A Comparison of the Sensitivity of Four Indirect Evaluation Measures to Evaluative Information

Yoav Bar-Anan^{a*} and Brian A. Nosek^b

^aDepartment of Psychology, Ben-Gurion University of the Negeve in Beer-Sheva, Israel;

^bDepartment of Psychology, University of Virginia and Center for Open Science, USA.

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

Indirect evaluation measures are used as a dependent measure to assess the impact of experimental interventions on shifting pre-existing attitudes or creating new attitudes. In four experiments (total $N = 13,894$), we compared the sensitivity of four indirect evaluation measures to evaluative information about two novel targets. Evaluative sensitivity was strongest for the Implicit Association Test (IAT). Other measures were more similar in their sensitivity, but the pattern, from stronger to the weakest was the Affect Misattribution Procedure (AMP), the Sorting Paired Features (SPF), and then Evaluative Priming task (EPT). To the extent that these findings are generalizable to related research applications, these results suggest that the measures differ in their research efficiency (power). For example, to achieve 80% power to detect the evaluative learning effect in the present studies, direct self-report would require 10 participants, the IAT 28 participants, the AMP 57, the SPF 79, and the EPT 131.

Keywords: Attitudes; Evaluation; Indirect Measures; Implicit Measures; Automatic Evaluation

Author's note: This project was supported by grants from the Israeli Science Foundation [779/16] to Y. B.-A, and from Project Implicit Inc. and the United States – Israel Binational Science Foundation [2013214] 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.” The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Evaluation of objects as positive or negative is a ubiquitous human behaviour. People judge whether they like or dislike politicians, job candidates, potential partners for friendship or romance, food, appliances, and scientific manuscripts. The study of the factors that determine likes and dislike is central to social psychology (Ajzen & Fishbein, 1980; Anderson, 1965; Asch, 1946; Petty & Cacioppo, 1986; Walther & Langer, 2008). Experimental investigations of evaluation require an evaluation measure as a dependent variable. Self-reports of evaluation are common, partly because they are relatively easy to implement and partly because participants' explicit, deliberate evaluation is the variable of interest. Alternative measures are indirect in that participants do not need to consciously reflect on their evaluation for assessment (Nosek & Greenwald, 2009). Indirect measures of evaluation are used, partly because participants might otherwise report the evaluation they want to appear to have (Jones & Sigall, 1971; Leary & Kowalski, 1990), or because they may not be aware of all aspects of their evaluation. It is common to assume that indirect measures of evaluation reflect automatic evaluation whereas most self-report measures capture deliberate evaluation (De Houwer, 2006; Fazio & Olson, 2003; Nosek, Hawkins, & Frazier, 2011).

To improve the knowledge about the various indirect evaluation measures, research on their psychometric is essential. Such research examines the reliability and validity of the measures (e.g., Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Spruyt, Hermans, De Houwer, Vandekerckhove, & Eelen, 2007), investigates what psychological processes influence those measures (Blaison, Imhoff, Hühnel, Hess, & Banse, 2012; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Gawronski, Deutsch, LeBel, & Peters, 2008), and what modifications could improve them (Nosek,

Greenwald, & Banaji, 2005; Payne, Cheng, Govorun, & Stewart, 2005). Most psychometric evidence comes from studies focused on a single indirect measure. However, in reality, researchers must weigh the evidence comparatively to decide which indirect measures are most appropriate for their research. Comparing separate studies of each measure is fraught with procedural, design, and sampling confounds. There are some comparative studies of indirect measures, but all so far have been correlational designs (e.g., Bar-Anan & Nosek, 2014; Bar-Anan, Nosek, & Vianello, 2008; Bosson, Swann, & Pennebaker, 2000; Cunningham, Preacher, & Banaji, 2001; De Houwer & De Bruycker, 2007; Karpinski & Steinman, 2006; Krause, Back, Egloff, & Schmukle, 2011; Payne, Govorun, & Arbuckle, 2008; Ranganath, Smith, & Nosek, 2008; Rudolph, Schröder-Abé, Schütz, Gregg & Sedikides, 2008). For example, Bar-Anan and Nosek (2014) measured attitudes about race, politics, and the self with seven indirect measures, and compared the measures on reliability and relations to direct evaluation measures, and to the other indirect measures.

In the present research, we compared the validity of indirect evaluation measures experimentally. We manipulated evaluative information about two novel target objects, and then measured the automatic evaluation of those objects. We compared four indirect measures: The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007), the Evaluative Priming Task (EPT; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Klauer, & Musch, 2003), the Affect Misattribution Procedure (AMP; Payne et al., 2005; Payne & Lundberg, 2014), and the Sorting Paired-Features task (SPF; Bar-Anan, Nosek, & Vianello, 2009). IAT, EPT and the AMP are the three most commonly used indirect evaluation measures (Nosek,

Hawkins, & Farzier, 2012). We included the SPF because, for some research purposes, especially in the domain of evaluative learning, it is a promising alternative to the IAT because it measures all the task conditions in the same block, and it has previously outperformed EPT (Bar-Anan et al., 2008; Bar-Anan & Nosek, 2014).

In a previous correlational comparison, the ranking of the four measures by their psychometric qualities was IAT, SPF, AMP, EPT (Bar-Anan & Nosek, 2014). However, in a follow-up study with a larger number of AMP trials, measuring only race attitudes, the AMP was similar to the IAT in most criteria, and showed the strongest relation with self-reported attitudes (Bar-Anan & Nosek, 2013). In a study that included the IAT, AMP, and EPT (Payne, Govorun, & Arbuckle, 2008) in a measurement of attitudes toward alcohol, the AMP was the only measure with reliable relations with six reported drinking behaviours, whereas the IAT and EPT were each related reliably only to one reported behaviour. However, the sample size was too small to reach reliable conclusions about comparative performance. Finally, in a direct comparison of the SPF and EPT for measurement of political attitudes, the SPF showed better reliability and validity than the EPT (Bar-Anan, Nosek, & Vianello, 2008).

The experimental induction of attitudes toward novel target objects for comparative purposes in the present studies has three main strengths. First, an experimental design provides better control of the variance of the evaluations, and is not reliant on pre-existing variation in evaluations. We provided evaluation information that ranged from one extreme (an extremely more positive information about object A than about object B) to another (an extremely more positive information about object B than

about object A), with equal likelihood for any extremity. This distribution is ideal for assessing measurement sensitivity.

