Assessing Test-Retest Reliability and Convergent Validity Among Four Implicit Association Measures

Ruo Ying Feng

Department of Psychology, McGill University

PSYC 380D: Honours Research Project and Seminar

Dr. Jordan Axt

Dr. Sarah Racine

April 22, 2021

Abstract

Over the last 25 years, the rise of research on implicit social cognition – attitudes, beliefs, or opinions that are comparatively automatic and resistant to conscious control – has been driven in part by the development of numerous methods to measure such processes. However, despite considerable use in the literature, relatively little is known about the comparative psychometric properties of different implicit measures across multiple attitudinal domains. This study uses a large online sample ($N > 5000$) to analyze the test-retest reliability and convergent validity of four implicit measures across ten topics. Correlational and meta-analytic results showed that the Single-Category Implicit Association Task (SC-IAT) had the best test-retest reliability, followed closely by the Implicit Association Task (IAT) and the Single Paired Features (SPF) task, whereas the Evaluative Priming Task (EPT) had significantly lower test-retest reliability. For convergent validity, the IAT was the measure most highly correlated with other implicit measures, while the EPT was the least highly correlated. When taking into account attitudinal domains, each measure showed substantial heterogeneity among true effects in meta-analytic estimates of test-retest reliability, except for the SC-IAT. Conversely, all measures retained assumptions of homogeneity in meta-analytic estimates of convergent validity, except for the EPT. Our results provide comparative knowledge about the psychometric strengths and weaknesses of various implicit measures, which can aid theoretical and practical advances in the study of implicit social cognition.

Keywords: Implicit social cognition, implicit measures, attitudes, stereotypes, test-retest reliability, convergent validity

Assessing Test-Retest Reliability and Convergent Validity Among Four Implicit Association Measures

Individuals hold beliefs, opinions, and attitudes about different topics. In psychological research, the convention has been to directly ask about them through self-report questionnaires, which tap into attitudes that are consciously endorsed and relatively more controlled (Buttrick et al., 2020; Cunningham et al., 2001). Although these explicit measures have been found to reliably predict associated behaviours (e.g., Bogart et al., 2004), many researchers in recent decades have been interested in studying implicit cognition, which focuses on attitudes or stereotypes that are more indirect, unconscious, and automatic (De Houwer et al., 2009). To assess these implicit attitudes, several measures such as the Evaluative Priming Task (EPT; Fazio et al., 1986) and the Implicit Association Task (IAT; Greenwald et al., 1998) have been developed, in which attitudes are inferred from comparisons of behavioural responses. As compared to explicit measures, implicit measures of attitudes or stereotypes have been found to be less sensitive to self-presentational biases (Olson et al., 2007) and less influenced by conscious goals (De Houwer et al., 2009). Moreover, implicit measures have been found to be a reliable predictor of relevant behaviour, independent of explicit attitudes (Buttrick et al., 2020; Kurdi et al., 2018).

Over the last two decades, the use of these indirect or implicit measures of attitudes, beliefs, and stereotypes has grown in both breadth and depth. Indeed, the study of implicit social cognition has not only increased in the number of published or cited articles (Forscher et al., 2019; Greenwald & Lai, 2020; Nosek et al., 2011), but also in its prevalence across multiple subfields of psychology (De Houwer et al., 2009). For example, although implicit measures originated within social psychology, they have now spread to subfields ranging from clinical (e.g., Nock et al., 2010) to consumer psychology (e.g., Maison et al., 2004). These measures of implicit cognition have also played a crucial role as a point of entry for larger discussions about intergroup disparities and biases. For instance, between 2002 and 2020, more than ten million sessions of the Race IAT (measuring implicit attitudes towards Black vs. White people) have been completed on Project Implicit's website (https://implicit.harvard.edu; Xu et al., 2014). Moreover, these data have been used to investigate important societal questions such as the inter- and intraregional variability of racial attitudes (Rosenbusch et al., 2020) or the associations between income inequality and racial bias (Connor et al., 2019).

The growing prevalence of these indirect measures within the literature has been highlighted by several reviews and meta-analyses. In particular, Nosek and colleagues (2011) reviewed current citation patterns of implicit measures in social cognition research and found 6282 total citations, in which 750 (~12%) were accounted by the year 2010 alone. Forscher and colleagues' (2019) meta-analysis examining a narrower question concerning the effectiveness of interventions to change implicit measures retrieved 4908 total relevant articles published between 1995 and 2015. Most recently, Greenwald and Lai's (2020) review of the implicit social cognition literature found further evidence for the increased use of implicit measures, which collectively have been cited 4580 times and used in 1423 studies between 2014 and 2018. Implicit measures seem to be growing in popularity and influence, especially in psychological research.

