Behav Res (2014) 46:668–688
DOI 10.3758/s13428-013-0410-6

A comparative investigation of seven indirect attitude measures

Yoav Bar-Anan · Brian A. Nosek

Published online: 14 November 2013

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, relationships with indirect and direct measures of the same topic, the reliability and validity of single-category attitude measurement, their ability to detect meaningful variance among people with nonextreme attitudes, and their robustness to the exclusion of misbehaving or well-behaving 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 sources of variation in the reliability and predictive validity of the 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 relationships 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. These results inform us on the validity of the 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

The emergence of implicit social cognition in the last three decades has been 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 being 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 US 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 they are preceded by a Democratic versus a Republican prime. Democrats may be faster to categorize good words (and slower to categorize bad words) when they are preceded by Democratic primes, whereas Republicans may be faster to categorize good words (and slower to categorize bad words) when they are 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).

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0410-6) contains supplementary material, which is available to authorized users.

Y. Bar-Anan (✉)
Psychology Department, Ben Gurion University of the Negev, Be'er Sheva, Israel
e-mail: baranany@bgu.ac.il

B. A. Nosek
University of Virginia, Charlottesville, VA, USA

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 the 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 an increasing diversity of indirect measurement methods, much less is known about the psychometric properties of measures other than the IAT and EPT. Moreover, little systematic knowledge is available on the comparative psychometric qualities and performance of different indirect measures. Relatively few studies have used multiple indirect measures in the same study and experimental setting, thus creating a vacuum of comparative knowledge regarding 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, Cheng, Govorun, & Stewart, 2005), Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009), Go–No-Go association task (GNAT; Nosek & Banaji, 2001), Single-Target Implicit Association Test (ST-IAT; 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 to enable precise estimation. The present investigation allowed for 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 provided 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 interrelations 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 showed weak relations among multiple indirect measures of self-esteem (Bosson, Swann, & Pennebaker, 2000). Part of the lack of relationship was attributable to the 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 et al., 2009; Karpinski & Steinman, 2006; Ranganath, Smith, & Nosek, 2008). In the present research, we provide evidence as to 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. On the basis of previous research and theory (Bosson et al., 2000; Nosek, 2005), we predicted that politics would elicit the strongest indirect–direct relations, and that self-esteem would show the weakest indirect–direct relations. However, no consistent pattern of results in previous research allowed for strong predictions 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 were to 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 nontrivial amount of time to administer and is mentally taxing to complete. Furthermore, large participant samples are necessary in order 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 overtaxing 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, and 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 measures 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 interindividually.

Nonetheless, there is an important challenge with using direct measures as evaluation criteria 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 nonextreme 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 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 website 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).

A total of 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 three measures, and 31 % completed four measures. 45.1 % completed more than four measures, of which 10 % completed more than ten 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 % multiracial, 1.8 % other, and 3.2 % did not identify. In all, 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 six pictures of white people (three females, three males), and 6 pictures of black people (three females, three 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: five Democrats (Barack Obama, Hillary Clinton, Bill Clinton, Al Gore, and John Kerry) and five 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 of 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 in light of the present knowledge and the practical constraints of the study (time, accuracy, and the need for clear and succinct instructions). Table 1 summarizes the key features of the measures and the particular procedures used. The supplemental materials provide full details. All of the tasks—exactly as they were administered in the study—can be experienced at http://opensciceframework.org/project/Qf9jX/node/YJQiq/.

Implicit Association Test The IAT procedure followed the one described in Nosek, Greenwald, and Banaji (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 quickly as 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 Blocks 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 nonfocal categories).

Go–No-Go association task The GNAT procedure was based on Nosek and Banaji (2001), designed for scoring on the basis of 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 +

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

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 the error rate was combined into the score)

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 US Democrats [Black people][yourself] and Republicans [White people][other people]?" Seven response options were presented, 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 Two item rating questionnaires were presented: 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 et al., 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. Although 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 US citizens, the two last words in the statement "Discrimination against Blacks is no longer a problem in the US" 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 ten 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?" The ten response options ranged from *All* to *None*.