Second, in correlational studies, the evidence depends on the reliability and validity of the criterion measures. For instance, Bar-Anan and Nosek (2014) found that the AMP was comparatively weakly related with other indirect measures. That could indicate a weakness of the AMP or that other indirect measures are all confounded by a non-evaluative construct that inflated their inter-correlations. In the present study, we examined how much variance each indirect measure shared with the evaluative information that we provided experimentally providing an independent assessment of measurement quality and sensitivity.

A third strength of the present research is that it is uniquely relevant for research on evaluative learning. Research on impression formation (Ratliff & Nosek, 2010; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007), persuasion (Horcajo, Briñol, & Petty, 2010; Smith, De Houwer, & Nosek, 2013), and other forms of evaluative learning (Olson & Fazio, 2001; Stahl & Unkelbach, 2009) often uses novel objects as stimuli and indirect measures as the dependent variable. Previous comparative studies of indirect measures addressed concepts that participants have known for days, weeks, and years. Knowledge and evaluation of such concepts is different from knowledge formed based on very recent information. For example, recently formed memories have not been fully consolidated (McGaugh, 2004; Stickgold & Walker, 2005). Therefore, newly formed attitudes might depend on unique retrieval and judgment processes. Perhaps indirect measures of evaluation vary in their relative sensitivity to different retrieval processes. If that is the case, previous correlational studies are of less relevance than the

present research to the reliability and validity of indirect evaluation measures in the context of evaluative learning research.

Overview of Experiments

We used the same procedure in all four experiments. Participants read eight behaviours, four positive and four negative. Four behaviours were attributed to a man named Kevin, and four attributed to a man named Brian. We varied between participants the number of positive and negative behaviours attributed to each man. Theory and empirical evidence suggest that just like deliberate evaluation, automatic evaluation has a straightforward reaction to such information: more positive behaviours cause more automatic liking (Bar-Anan & Nosek, 2012; Ratliff & Nosek, 2010; Rydell et al., 2007; Nosek, 2005; Smith et al., 2013). That manipulation is expected to cause clear effects on automatic evaluation regardless of the measure.

We considered inducing automatic evaluation that is distinct from deliberate evaluation because researchers often use indirect measures to tap a construct distinct from deliberate evaluation. However, there is currently no well-replicated method to experimentally induce automatic evaluation that is distinct from deliberate evaluation. In recent research that demonstrates this problem, Moran, Bar-Anan, and Nosek (in press) used an evaluative learning procedure that had a discrepant effect on indirect and direct evaluation measures (Gregg, Seibt, & Banaji, 2006; Peters & Gawronski, 2011). In three pairs of studies, each pair used a different indirect measure. Each measure showed different results, and only one measure showed evidence of discrepancy from deliberate evaluation. In the present research, although we provided straightforward evaluative

information about the targets, for exploration, we still examined whether the indirect measures captured any effect not captured by the deliberate evaluation.

In each of the present experiments, after providing the evaluative information, we measured participants' attitudes toward Kevin and Brian, using one indirect measure, and a direct self-report measure. We compared all four measures in Experiment 1. As replications with a few methodological improvements, we compared the AMP to the IAT in Experiment 2, the SPF to EPT in Experiment 3, and all four measures in Experiment 4¹.

Method

Participants

Participants were volunteers at the Project Implicit research website (Nosek, 2005), who registered for research and were randomly assigned to the study from a pool of available studies. In Experiments 1-3, we planned to run about 1,000 participants for each indirect measure in each study to avoid exhausting the shared resource of participants in Project Implicit (<https://implicit.harvard.edu/>). In Experiments 1-3, power was about 80% for detecting a correlation difference between .4 and .5, but only 65% to detect a difference between .15 and .25. In Experiment 4, we planned to recruit at least 6000 participants (around 1,500 participants for each measure), to achieve 90% power to detect a correlation difference between .2 and .3.

¹ We report all data exclusions, manipulations, and measures, and how we determined sample sizes. Decisions to stop collecting data did not depend on the obtained results. Materials, data and links to experience the studies are in <https://osf.io/sn9jp/>.

The total number of participants who responded to all the self-report questions and completed the indirect measure was 4,041, 2,145, 1,949, and 6,406 in Experiments 1-4, respectively ($M_{age} = 34.88, 39.03, 36.24, 35.37$; $SD_{age} = 13.96, 16.79, 15.26, 14.56$; $N_{women} = 2,621, 1,395, 1,267, 3,885$)². Of those, the analyses did not include participants whose performance in the indirect measure suggested lack of ability or motivation to perform the task properly. The exclusion rules were those recommended in previous work with the present data source, for the IAT (Greenwald, Nosek, & Banaji, 2003), SPF (Bar-Anan & Nosek, 2014), and EPT (Bar-Anan, 2010; Bar-Anan & Nosek, 2014; Bar-Anan, 2016), and the rules recommended based on various participant populations for the AMP (Payne & Lundberg, 2014). Specifically, the analyses did not include participants who responded faster than 300ms in more than 10% of the IAT trials (exclusion rates: 7/1,043, 3/1,073, 21/1,653 in Experiments 1, 2, and 4), or the SPF trials (13/976, 30/957, 37/1,454 in Experiments 1, 3, and 4), participants who had less than 40% success-rate in EPT (5/1013, 7/992, 11/1,579 in Experiments 1, 3, and 4), and participants who used only one of the two responses throughout the AMP, or reported having any knowledge of Chinese (108/1006, 120/1,067, 199/1,639 in Experiments 1, 2, and 4). In total, we included in the analyses 3,902, 2,017, 1,912, and 6,063 participants in Experiments 1-4, respectively.

² 7,993 participants started Experiment 1, but some data of 1,389 participants was lost for technical problems in the data recording, and 2,564 participants did not finish the study (a drop-out rate of 39%, which is a typical dropout rate in Project Implicit's volunteer-based participant pool). In Experiment 2-4, the dropout rates were 1,346/3,491, 1,310/3,259, 2,849/9,234, respectively.

Procedure

The experiments were administrated over the Internet using PIPE (Version 0.3; Zlotnick, Dzik, & Bar-Anan, 2015). Participants who previously registered for the Project Implicit participant pool logged in voluntarily and were randomly assigned to complete this study from a pool of available studies. Once assigned to any of the four experiments, they could not be assigned to any of the four experiments on future visits.

The consent form stated that studies at Project Implicit examine attitudes, preferences, and memory. Participants observed the learning task, completed a questionnaire about their evaluation of the two target men, a questionnaire about their evaluation of each behaviour that was presented in the learning task (removed from Experiment 4), and an indirect measure of evaluation toward the two targets. The behaviour rating questionnaire appeared before the learning task or at the end of the experiment (randomized between participants). The direct and indirect evaluation measures always appeared after the learning task, in random order. Participants were randomly assigned to complete one indirect measure.

