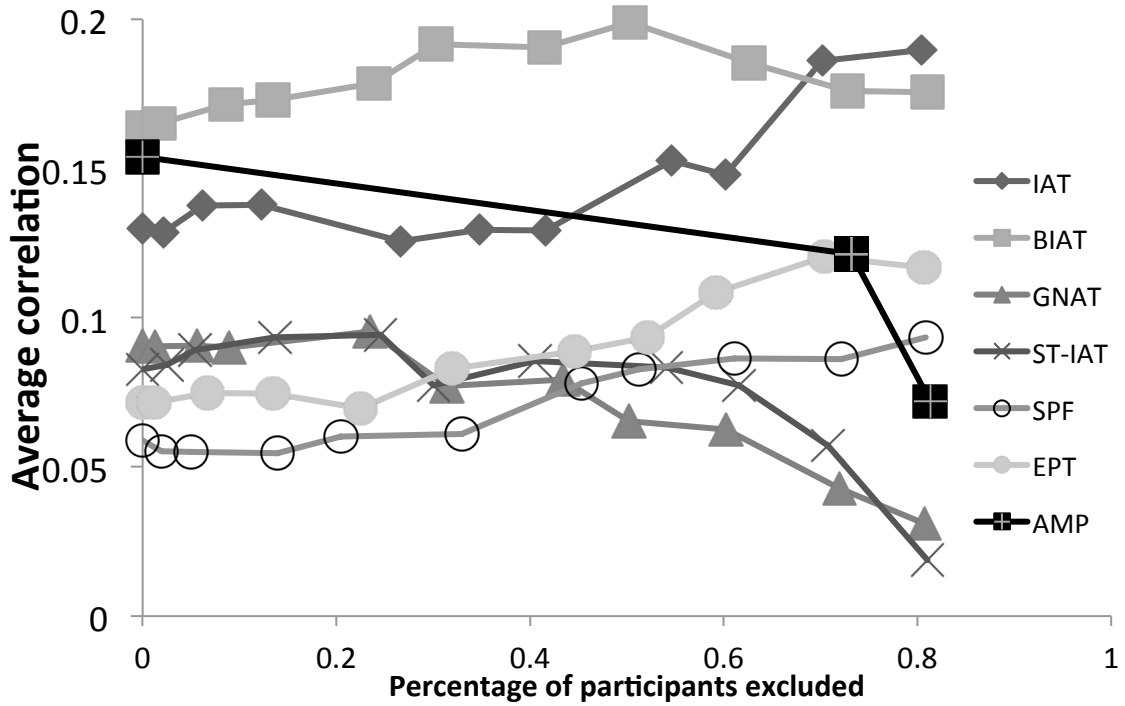


Figure D7. The effect of excluding "well-behaved" participants (participants with small number of suspect trials) on the average relationship with direct measures. Panel A: Race; Panel B: Politics; Panel C: Self-esteem.