Voting behavior Participants reported whether they had voted and which candidate they had voted for in the most recent past US 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 eight 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 min. Measures were randomly selected for each session, with the constraint that two "long-duration" measures and two "short-duration" measures were always presented, and the same measure (i.e., the same method and topic) could not be selected twice in a single session. Otherwise, we imposed no constraint on the repetition of topics or methods 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 the long-duration and short-duration measures and illustrates their selection for a session. Participants could initiate additional sessions and could receive identical measures from previous sessions, in order to facilitate test–retest comparisons. At the end of each session, the purpose of the study was explained to the participants, and they received result feedback for the indirect measures that 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 of the indirect measures except the AMP (d = –0.23) indicated a preference for White rather than 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 item 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.

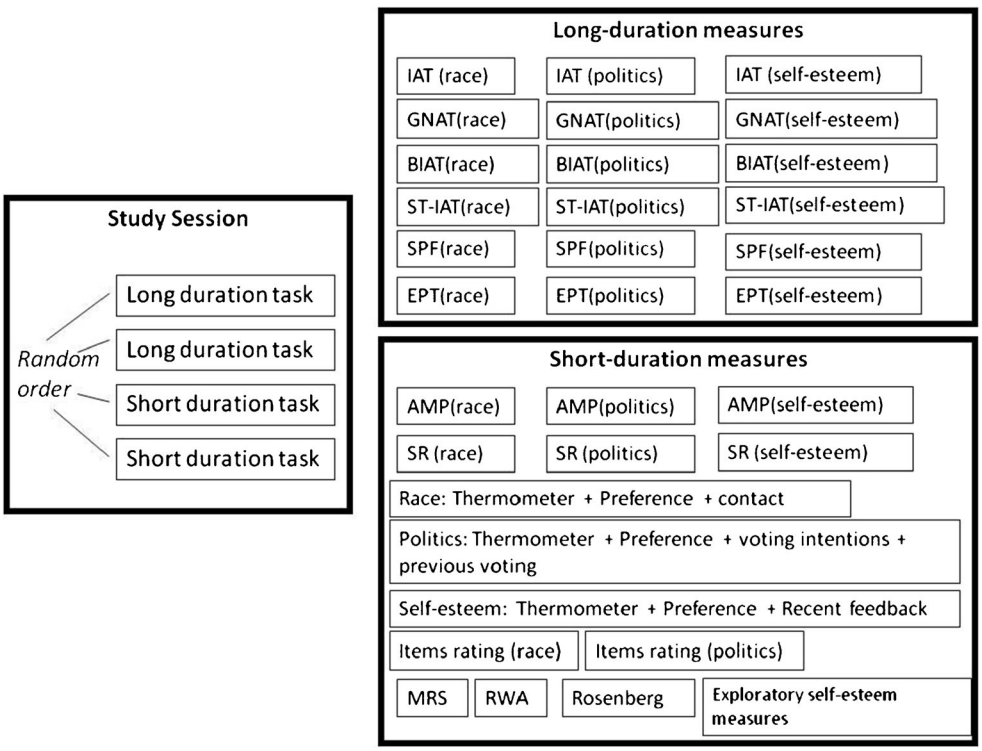

Fig. 1 Illustration of the study procedure. Each study session included two long and two short tasks, selected and ordered randomly. The tasks are listed on the right. Measures that share a rectangle were presented together in the same questionnaire

All of 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 of 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 the 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 can be seen in Table 2, the scores of the IAT, BIAT, and GNAT were the most sensitive to a 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 and 1.04, respectively), and the AMP and the EPT were the least sensitive, about 35 %–50 % weaker than the strongest measures (ds = 0.88 and 0.73, respectively).