The instructions before the learning task were: *In the next task, you will get to know two men. We will tell you a few facts about each of these men. Later, we will ask you a few questions about what you think of them.* The learning task had eight trials. On each trial, the man's image, his name, and a behaviour he performed were displayed together for 5 seconds, followed by a 750ms blank screen. The trials alternated from one man to the other, for a total of four trials per man. For each participant, four positive and four negative behaviours were randomly selected from a pool of 23 positive and 23 negative behaviours. The five between-subjects randomly-assigned conditions were a

ratio of positive to negative behaviours attributed to Brian of 0:4, 1:3, 2:2, 3:1, and 4:0, always mirrored by the ratio of positive-to-negative behaviours attributed to Kevin (4:0, 3:1, 2:2, 1:3, and 0:4). All procedures are available for self-administration at osf.io/jxpb8 to review details of the methodology.

Measures

Self-reported evaluation of the behaviours. For each behaviour, the participants answered the question *How positive or negative is this behaviour?* on a 9-point scale with the options 'Extremely positive', 'Very positive', 'Moderately positive', 'Slightly positive', 'Neutral', 'Slightly negative', 'Moderately negative', 'Very negative', and 'Extremely negative'. The eight questions were presented one at a time, in a random order, with no mention that the men performed them (approximately half of the participants completed this questionnaire before being presented with the men). We used those ratings to create a corpus of behaviour ratings (see Table S4 in the web supplement), and do not discuss this measure any further in the main text.

Self-reported evaluation of the men. In Experiments 2-4, participants answered questions in the format of “How trustworthy is Kevin?”, for the traits *friendly*, *trustworthy*, and *likable*, on a 9-point scale ranging from 1 – Extremely *unfriendly* to 9 – Extremely *friendly*. In Experiment 1, as a part of a methodological investigation of question formats, each participant saw the questions in one of six formats (e.g., the scale responses appeared horizontally or vertically; one at a time in separate pages or all in the same page), all with identical or very similar wording and five of them with 9-point response scales (a sixth used visual analogue scale that displayed only the most extreme

labels on each size of the scale, and was divided to 201 responses). This methodological investigation is not discussed further.

Affect Misattribution Procedure. Each trial in the AMP displayed stimuli in the following sequence: A photograph of one of the two targets appeared for 100 ms, followed by a blank screen for 100 ms, followed by a Chinese pictograph for 100 ms, followed by a pattern mask of black-and-white noise that appeared until participants responded. Upon presentation of the mask, participants indicated if the Chinese pictograph is more pleasant or more unpleasant than the average Chinese pictograph using two response keys on the keyboard signifying *pleasant* and *unpleasant*. Following Payne et al. (2005), the instructions informed participants the pictures appearing before the Chinese pictographs may bias responses and that they should not let the pictures influence their judgments. The task started with three example trials and then participants completed three blocks of 40 trials, each with 20 primes for each man.

Evaluative Priming Task. In Experiments 1 and 3, each trial in the EPT started with a prime (i.e., a photograph of Brian or Kevin) displayed for 200 milliseconds; then a positive or negative target word appeared until participant categorized it as positive or negative using one of two response keys on the keyboard. If the response was incorrect, a red “X” appeared for 750 milliseconds. Last, a blank screen appeared for an inter-trial interval (ITI) randomly selected from a range of 300ms to 900ms. In comparison to the other measures, there is less consensus about the procedural parameters of the EPT. However, inspection of recent literature found only minor differences. In Experiments 1 and 3 we used parameters from our previous work (Bar-Anan, 2009; Bar-Anan & Nosek, 2014; Moran, Bar-Anan, & Nosek, in press). In Experiment 4, we used parameters from

research by Bertram Gawronski, who has published many recent evaluative learning papers using the EPT (specifically, we used parameters from Gawronski, Gast, & De Houwer, 2015). The modifications relative to the procedure in the previous experiments were that the trial started with a 500ms display of a fixation stimulus before the prime, the error feedback appeared for 1500ms, and the ITI was fixed to 500ms.

The rationale behind the EPT is that liked primes facilitate the classification of positive targets, and impede the classification of negative targets; whereas disliked primes facilitate the classification of negative targets, and impede the categorization of positive targets. Therefore, the automatic preference for one prime over another is inferred by comparing the facilitation in responses to positive targets versus negative targets after one prime versus the other prime. In Experiments 1 and 4, the EPT consisted of three blocks of 60 trials (15 trials for each prime-target combination). In Experiment 3, each block consisted of 40 trials (10 trials for each prime-target combination). Participants were instructed to categorize the target words as quickly and accurately as they can, and ignore the image that appeared before them.

Implicit Association Test. In the IAT, participants categorize stimuli using two computer keys. In each trial, the target stimulus remained on screen until correct response. Participants had to fix incorrect responses with the correct response. The intertrial interval was 250ms. In the critical blocks, participants respond with the left key to stimuli of two categories (e.g., "Kevin" and "Good words"), and with the right key to stimuli of two other categories (e.g., "Brian" and "Bad words"). In two of these blocks, Kevin and Good words shared the same response key, and in the other two critical block, Brian and Good words shared the same response key. When "Kevin" and "Good words"

shared the same key, people with more positive associations for Kevin than Brian should respond more quickly. Therefore, the difference between the average response latencies in blocks that assigned Kevin and Good words to the same key and blocks that assigned Brian and Good words to the same key is interpreted as automatic preference. The seven-block IAT followed the procedure described in Nosek, Greenwald, and Banaji (2005). The IAT had 120 critical trials for scoring the Brian-Kevin preference score.

Sorting Paired Features Task. In the SPF, two pairs of category labels appear in each of the four corners of the screen. In each trial, participants use four keyboard keys to sort pair stimuli (e.g., an image of Kevin and the word *wonderful*) to one of four category pairs (e.g., *Kevin+pleasant*, *Kevin+unpleasant*, *Brian+pleasant*, and *Brian+unpleasant*). Faster sorting indicates a stronger association between the pair categories. Therefore, a person who prefers Kevin over Brian will be faster on trials that present Kevin with pleasant words or Brian with unpleasant words than on trials that present Kevin with unpleasant words or Brian with pleasant words. We used three blocks of 40 trials. In each trial, the target stimulus appeared until correct response. Participants had to fix incorrect responses with the correct response. The intertrial interval was 250ms.