Across the psychological literature, the most common implicit measures have been the Implicit Association Task (IAT; Greenwald et al., 1998) and the Evaluative Priming Task (EPT; Fazio et al., 1986). For instance, in the most recent review by Greenwald and Lai (2020), the IAT was cited 2116 times (46.20% of all implicit measures) and used in 767 studies (53.90%) from

2014 to 2018, while the EPT was cited 359 times (7.84%) and used in 103 studies (7.24%). Though both measures seek to assess implicit associations, they adopt different approaches for doing so. For instance, the IAT involves showing four categories on the screen, where one attribute label (e.g., good vs. bad) and one attitude category label (e.g., Black vs. White people) are presented together on each side of the screen. Participants are then told to quickly and accurately categorize words and images to one of the two combinations of categories. The underlying principle of the IAT is that categorization speed and performance should indicate the extent to which two concepts are more closely associated implicitly. The EPT also involves categorizing target words into two labels (e.g., good vs. bad), but one important difference is that primes (e.g., images of Black vs. White faces) immediately precede target words, which might facilitate the identification of whether the target word is positive or negative. As a result, while participants are instructed to attend to all four categories and attributes in the IAT, instructions for the EPT tell participants to only focus on categorizing the two classes of stimuli that follow the to-be-ignored primes.

Although the IAT and EPT represent the two most frequently used implicit measures, review articles also demonstrate an increasing diversity in the measurement of implicit attitudes via variants of the IAT (e.g., the Single-Category Implicit Association Test or SC-IAT; Karpinski & Steinman, 2006) or more behavioural methods (e.g., mouse-tracking; Freeman & Ambady, 2010). Yet, despite the growing prevalence and diversity of implicit measures in different subfields of psychology, the psychometric properties of measures other than the IAT or the EPT, as well as comparisons of these measures across topics, are not well understood. Specifically, little is known about the associations between two administrations of the same measure (i.e., test-retest reliability) and the relationship between two measures seeking to assess the same implicit construct (i.e., convergent or construct validity). In the next section, we explore the importance of test-retest reliability and convergent validity when considering the broader use and validity of such measures in research on implicit cognition.

The Importance of Test-Retest Reliability and Convergent Validity

Test-retest reliability refers to consistency or stability of results across multiple administrations of the same test over time. A measure with strong test-retest reliability indicates that it is subject to less random error and thus more internally valid. In the study of implicit attitudes and implicit self-esteem, this psychometric quality is especially important because it allows researchers to meaningfully assess individual differences and to predict conceptually related outcomes (Bosson et al., 2000; Rae & Olson, 2017). Simultaneously, weaker test-retest reliability might also convey significant properties about relevant constructs. For instance, Gawronski and colleagues (2017) found that, contrary to common beliefs, explicit measures of self-concept, racial attitudes and political attitudes were more resistant to change over time (i.e., showed greater test-retest reliability) than implicit measures of the same domains. As such, weaker test-retest reliability might not be a threat to construct validity (Cunningham et al., 2001) but might indicate differences in temporal stability or other properties of these constructs.

Regardless, as test-retest reliability can indicate the precision of a measure (Greenwald & Lai, 2020), more investigation into this psychometric quality is needed to reduce potential measurement error, which remains an important issue in the field of implicit cognition. Notably, Connor and Evers (2020) highlighted how the issue of measurement error may have led to theoretical confusion in Payne and colleagues' (2017) bias-of-crowds model, which reconceptualizes implicit bias as being an aspect of social situations rather than of the individual. Additionally, while implicit and explicit measures may vary in test-retest reliability due to inherent construct-related differences, better measures of such constructs will have better test-retest reliability because they will be less impacted by measurement error (Axt, 2018). In other words, it is imperative to determine which implicit measures possess the best test-retest reliability to minimize measurement error and potential theoretical misunderstandings.

Similarly, construct validity, or the extent to which a test measures what it claims to measure, is a psychometric property essential to the overall validity of a test, as it contributes to the extent to which a test's inferences are appropriate, meaningful, and useful (Coulacoglou & Saklofske, 2017). Construct validity is composed of two important aspects – convergent and discriminant validity. A test that has high convergent validity should correlate well with tests that claim to measure the same construct, while poorly correlating with tests that do not measure the same construct (i.e., discriminant validity; Krabbe, 2017). Literature on implicit cognition thus far has largely supported a dual-attitude perspective, where implicit measures of the same attitude or stereotype should be more strongly correlated with one another than with parallel explicit measures, while explicit measures should be more strongly correlated with one another than with implicit measures (Bar-Anan & Vianello, 2018).