In summary, the IAT and the BIAT showed the best sensitivity to detect expected effects 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 the 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 nonoverlapping. 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 most consistent measures, respectively. Comparatively, the 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. The IAT and BIAT had R-squared values of 77 % and 69 % for average internal consistency, whereas the EPT and SPF had 32 % and 28 %, less than half 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 h 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 the internal consistency of the measures, rather than indicating their stability over time. Table 2 presents the test–retest correlations both 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 had 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 relative to the 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 creates 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, which we pursue in the next section. Concern (b) can be addressed by examining whether clusters of strong covariation occur. An in-depth examination of the structural relations among indirect measures goes beyond the scope of this article, but it is taken up in detail with these data by Bar-Anan, Shahar, and Nosek (2013).

The average correlations of each measure with the other indirect measures are presented in Table 2, and the correlation

Table 2 Summary of 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
Average											
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*
	Mean Effect Size	Known-Groups Effect	Alpha	95 % CI	Correlation	95 % CI					
Race											
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

Table 2 (continued)

		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
SPF	0.96	*.48*[f]	.45–.51	.39[a]	.14	*.06*	19	1.4	**0.1**
EPT	0.43	.54[e]	.51–.56	.29[a]	*.07*	.08	22.5	**0.3**	**0.1**
AMP	0.16	.55[e]	.53–.57	.35[a]	.10	.16	*29.3*	1.1	*1.9*

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 the Feldt (1969) test; the test–retest values are correlations between the first and the second tests

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 none of the correlations were inordinately influential (see Table C3 in the supplemental materials).

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, and 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).

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

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

Average correlations were calculated after applying Fisher's transformation, and then were transformed back to correlation coefficients. **Bold** font = best performance in the relevant criterion; *underlined italics* font = worst performance in the relevant criterion. In the by-topic sections, correlations in each column that do not share a superscript are significantly different from each other

weakest relations). The main difference between the performance of the indirect measures in this criterion in comparison to the previous criteria 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, Brown-Iannuzzi, and Payne (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 all of 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

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*	–	–	–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}			

In each column of each topic section, correlations that do not share a superscript are significantly different from each other. The Thermometer and Items columns refer to difference scores. **Bold** = the strongest correlation of the relevant criterion; *underlined italics* = the weakest correlation of the relevant criterion. * The effect size of the race items-rating score is negative, indicating a preference for Black people over White people (opposite to the effect of the other direct measures)

associated with the category. The online supplemental materials provide more details about the computations.

Little research has focused on the quality of indirect measures in separate measurements of two attitude objects. Prior research found poor discriminant validity for single-attitude measurement with the IAT (Nosek et al., 2005), good discriminant validity for the ST-IAT (Karpinski & Steinman, 2006), and possible threats to nonrelative 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 a direct test between the measures that are constrained by relative measurements and the measures that, at least theoretically, seem to provide a separate measurement for each attitude object.

We tested the single-category scores for known-group effects, internal consistency, test–retest correlation, and the relationship to other indirect and direct measures of the same category. In addition, we looked at the 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 to 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 Rosenberg, MRS, and RWA were more strongly related to direct measures of Self, Black people, and Republicans than to direct measures of Other, White people, and Democrats, respectively (Table C4 in the online supplement materials).

Finally, and perhaps most importantly, 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 measure 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,

Table 5 Summary of single-category measurement criteria

	Reliability		Convergent Validity							
			Known-Groups		Correlation With Other Measures					
Overall	Alpha Cronbach	Test–Retest	Race	Politics	With Indirect	With Direct	With Rosenberg	With MRS	With RWA	Discriminant Validity
IAT	.77	.41	**1.03**	**1.37**	**.29**	**.26**	**.18**	–.25	.43	*0*
BIAT	.67	.53	0.65	1.27	**.29**	**.26**	.17	**–.27**	**.53**	*0*
GNAT	.64	*.29*	0.44	1.06	.27	.24	.15	–.26	.41	.06
ST-IAT	.76	.30	0.25	0.77	.23	.22	.14	–.25	.31	.07
SPF	*.44*	.34	0.54	0.82	.20	.17	*0*	–.21	.29	.04
EPT	.63	.32	0.36	*0.50*	.16	*.15*	*0*	–.21	.27	.05
AMP	**.82**	**.59**	*0.18*	*0.50*	*.12*	.21	.13	*–.18*	*.25*	**.09**

For 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** font = best performance in the relevant criterion; *underlined italic* font = worst performance in the relevant criterion

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.