Materials

Target men. In Experiment 1, one photo of a White man was always Brian, and another photo of a White man was always Kevin (the photos were taken from Minear & Park, 2004). In Experiments 2 and 4, to increase the stimulus sample, for each participant, two of 12 possible White man photos (from Payne, 2001) were randomly assigned to each man. In Experiment 3, the two images were selected from ten photos of Black men (from Gawronski, Cunningham, LeBel, & Deutsch, 2010). For the indirect

measures, we used five stimuli for each man: *KEVIN*, *kevin* [*BRIAN*, *brian*] and three image stimuli: the photo used in the induction, a vertical flip of the photo (mirror image), and a grey-scale version of the photo.

Evaluative stimuli. In Experiment 1-3, word selection was based on previous work on the Project Implicit website (Bar-Anan & Nosek, 2014; Nosek et al., 2007). The EPT used the same words as the SPF (*Pleasure, Love, Wonderful, Paradise, Cheer, Splendid, Bomb, Abuse, Sadness, Pain, Poison, Grief*), with the addition of 16 words (*Friend, Glee, Smile, Enjoy, Delight, Beautiful, Attractive, Likeable, Ugly, Dirty, Stimtmnnk, Noxious, Humiliate, Annoying, Disgusting, Offensive*). The words used in IAT were *Pleasure, Love, Wonderful, Laughter, Happy, Glorious, Joy, Peace, Awful, Failure, Agony, Hurt, Horrible, Terrible, Nasty, and Evil*. In Experiment 4, the SPF, EPT, and the IAT all used the same words stimuli taken from Gawronski et al. (2015): *pleasant, good, outstanding, beautiful, magnificent, marvelous, excellent, appealing, delightful, nice, unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful, ugly*. For the IAT, SPF, and EPT, the labels of the attribute categories were *Pleasant words* and *Unpleasant words* in Experiments 1-3, and *Positive words* and *Negative words* in Experiment 4.

Neutral stimuli. The Chinese pictographs in the AMP were retrieved from Keith Payne's website (<http://bkpayne.web.unc.edu/research-materials/>).

Behaviours. The behaviours were selected randomly for each participant from a pool of 23 positive and 23 negative behaviours (see web supplement), collected from previous impression formation research (Bar-Anan & Nosek, 2012; Ranganath & Nosek, 2008).

Results

Data Processing

For each measure, we computed a preference for Brian over Kevin.

EPT. Because there are many variations for computing EPT scores, we selected the scoring method that showed the best performance out of 1,536 scores created by enumerating a number of scoring parameters (Bar-Anan, 2016). We deleted error responses and replaced latencies longer than 800ms with 800ms (winsorizing). For each of the three blocks, we then subtracted the average log-transformed latencies of conditions compatible with preference for Brian (*Brian-good*; *Kevin-bad*) from the average log-transformed latencies of the conditions incompatible with that preference (*Kevin-good*; *Brian-bad*). We then divided each block's difference score by the standard deviation of the latency in that block. The preference score was the average of the standardized scores of the three blocks. Because this scoring has never been used before, we also report two scores that were used in the research from which we adapted the EPT procedures. Bar-Anan and Nosek (2014) used a similar scoring method, but instead of winsorizing slow responses they deleted trials with response latencies faster or slower more than 2 standard deviations from the average latency of the trial's prime-target condition. Gawronski et al. (2015) used raw latencies (rather than log-transformation), deleted error-trials and trials with response latencies longer than 1,500ms (no winsorizing), computed one score (rather than averaging across blocks), and did not divide it by standard deviation.

AMP. The proportion of *pleasant* judgments after Kevin were subtracted for the proportion of *pleasant* judgments after Brian.

IAT. We used the D600 algorithm recommended by Greenwald et al. (2003). The overall IAT *D* score was an average of two *D* scores computed for each pair of critical blocks (blocks 3 versus 6, and blocks 4 versus 7). The *D* score for each pair of blocks was the difference between the average latency of each of the two opposing blocks (e.g., block 3 versus block 6) divided by the standard deviation of the trials in both blocks together. A positive *D* score indicated faster responding on average when Brian was paired with good words and Kevin with bad words compared to the reverse, interpreted as a preference for Brian compared to Kevin. *D* was calculated after removing response latencies under 400 ms or over 10000 ms, and included all other trials. Latencies of error responses were replaced with the block mean of correct latencies plus 600 ms.

SPF. We tested two scores for the SPF. One was recommended by Bar-Anan et al. (2009): A *D* score for each of the four trial conditions was computed within each block as the difference between the average latency of the trial condition and the average latency in the whole block, divided by the overall standard deviation of that block. Then, single-attitude object scores were computed as the difference between the good and bad *D* scores of each category (e.g., the difference between the *D* scores of *Kevin-Good* and *Kevin-Bad*). A preference score for each block was computed as the difference between the two single-category scores. The SPF preference score was the average of the preference scores of the three blocks. We also explored a modification of this scoring method, copying the same method but using log latencies instead of raw latencies and winsorizing of latencies faster than 300ms or slower than 3,300ms.

Figure 1. Experiments 1-4: Vertical probability density (violin) plots of the standardized preference score (for Brian over Kevin) of each measure by the number of positive behaviours attributed to Brian (x-axis). Standardization was across all induction conditions (for each measure in each study). The black circles are the mean standardized preference, and the error-bars are 95% confidence intervals around each mean.

Direct evaluation of the men. The preference score was the difference between Kevin's average trait rating and Brian's average trait rating. In Experiment 1, the ratings were standardized within each question format condition because response values on the VAS ranged from -100 to 100.

Behaviour Evaluation

In Experiments 1-3, all the positive behaviours were rated as more positive than all the negative behaviours, with a minimum effect size of $ds = 1.80, 1.68, 2.01$.

Sensitivity to Evaluative Information

Figure 1 presents the average standardised preference as a function of number of positive behaviours attributed to Brian, for each measure in each experiment. In the web supplement, Figure S1 shows the same graph with unstandardized preferences. In Experiment 1, the measures were slightly biased by pre-existing preference for the combination of the name 'Brian' and the image selected for Brian over 'Kevin' and his image. In the other experiments, because the face images were selected randomly from a pull of images, pre-existing attitudes did not influence the overall means.