Nonetheless, as mentioned previously, the current literature examining test-retest reliability and convergent validity across different topics and types of implicit tests is still limited. In particular, only one study (Bar-Anan & Nosek, 2014) has comparatively investigated these psychometric properties across multiple attitudinal domains. Furthermore, to our knowledge, only three studies have examined the test-retest reliability of the SC-IAT (Chevance et al., 2017; Galdi et al., 2012; Stieger et al., 2010), and only two have examined the functionally similar Single-Target Implicit Association Test (ST-IAT; Bar-Anan & Nosek, 2014; Bluemke & Friese, 2008). There have also only been two studies looking at test-retest reliability of the

Sorting Paired Features (SPF; Bar-Anan et al., 2009; Bar-Anan & Nosek, 2014). In the next section, we review current literature investigating psychometric properties of implicit measures.

Psychometric Properties of Implicit Measures

Thus far, there has been substantial research demonstrating the psychometric properties (e.g., internal consistency, test-retest reliability, construct validity) of some implicit measures, especially the IAT and the EPT. However, most of these investigations have been limited to a single topic or domain, such as race (e.g., Cunningham et al., 2001), self-esteem (e.g., Krause et al., 2011), or smoking (e.g., Spruyt et al., 2015). Although the psychometric properties of less-commonly used implicit measures have yet to be explored to the same extent, there has been some progress in the last decade. For example, Chevance and colleagues (2017) measured the test-retest reliability of the IAT and the SC-IAT for physical activity and sedentary behaviour and found that the IAT showed better internal consistency and test-retest reliability ($\alpha = .93$, $r = .75$) than both the physical activity SC-IAT ($\alpha = .71$, $r = .33$) and the sedentary behaviour SC-IAT ($\alpha = .76$, $r = .19$).

Despite better understanding of the measurement properties of diverse implicit measures, Greenwald and Lai's (2020) meta-analysis highlights that only one study has simultaneously investigated the comparative psychometric qualities of several common measures across multiple domains (Bar-Anan & Nosek, 2014). In this work, the researchers compared seven implicit measures across three domains (i.e., race, politics, self-esteem). Results showed that the IAT and the Brief IAT (BIAT; Sriram & Greenwald, 2009) had the best psychometric qualities overall, especially when looking at internal consistency ($\alpha_{\text{IAT}} = .88$, $\alpha_{\text{BIAT}} = .83$) and test-retest reliability ($r_{\text{IAT}} = .45$, $r_{\text{BIAT}} = .63$). The BIAT was on average the most highly correlated with other implicit measures ($r = .41$), with the IAT having similar convergent validity ($r = .39$). The measures with the worst overall psychometric qualities were the EPT (α = .57, *r* = .33), the SPF (α = .53, *r* = .46), and the Affective Misattribution Procedure (AMP; Payne et al., 2005; α = .69, *r* = .50). Likewise, the EPT (*r* = .25) and the AMP (*r* = .26) also had the worst convergent validity when averaged across topics. Overall, average internal consistencies were above α = .70 for four out of the seven measures while average test-retest reliabilities were over *r* = .40 for all measures except the EPT (*r* = .33). For convergent validity, average correlations with other implicit measures were over *r* = .30 for all except the EPT and the AMP.

More recently, Greenwald and Lai's (2020) meta-analysis similarly revealed that despite substantial variation among implicit measures, internal consistencies across topics were acceptable (especially for the IAT) while test-retest reliabilities were modest (*r* < .30) for several measures (e.g., EPT, SC-IAT). Lastly, Greenwald and Lai (2020) also highlighted the need for additional insight into interrelations among implicit measures, which is needed to advance knowledge about convergent validity.