We did not find an advantage for the AMP and the ST-IAT in predicting scales that were 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, whereas 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 did 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, separate direct evaluation would be more strongly related to measures that induce a similar context by contrasting the two categories than to 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 nonextreme attitudes

Participants with extreme attitudes may contribute to the psychometric qualities of measures more than do 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.

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 is 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 α = .66 to .10, and the average correlation between the AMP and the direct race measures declined from r = .31 to .13. Compare that with the race IAT's psychometric resistance to trimming the extreme scores: a small decrease from α = .86 to .78 in internal consistency, and from r = .27 to .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 percentages 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 % cutoff.

After the AMP, the ST-IAT was the measure 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 misbehaving participants or without the most well-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

In the present research, we compared the psychometric qualities of seven indirect attitude measures across three topics (racial attitudes, political attitudes, and self-esteem) using

Table 6 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 Alpha	Middle 90 %		All Cases *r*	Middle 90 %		All Cases *r*	Middle 90 %	
		Alpha	% Loss		*r*	% Loss		*r*	% Loss
Overall									
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
GNAT	.77	.59	19.5	.40	**.34**	3.5	.33	.29	3.2
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*

The average % loss is the average loss of shared variance. **Bold** font = best performance in the relevant criterion; *underlined italic* font = worst performance in the relevant criterion

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 nonextreme attitudes, and robustness to the exclusion of misbehaving or well-behaving 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) able to predict 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 of the attitude research based on indirect attitude measurement—either to increase the predictive validity of attitudes (complementing direct measures) or as a separate measure to assess nonexplicit evaluation—has employed only one indirect measure. Standard practice is to interpret the results of any indirect measures as being assessments 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 relationships amongst themselves (e.g., Bosson et al., 2000; Olson & Fazio, 2003; Payne, Govorun, & Arbukle, 2008). In the present study, we 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 interrelations between indirect measures, the strength and breadth of the present findings provides confidence that indirect attitude measures are interrelated, but that this relation varies across attitude domains. This finding reduces the concern raised by studies that have failed to find interrelations.

Variations across attitude domains

In the present study, we 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. Furthermore, 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 interrelations. Until further evidence, we can only speculate that the concept *self* is more multifaceted 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 effects of attitude topic on the 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 vs. 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 of 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 interrelations 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 for the assumption that indirect and direct measures of social cognition are sensitive to different theoretical constructs or different psychological processes.

Because much previous research has supported the assumption that indirect measures (more than direct measures) tap into implicit cognitions (e.g., Cameron et al., 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 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 interrelation correlations might reflect the lower reliability of most indirect measures (in comparison to direct measures) that prevents strong correlations among indirect measures. Alternatively, the present results might reflect variability in the methodological or theoretical sources of variance that influence indirect measures. We will 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 the GNAT (3.74), ST-IAT (4.26), SPF (4.39), AMP (5.04), and EPT (5.30).

At 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 at an 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 the 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, in a follow-up study (Bar-Anan & Nosek, 2013), we used stimuli selected especially for the AMP and compared them experimentally with the stimuli from the main study. In the follow-up study we 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 (as in the main study), we found that the stimulus set influenced the average preference scores 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 stimulus 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 those of the best-performing measures. As 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 measures that have received less empirical scrutiny than 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 those of the strongest performers. Additionally, the psychometric qualities of most measures were not very sensitive to the exclusion of extreme scores, or to the exclusion of well-behaved or misbehaving participants.

In the present research, we 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 those of any other measures in predicting single-category evaluation, or in predicting scales that were supposed to relate to one category evaluation more strongly 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 of 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's 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 of assessment.