Table 1 presents the internal consistencies of the measures and their inter-correlations. The measure of sensitivity to the induction was the correlation between each measure's score and the number of positive behaviours performed by Brian (*Induction Condition* in Table 1). Higher correlations indicate strong sensitivity to the evaluative information. For statistical inference from these results, we compared the Pearson correlations (Table 1). The IAT outperformed all the other measures in each experiment. The statistics for the comparison of the correlations of the induction condition with the IAT versus with the second best measure (AMP, AMP, SPF) were $Zs = 2.73, 4.01, 5.84$, $ps = .006, .0001, < .0001$ in Experiments 1, 2, and 4, respectively.

The AMP was significantly more sensitive than the SPF in Experiment 1, $Z = 2.29, p = .021$, but numerically less sensitive in Experiment 4, $Z = 0.92, p = .358$. We speculate that the SPF improved because of the word stimuli used in Experiment 4 (the IAT and the EPT also showed better sensitivity using those words). Although the SPF showed numerically stronger correlations than the EPT in Experiments 1, 3, and 4, the

differences reached statistical significance only in Experiment 4, $Z = 3.49$, $p = .0005$, and not in Experiments 1 and 3, $Z_s = 0.82$, 1.15 , $p_s = .409$, $.248$.

Table 1. The correlations of the evaluation measures with evaluative information and self-report

	IAT	AMP	SPF	SPF _w	EPT	EPT _{BN}	EPT _G	Explicit
Experiment 1								
N	1,036	898	963	963	992	1,008	1,010	3,902
Internal consistency	.71 _c	.85 _b	.41 _d	.48 _d	.29 _e	.33 _e	.19 _f	.91 _a
Induction condition	.467 _b	.364 _c	.248 _d	.268 _d	.233 _{de}	.219 _{de}	.181 _e	.756 _a
Explicit preference	.525 _a	.449 _b	.251 _c	.281 _c	.279 _{cd}	.253 _{cd}	.188 _d	-
Experiment 2								
N	1,070	947	-	-	-	-	-	1,980
Internal consistency	.78 _c	.87 _b	-	-	-	-	-	.93 _a
Induction condition	.517 _b	.373 _c	-	-	-	-	-	.762 _a
Explicit preference	.565 _b	.474 _c	-	-	-	-	-	-
Experiment 3								
N	-	-	927	925	933	983	985	1,898
Internal consistency			.33 _c	.46 _b	.12 _d	.32 _c	.07 _d	.94 _a
Induction condition			.234 _{bc}	.285 _b	.235 _{bc}	.192 _c	.176 _c	.773 _a
Explicit preference			.251 _{ab}	.293 _a	.263 _{ab}	.236 _{ab}	.208 _b	-
Experiment 4								
N	1,632	1,446	1,417	1,415	1,518	1,563	1,567	6,063
Internal consistency	.75 _c	.84 _b	.47 _e	.57 _d	.29 _f	.43 _e	.25 _f	.93 _a
Induction condition	.540 _b	.349 _c	.337 _c	.373 _c	.257 _d	.231 _{de}	.172 _e	.790 _a
Explicit preference	.564 _a	.449 _b	.375 _c	.421 _{bc}	.280 _d	.257 _{de}	.192 _e	-
Avg. with induction	.508	.362	.274	.309	.242	.214	.176	.770
N for 80% power	28	57	101	79	131	168	250	10
N for 99% power	61	130	234	182	242	391	583	20

Notes. SPF_w is the score based on winsorized latencies. EPT_{BN} is the EPT score recommended by Bar-Anan and Nosek (2014); EPT_G is the score adapted from Gawronski et al. (2015). All the correlations are significantly different from zero; In each row, a different subscript indicates a significant difference ($p < .05$, two-sided), computed following instructions from Zou (2007) for correlations and from Feldt (1969) for internal consistency; Internal consistency was computed from three parcels; Power to detect each correlation was computed for the average correlations, with $p = .05$, two-tailed.

In the web supplement, we report two alternative tests for significant differences: one based on Spearman correlations and the other based on multiple regressions. The results were similar with one exception: Spearman correlations harmed the AMP more than the other measures, increasing the IAT's advantage over the AMP in Experiments 1, 2, and 4, eliminating the AMP's advantage over the SPF in Experiment 1, and showing a significant inferiority in comparison to the SPF in Experiment 4. Because the Spearman correlation is based on ranking instead of exact scores, the decrease in the AMPs performance might suggest that the AMP's overall sensitivity is more dependent on extreme scores than other measures, implying lower sensitivity for scores in the less extreme range.

Predicting induction condition above and beyond self-reported preference

We explored whether each of the indirect measures shared any variance with the induction condition that was not shared with the self-report. Table 2 shows the results of two analysis strategies. One was a multiple regression predicting the induction condition (number of positive behaviour attributed to Brian) from the indirect measure and the self-report measure. The regression tested whether the indirect measure added any explained variance to the model. Another analysis was the same multiple regression computed as a structural equation model (SEM), specifying measurement models for each predictor, to predict the outcome variable from the latent variables, thus clearing the error-variance associated with each factor (Westfall & Yarkoni, 2016). This analysis tested whether the explained variance gained by adding the indirect measure to the model can be attributed to a theoretical construct different than the construct measured with the self-report measure (as opposed to attributing the gain to limits in the reliability of the self-report measure). Based on previous research and theory pertaining to the effects of simple evaluative information, we expected no difference in the effect of the present manipulation on the two constructs. However, because the reliability of the self-report was high but not perfect, it was reasonable to assume that highly reliable indirect measure could improve prediction in the multiple regression analysis.

Table 2

Standardised parameter estimates from simultaneous prediction of the induction condition from direct and indirect measures (in parentheses: p-value)

Experiment	Indirect Measure	Multiple Regression		SEM	
		Indirect Measure	Direct Measure	Indirect Measure	Direct Measure
1	IAT	.096 (< .0001)	.707 (< .0001)	.063 (.073)	.751 (< .0001)
1	AMP	.019 (.413)	.767 (< .0001)	-.021 (.442)	.818 (< .0001)
1	SPF	.057 (.009)	.742 (< .0001)	.076 (.053)	.754 (< .0001)
1	EPT	.023 (.296)	.732 (< .0001)	.022 (.320)	.760 (< .0001)
2	IAT	.124 (< .0001)	.690 (< .0001)	.112 (< .0001)	.713 (< .0001)
2	AMP	.020 (.410)	.751 (< .0001)	-.011 (.687)	.789 (< .0001)
3	SPF	.068 (.002)	.761 (< .0001)	.093 (.014)	.743 (< .0001)
3	EPT	.031 (.148)	.757 (< .0001)	.023 (.258)	.802 (< .0001)
4	IAT	.140 (< .0001)	.698 (< .0001)	.135 (< .0001)	.722 (< .0001)
4	AMP	-.006 (.740)	.792 (< .0001)	-.038 (.071)	.831 (< .0001)
4	SPF	.046 (.009)	.788 (< .0001)	.032 (.262)	.799 (< .0001)
4	EPT	.039 (.018)	.783 (< .0001)	.008 (.634)	.811 (< .0001)