Although there is some existing literature on the convergent validity of implicit measures, it is severely limited by the number of attitudinal domains investigated. Considering the breadth of domains to which these implicit measures have been applied to (Greenwald & Lai, 2020), the number of studies that have accounted for multiple domains while assessing interrelations among these measures is relatively low. For instance, studies looking the convergent validity of implicit measures have only assessed race (e.g., Cunningham et al., 2001), self-esteem (e.g., Bosson et al., 2000; Rudolph et al., 2008), or feelings of threat (e.g., Reinecke et al., 2010). In these studies, the average convergent validity ranged from weak and non-significant for self-esteem (mean *r* = -.09 for the IAT; Bosson et al., 2000) to moderately strong for race (mean *r* = .54 for the IAT; Cunningham et al., 2001). Furthermore, in Bar-Anan and Nosek's (2014) previously mentioned comparative analysis, the authors found significant variability across topics. Politics (mean $r = .53$) had by far the strongest convergent validity among all seven measures, followed by race (mean $r = .33$) and self-esteem (mean $r = .21$).

This variability observed across attitudinal domains in individual studies as well as in Bar-Anan and Nosek's (2014) comparative analysis raises major concerns for the field of implicit social cognition. Indeed, if implicit measures are expected to vary substantially in convergent validity depending on the domain, researchers may be faced with significant uncertainty when interpreting the results of their own studies. That is, without greater knowledge of how convergent validity may vary across topics when using implicit measures, researchers may have difficulty discerning whether they are assessing a domain where implicit measures have high or low convergent validity, information that is crucial for determining whether such measures have accurately assessed the desired construct. This potential uncertainty with regard to the interpretation of findings may in turn have important implications for the wider use of implicit measures in both theoretical and applied settings. As such, to better understand whether this variability in convergent validity is restricted to the specific domains chosen in previous studies (i.e., Bar-Anan & Nosek, 2014) or represents a broader issue in the study of implicit social cognition, it is crucial to further investigate the interrelations among implicit measures across a wider range of domains.

The Present Study

Using a large online sample, we investigate the a) test-retest reliability and b) convergent validity of four implicit measures across ten topics. Based on prior meta-analyses and comparative studies (Bar-Anan & Nosek, 2014; Greenwald & Lai, 2020), we hypothesize that a) across topics, although all implicit measures will have moderate test-retest reliability, the EPT will have the lowest level of test-retest reliability and the IAT will have the highest level. Relatedly, we hypothesize that b) the EPT will have the worst convergent validity across topics (i.e., be the least correlated with other implicit measures), because it procedurally distinguishes itself from other implicit measures by not requiring participants to categorize primes (Bar-Anan & Nosek, 2014). However, as we have noted previously, the current literature comparing the psychometric properties of these implicit measures, especially across multiple topics, is limited. Therefore, it is possible that our analysis diverges from previous comparative studies. Moreover, because there is little systematic knowledge about the test-retest reliability and convergent validity of the SC-IAT and SPF, it is difficult to hypothesize whether they will perform better or worse than other, more common implicit measures (i.e., EPT, IAT).

To our knowledge, the present study is the most comprehensive investigation of test-retest reliability and convergent validity of implicit measures across multiple attitudinal domains. Additionally, through our large sample sizes and breadth of topics, we contribute significantly to the literature on comparisons between these specific measures and on the psychometric properties of implicit measures as a whole.

Contribution to Literature on Test-Retest Reliability

The large sample presented here greatly increases the amount of data available on the question of test-retest reliability. Our sample size would represent 43.98% (*n* = 683) of the total literature looking at test-retest reliability for the EPT, 11.36% (*n* = 704) of the total literature for the IAT, 120.42% (*n* = 808) of the total literature for the SC-IAT, and 138.89% (*n* = 675) of the total literature for the SPF. In addition, aside from expanding the available data concerning test-retest reliability of these implicit measures, this analysis also expands the question of test-retest reliability into several new domains for all four implicit measures (i.e., Age, Arab-Muslim, Gender-Career, Gender-Science, Religion, Sexuality, Weapons, and Weight), as well the Disability domain for the EPT, SC-IAT, and SPF.

Contribution to Literature on Convergent Validity

A prior review of the literature on implicit social cognition found that 46 studies included more than one implicit measure (Greenwald & Lai, 2020). Although most of these studies only considered two implicit measures, Zenko and Ekkekakis (2019) included nine measures of automatic exercise associations, Bar-Anan and Nosek (2014) and Bosson and colleagues (2000) conducted comprehensive analyses using seven implicit attitude measures each, and Krause and colleagues (2011) investigated five implicit measures related to self-esteem. As such, our study would be the fifth-largest study in terms of the number of implicit measures assessed simultaneously. However, our study would be the largest in terms of sample size that includes multiple measures of implicit attitudes or stereotypes. In addition, our study would be the largest comparative analysis of psychometric qualities in terms of attitudinal domains (*n* = 10), which would significantly expand the number of areas to which convergent or construct validity has been investigated.