IAT

Among the indirect measures, the IAT has earned its status as the most popular tool because of its 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, Bar-Anan, Sriram, & Greenwald, 2013; Sriram & Greenwald, 2009). In particular, although this test shares 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, both making it easier to learn how to do the task and allowing it to be completed with fewer trials, but this test 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, it has unique qualities. Its relatively good performance in the present study was surprising, considering that past evidence had 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 well as the IAT and BIAT. On the negative side, the present research showed 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 (i.e., not responding too quickly 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 highest 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 that are less tolerant of time pressure tasks. It might also mean that the GNAT is more sensitive to extraneous influences, such as individual differences in cognitive capacity, making it more difficult to compare across age groups (children as well as young and older adults) or other groups that could differ on these variables. Whether this is the case will require additional empirical evidence.

ST-IAT

The ST-IAT was developed to measure attitudes toward a single object in a noncomparative 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 relationships of the ST-IAT's preference score to other indirect measures and to direct measures were usually better than those for some of the measures (SPF, EPT, and sometimes AMP), and often not far behind those of the IAT, BIAT, and GNAT. In our comparison of single-category measurement quality, the ST-IAT showed better evidence for discriminant and convergent validity than did most other measures.

Because it is a relatively easy task, the ST-IAT may seem more vulnerable than other measures to nonautomatic processes (Stieger, Göritz, Hergovich, & Voracek, 2011). Indeed the ST-IAT had the lowest error rate of all of 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 to 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 measures more than to direct measures (Table 2). Additionally, the ST-IAT was usually the fourth-best measure on the two main validity criteria (relationships 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 use, 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 associations between blocks (IAT, GNAT, BIAT, and ST-IAT), as well as to the extraneous effects of block order that 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, in the present research we found that the SPF has worse psychometric qualities than all of the "blocked" measures. It was consistently superior only to EPT, and sometimes to 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 are 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 will 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, no empirical evidence has indicated 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—for a variety of research applications categorization tasks are not appropriate, and EPT may be the best available measure. Because EPT has low reliability, the use of this measure will be most effective by increasing statistical power via other means, such as using larger samples than would be necessary 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. In previous research it has shown good internal consistency (Payne et al., 2005) and good validity (Cameron et al., 2012), and it seems to have straightforward procedural validity: Attitudes affect performance, despite participants' intention to prevent this. 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. Also 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, still no research has investigated 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, however, 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), little evidence suggested 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 has been used in most past AMP research, and researchers should also examine whether the results are dependent on the extreme scores rather than being 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 being sensitive to automatic more than to 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 will require

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 had 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 carryover effects, including fatigue, loss of interest, improvement in performance, and improved understanding of the study's general purpose (attitude measurement). All of these factors may limit 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

For the present study, we compared seven indirect measures on a variety of psychometric qualities. We found strong evidence for interrelations among all of the 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 relative to the other measures. We believe that further multimeasure research could help us 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 constructs measured by indirect versus direct measures.

Author note This project was supported by grants from the European Union (PIRG06-GA-2009-256467), and the Israeli Science Foundation (1012/10) to Y. B.-A, and from Project Implicit Inc. to Y.B.-A and B.A.N.. B.A.N. is an officer of Project Implicit Inc., a nonprofit organization that provided financial and technical support to this project, and includes in its mission "To develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors."

References

Altemeyer, B. (1981). *Right-wing authoritarianism*. Winnipeg, MB: 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://openscienceframework.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). A multitrait-multimethod structural analysis of direct and indirect attitude measures. Unpublished raw 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,* 163–170.

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.

Feldt, L. S. (1969). A test of the hypothesis that cronbach's alpha or kuder-richardson coefficent twenty is the same for two tests. *Psychometrika, 34*(3), 363–373.

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. doi:10.1037/0033-2909.132.5.692

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 ed.). New York, NY: Cambridge University Press.

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

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

Greenwald, A. G., 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. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129,* 339–375. doi:10.1037/0033-2909.129.3.339

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. doi:10.1037/0096-3445.134.4.565

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, NY: 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., . . . 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, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York, NY: 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.