*Notes. Each multiple regression predicted the induction condition from the indirect measure and the self-report measure; The SEM analysis was conducted with R's package lavaan (version 0.5-20, Rosseel, 2012); In **bold**: significant prediction for the indirect measure.*

The SEM analysis used three indicators for each predictor – three parcels for each indirect measure (using three thirds of the relevant trials for each task, divided sequentially; EPT scores were computed with the winsorizing-based score), and the self-reported preferences on the three traits (trustworthiness, likeability, friendliness). The IAT had the best performance, significantly improving prediction in all three experiments. In two of the three experiments, the IAT's shared variance with the self-report remained significant even after removing error-variance with SEM. The SPF showed significant improvement of the prediction in all three experiments. With SEM, the SPF's improvement remained significant in only one of the three experiments. There

was no statistically significant evidence that the EPT and the AMP improved the prediction in any of the experiments.

General Discussion

Four experiments compared the sensitivity of four indirect evaluation measures to clear explicit evaluative information. The results revealed that the IAT was the most sensitive. Averaged across experiments, the AMP showed the second-best sensitivity. However, it was not superior to the SPF in Experiment 4, and was somewhat inferior to the SPF when using ranked values (Spearman correlation) rather than raw data (Pearson correlation) for testing sensitivity. Therefore, the AMP's advantage over the SPF is less certain. EPT was always inferior to the IAT and the AMP, but the SPF was significantly stronger than the EPT only in Experiment 4³. The superiority of the IAT and the inferiority of the EPT are mostly compatible with the evidence collected earlier in comparative and non-comparative studies using those measures (e.g., Bar-Anan & Nosek, 2014). The SPF performed slightly better than in the past, especially in Experiment 4, whereas the AMP performed somewhat worse, especially in Experiment 4.

All studied indirect measures were sensitive to evaluative information, but differed substantially in the sample sizes required for reliably detecting effects. Table 1 provides post-hoc computations of sample-size required to achieve 80% and 99% power to detect a statistically reliable effect (a correlation with the induction) for each measure. For equal probability to detect the same effect the measures require substantially different numbers of participants. Taking 99% power, for example, the self-report measure requires just 20 participants, whereas the indirect measures require, 61 for the IAT, 130 for the AMP, 182 for SPF, and 242 for EPT. These dramatic differences have significant implications for research efficiency. If generalizable, using EPT requires 300% more participants than the IAT to conduct equivalently high-powered research in this context.

As Table 1 shows, the SPF and the EPT performed much better when we used scoring methods that we identified as high performing approaches, but are different than those used in previous investigations. Particularly, it seems that winsorizing slow

³ In a correlational comparative study, Bar-Anan and Nosek (2014) found that the Brief-IAT (BIAT; Sriram & Greenwald, 2009)—a recent modification of the IAT—showed psychometric qualities comparable or better than the IAT. An unpublished experiment (Bar-Anan & Nosek, 2016) found that the BIAT is more sensitive than the AMP to direct evaluative information.

responses improves performance in both cases. In future work, we will investigate that possibility with a wider sample of populations, settings, stimuli, study designs, and attitude domains. For now, we recommend researchers to explore winsorizing-based scores, when using the SPF and the EPT (see Richetin, Costantini, Perugini, & Schönbrodt, 2015, for initial findings that winsorizing might also be useful for scoring the IAT).

The IAT and the SPF had unique variance with the evaluative information above and beyond the variance captured by the self-report measure. The SEM analyses found some evidence for divergent validity, suggesting that evaluative information has unique influence on constructs captured by the IAT and the SPF, not captured by the self-reported evaluation. This might suggest that even straightforward evaluative information causes some discrepancy between deliberate and automatic evaluation. As a speculation, perhaps automatic evaluation is more sensitive than deliberate evaluation to simple statistical contingencies, and less sensitive to information integration processes (Ratliff & Nosek, 2010; Rydell et. al, 2007). The fact that the AMP and the EPT did not capture that unique variance might suggest that the IAT and SPF are more suitable for studying discrepancies between automatic and deliberate evaluation (Bar-Anan & Nosek, 2012). However, it is premature to conclude that that unique variance is evaluative. Perhaps the IAT and the SPF showed sensitivity to non-attitudinal effects of the induction procedure. More generally, we manipulated evaluative information as a test of validity of the measures because it is known that variance in evaluative information influences evaluation. Nevertheless, variance in evaluative information might also influence non-evaluative constructs. Therefore, it is not guaranteed that better measurement sensitivity to the present manipulation always reflected better sensitivity to evaluation. Like most research, the interpretation of the present results could change depending on future findings; in the present case, this would occur if future research finds that evaluative information influences a non-evaluative construct captured by the IAT and the SPF.

Of the four indirect measures, the AMP was the only that showed considerable decline in its sensitivity when the correlation was computed from ranked values (Spearman correlation; Table S1) rather than raw values (Pearson correlation; Table 1). This decline might suggest that the AMP's psychometric quality depends on extreme

values more than the other indirect measures. Supporting that interpretation, in Figure 1 the distributions of the AMP are quite similar regardless of the induction condition, and the differences in the average are mostly the result of wide tails in each distribution. That finding provides converging support for previous finding in a comparative research using pre-existing attitudes (Bar-Anan & Nosek, 2014): After removing the most extreme values, the AMP showed much sharper decline in psychometric qualities than any of the other six indirect measures tested in that study.

In the present research, one possible reason for the superiority of the IAT and the relatively good performance of the SPF is that these two measures are designed specifically for the measurement of preferences between two attitude objects. The AMP and the EPT have the advantage of being able to measure evaluations of more than two objects. There is currently only initial evidence that the SPF (Bar-Anan et al., 2009) or IAT variants (Axt, Ebersole, & Nosek, 2014) can provide valid measures for more than two objects. Further research is needed for comparing the quality of indirect measures in capturing attitudes toward more than two objects.

The present research focused on sensitivity to evaluation. There are other important considerations in the selection of indirect measures that the present research has not addressed. Perhaps the most important features are the ability to capture automatic evaluation, predict behaviour, and overcome faking efforts. We are unaware of any research that directly compared the sensitivity of the measures we tested to faking efforts. In the only known comparative faking study, participants were more successful in faking the Single-Target IAT (Karpinski & Steinman, 2006) than the IAT (Stieger, Göritz, Hergovich, & Voracek, 2011). Research that compared different indirect measures on predicting automatic and non-automatic behaviour is scarce and has always suffered from small sample sizes that prevented strong conclusions (e.g., Payne et al., 2008; Spruyt et al., 2007). So far, we have not been able to conduct such a study because we have not found any behavioural measure that consistently shows stronger relations with automatic evaluation than with deliberate evaluation.

Summary

For investigating sensitivity to experimentally formed attitudes, we found clear evidence that the IAT is more sensitive than the AMP, SPF, and EPT, and relatively clear

evidence that the AMP is more sensitive than the EPT. Because these measures are important tools in investigating attitudes and automatic evaluation, further comparative studies about the relative psychometric qualities of these measures and about their sensitivity to various forms of evaluative information are essential.

References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, 70(4), 394-400.
- Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3), 258-290.
- Axt, J. R., Ebersole, C. R. & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, 25, 1804-1815.
- Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für experimentelle Psychologie*, 48(2), 145-160.
- Bar-Anan, Y. (2010). Strategic modification of the evaluative priming effect does not reduce its sensitivity to uncontrolled evaluations. *Journal of Experimental Social Psychology*, 46(6), 1101-1104.
- Bar-Anan, Y. (2016). Exploring Scoring Algorithms for the Evaluative Priming in Evaluative Learning Experiments. *Unpublished Manuscript*. <https://osf.io/7yzzu/>
- 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(9), 1194-1208.
- 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. *Unpublished Manuscript*. <https://osf.io/dspar/>.
- Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior research methods*, 46(3), 668-688.

- Bar-Anan, Y., & Nosek, B. A. (2016). An Experimental Comparison of the Affect Misattribution Procedure and the Brief Implicit Association Test. *Unpublished manuscript*. <https://osf.io/4yn3r/>
- Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2008). A comparison between the SPF and evaluative priming. *Unpublished manuscript*. <http://www.briannosek.com/spf/>.
- Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, 56, 329–343.
- Blaison, C., Imhoff, R., Hühnel, I., Hess, U., & Banse, R. (2012). The affect misattribution procedure: Hot or not?. *Emotion*, 12(2), 403.
- 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.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of personality and social psychology*, 89(4), 469.
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological science*, 12(2), 163-170.
- De Houwer, J. (2006). What are implicit measures and why are we using them. *The handbook of implicit cognition and addiction*, 11-28.
- De Houwer, J., & Bruycker, E. (2007). The implicit association test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, 46(2), 401-421.
- Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of personality and social psychology*, 83(6), 1441.
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual review of psychology*, 54(1), 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 coefficient twenty is the same for two tests. *Psychometrika*, 34(3), 363–373.
- Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, 24(4), 218-225.
- Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects?. *Cognition and Emotion*, 24(6), 1008-1025.
- Gawronski, B., Gast, A., & De Houwer, J. (2015). Is evaluative conditioning really resistant to extinction? Evidence for changes in evaluative judgements without changes in evaluative representations. *Cognition and Emotion*, 29(5), 816-830.
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90(1), 1-20
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197-216.
- Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology & Marketing*, 27(10), 938-963.
- Jones, E. E., & Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychological Bulletin*, 76(5), 349.
- 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.
- Klauer, K. C., & Musch, J. (2003). Affective priming: Findings and theories. *The psychology of evaluation: Affective processes in cognition and emotion*, 7-49.

- Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2011). Reliability of implicit self-esteem measures revisited. *European Journal of Personality*, 25(3), 239-251.
- Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1), 34.
- McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.*, 27, 1-28.
- Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4), 630-633.
- Moran, T., Bar-Anan, Y., & Nosek, B. A. (in press). The effect of the validity of co-occurrence on automatic and deliberate evaluation. *European Journal of Social Psychology*.
- Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, 134, 565-584.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265-292). New York: Psychology Press.
- Nosek, B. A., & Greenwald, A.G. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin*, 135, 373-376.
- 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(2), 166–180.
- Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in cognitive sciences*, 15(4), 152-159.
- Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2012). Implicit social cognition. *Handbook of social cognition*, 31-53.
- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, 12(5), 413-417.

- Payne, B. K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, 81(2), 181.
- 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., & Arbuckle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition and Emotion*, 22, 238–271.
- Payne, B. K., & Lundberg, K. (2014). The Affect Misattribution Procedure: Ten Years of Evidence on Reliability, Validity, and Mechanisms. *Social and Personality Psychology Compass*, 8(12), 672-686.
- Peters, K. R., & Gawronski, B. (2011). Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations. *Personality and Social Psychology Bulletin*, 37(4), 557-569.
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, 19, 123-205
- Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, 19(3), 249-254.
- 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.
- Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology*, 46(5), 721-728.
- Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling Implicit Association Test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PloS one*, 10(6), e0129601.

- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment*, 24(4), 273-281.
- Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, 37(5), 867-878.
- Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the Source Persuasion of Implicit Evaluations Is Moderated by Source Credibility. *Personality and Social Psychology Bulletin*, 39(2), 193-205.
- 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(3), 383.
- Scherer, L. D., & Lambert, A. J. (2012). Implicit race bias revisited: On the utility of task context in assessing implicit attitude strength. *Journal of Experimental Social Psychology*, 48(1), 366-370.
- Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., & Eelen, P. (2007). On the predictive validity of indirect attitude measures: Prediction of consumer choice behavior on the basis of affective priming in the picture–picture naming task. *Journal of Experimental Social Psychology*, 43(4), 599-610.
- Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, 56, 283–294.
- Stahl, C., & Unkelbach, C. (2009). Evaluative learning with single versus multiple unconditioned stimuli: the role of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(2), 286.
- Stickgold, R., & Walker, M. P. (2005). Memory consolidation and reconsolidation: what is the role of sleep?. *Trends in neurosciences*, 28(8), 408-415.

- Stieger, S., Göritz, A.S., Hergovich, A., Voracek, M., 2011. Intentional faking of the Single Category Implicit Association Test and the Implicit Association Test. *Psychological Report*, 109, 219–230.
- Walther, E., & Langer, T. (2008). Attitude formation and change through association: An evaluative conditioning account. *Attitudes and attitude change*, 87-109.
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3), e0152719.
- Zlotnick, E., Dzik, A., Bar-Anan, Y. (2015). PIPE (Version 0.3) [Computer software].
- Zou, G.Y. (2007). Toward using confidence interval to compare correlations. *Psychological Methods*, 12, 399-413.

Supplement Tables and Figures

Table S1

Spearman correlations of the evaluation measures with evaluative information

	IAT	AMP	SPF	SPF_w	EPT	EPT_{BN}	EPT_G	Explicit
Experiment 1	.461 _b	.305 _c	.238 _{cd}	.257 _{cd}	.212 _d	.209 _d	.186 _d	.759 _a
Experiment 2	.519 _b	.286 _c						.763 _a
Experiment 3			.233 _{bc}	.283 _b	.203 _{bc}	.172 _c	.157 _c	.780 _a
Experiment 4	.538 _b	.281 _{de}	.334 _{cd}	.368 _c	.239 _e	.214 _{ef}	.161 _f	.794 _a

Notes. SPF_w is the score based on winsorized latencies. EPT_{BN} is the EPT score recommended by Bar-Anan and Nosek (2012); EPT_G is the score used by Gawronski et al. (2015). All the correlations are significantly different than zero; In each row, different subscript indicates a significant difference ($p < .05$, two-sided), computed following instructions from Zou (2007).

Table S2

Comparison of the measures using multiple regression analyses: Standardized estimated parameters for the interaction term (in parenthesis: p-value)

Experiment 1	IAT	AMP	SPF
AMP	.047 (.024)		
SPF	.103 (<.0001)	.055 (.012)	
EPT	.119 (<.0001)	.070 (.001)	.016 (.463)
Experiment 2	IAT	Experiment 3	SPF
AMP	.07 (.0004)	EPT	.023 (.304)
Experiment 4	IAT	AMP	SPF
AMP	.102 (<.0001)		
SPF	.089 (<.0001)	.012 (.471)	
EPT	.143 (<.0001)	.043 (.014)	.055 (.002)

Notes. The interaction terms were computed from multiple regressions that predicted the number of positive behaviors performed by Brian from a standardized preference score, a dummy variable indicating which indirect measure the participant completed, and the multiplication of the two predictors. Each multiple regression compared only two indirect measures by including data only from participants who completed one of the two measures. The interaction term reflected the difference between the two indirect measures in predicting the outcome; We used the best scores for SPF and EPT (i.e., the winsorized scores).

Table S3

Pearson correlations of the evaluation measures with evaluative information, after controlling for self-reported evaluation

	IAT	AMP	SPF_w	EPT
Experiment 1	.106 ^{**} _a	.024 _a	.084 _a	.033 _a
Experiment 2	.131 ^{**} _a	.024 _b		
Experiment 3			.096 [*] _a	.046 _a
Experiment 4	.154 ^{**} _a	-.008 _b	.062 _b	.058 _b

*Notes. SPF_w is the score based on winsorized latencies; * the correlation is significantly larger than 0 with $p < .05$; ** the correlation is significantly larger than 0 with $p < .01$; In each row, different subscript indicates a significant difference ($p < .05$, two-sided), computed following instructions from Zou (2007).*

Table S4

Behavior ratings for the behaviors used in the experiments

Behavior	Mean	SD	
Sheltered a family that had lost their home in a fire	8.60	0.85	9
Brought a blanket and a hot meal to a homeless person	8.43	0.91	9
Volunteers to visit terminally ill children in the hospital twice a month	8.27	0.97	8
Pays college tuition for his sisters children	8.22	1.11	9
Helped an elderly man who dropped some packages	8.17	0.96	9
Collected toys for underprivileged children	8.13	1.12	9
Offered his place in the line to an elderly person	8.10	1.05	9
Mailed back an expensive item that was delivered to him by mistake	7.99	1.25	9
Cooked a nice dinner for his friends	7.87	1.00	9
Always promptly returns what he's borrowed	7.86	1.24	9
Fixed the door knob for his neighbor	7.86	0.97	9
Always makes people laugh when he notices that they are sad	7.76	1.23	9
Rides his bicycle to work to help reduce pollution	7.71	1.20	9
Threw a surprise party for a friend	7.68	1.08	9
Picked up a friend from work late at night	7.61	1.27	9
Took his younger brother to a movie	7.55	1.14	9
Usually smiles at people he passes on the sidewalk	7.49	1.21	9
Brings fresh bagels to everyone at work every morning	7.47	1.33	9
Knew how to go about solving the problems his co-workers had trouble with	7.38	1.40	9
Risked his job by protesting against an unfair personnel practice in his company	7.34	1.60	9
Complimented a friend on his new clothes	7.22	1.13	9
Occasionally works overtime for no extra pay in order to do a good job	6.73	1.73	9
Skipped lunch to work on a company project	5.93	1.53	9
Did not offer his guest anything to drink	3.76	1.15	9
Embarrassed a friend by playing a prank on him	3.61	1.39	9
Whispered during a movie even though he knew it disturbed others	3.51	1.18	8
Called in sick for work when he was well	3.33	1.35	9
Almost crowded someone off the sidewalk in his hurry to cross the street	3.07	1.12	9
Turned in a report to his boss four days late	3.03	1.28	9
Cheated in a game of cards	3.01	1.26	9
Criticized his friend harshly	2.99	1.47	9
Disturbed his neighbors by playing loud music at night	2.80	1.26	9
Yelled at his parents over the phone	2.77	1.30	8
Cut in front of some people in line to buy tickets	2.56	1.15	1
Became upset and broke some dishes	2.52	1.21	9
Shoplifted an inexpensive item from a store	2.50	1.29	9
Criticized others because they weren't wearing fashionable clothes	2.41	1.20	9
Dented the fender of a parked car and didn't leave his name	2.33	1.21	9

Insulted his assistant at work	2.07	1.06	9
Was arrested for selling heroin to a teenager	2.05	1.95	9
Drove through a red light at a potentially dangerous intersection	2.04	1.18	8
Threw a rock at a dog that was barking	1.93	1.22	9
Smoked on a crowded bus	1.90	1.11	9
Made fun of a colleague who has a speech impediment	1.62	0.99	8
Stole money and jewelry from his relatives	1.54	1.00	9
Likes to pick fights with people that are physically weaker than him	1.48	0.97	9

Notes. These statistics were computed from the ratings in Experiment 1-3 by participants who rated the behaviors before the behaviors were attributed to the men. The experiments were identical at that stage (before seeing the men).

Figure S1

Figure S1. Experiments 1-4: Vertical probability density (violin) plots of the preference score of each measure as a function by the number of positive behaviours attributed to Brian (x-axis). The black circles are the mean preference, and the error-bars are 95% confidence intervals around each mean.