This analysis would then greatly contribute to the existing literature on convergent validity. Indeed, our sample size (collapsing across domain) represents about 105.70% (*n* = 1112) of the sample size of published studies examining the association between the EPT and IAT, and also represents 311.39% (*n* = 1121) and 436.33% (*n* = 1309) of the existing literature on the correlations between the IAT and SPF as well as the EPT and SPF, respectively. Moreover, when taking into account the ST-IAT, which is functionally similar to the SC-IAT, our study would represent approximately 341.67% (*n* = 1025) of the sample size on the correlations between the EPT and SC/ST-IAT, 333.67% (*n* = 1001) of the correlations between the IAT and SC/ST-IAT, and 311.33% (*n* = 934) of the correlations between the SC/ST-IAT and SPF.

Methods

Participants

Participants voluntarily completed the study on Project Implicit (http://implicit.harvard.edu) between September 27, 2019 and December 4, 2020. The study served as a "background study" that was assigned to participants only after they had completed all of the other studies in the research pool for which they were eligible. Participants were randomly assigned to one of ten topics (i.e., Age, Arab-Muslim, Disability, Gender-Career, Gender-Science, Race, Religion, Sexuality, Weapons, Weight), which each consisted of one of four implicit measures (i.e., IAT, EPT, SC-IAT, SPF). In addition, each study session included five self-report measures of explicit attitudes or stereotypes, 25 self-report outcome measures, and a demographics questionnaire. Participants were able to complete multiple sessions, which allowed us to assess test-retest reliability if they were randomly assigned the same measure and topic across multiple study sessions, and to assess convergent validity if they were assigned to the same topic but different measures across multiple study sessions. All materials and procedures were approved by the University of Virginia's Institutional Review Board.

The data for this analysis came from 206,290 study sessions that completed at least one measure of implicit associations, which represents 120,882 participants (64.8% female, 70.6% White, 74.8% US citizens, M_{Age} = 36.58, *SD* = 15.26). Among these participants, 36,072 (29.84%) completed at least two implicit measures.

Stimuli

For all four implicit measures (i.e., IAT, EPT, SC-IAT, SPF), the attribute category labels consisted of *Good* (items: *Friend, Smiling, Adore, Joyful, Pleasure, Friendship, Happy, Attractiv*e) and *Bad* (items: *Bothersome, Poison, Pain, Nasty, Dirty, Hatred, Rotten, Horrific*), with the exception of the Gender-Career, Gender-Science, and Weapons topic conditions. In these topics, the target labels were, respectively, *Career* (items: *Career, Corporation, Salary, Office, Professional, Management, Business*) and *Family* (items: *Wedding, Marriage, Parents, Relatives, Family, Home, Children*), *Science* (items: *Astronomy, Math, Chemistry, Physics, Biology, Geology, Engineering*) and *Liberal Arts* (items: *History, Arts, Humanities, English, Philosophy, Music, Literature*), and *Weapons* (items: images of grenade, axe, cannon, mace, revolver, rifle, sword) and *Harmless Objects* (items: images of bottle, camera, coke, ice cream, phone, Walkman, wallet). See Appendix A for more details about attitude category labels and specific stimuli by topic.

Measures

In each topic condition, participants were randomly assigned one of four implicit measures of attitudes (e.g., “Race” measures implicit evaluations of Black versus White people) or stereotypes (e.g., “Weapons” measures the strength of implicit associations between Black and White people with guns versus objects). All implicit measures were selected on the basis of prior research and prominence in existing literature on implicit cognition (Bar-Anan & Nosek, 2014; Greenwald & Lai, 2020).

Implicit Association Task (IAT)

We followed the IAT procedure described by Nosek and colleagues (2007). One at a time, words and/or images appeared at the center of the screen. Participants were instructed to categorize attitude items into category labels on the top-right and top-left of the screen, while being as quick and as accurate as possible. A total of 120 critical trials were administered within seven blocks. In Block 1, which also served as a practice block, participants categorized items corresponding to the two attitude items (e.g., Black vs. White faces). In Block 2, participants did the same, but with good and bad words (e.g., “friendship” vs. “hate”). Blocks 3 and 4 combined the first two blocks by grouping, for example, Black faces and good words on one key and White faces and bad words on the other key. Blocks 5, 6 and 7 were the same as Blocks 1, 3 and 4, but the attitude objects switched side – for instance, Black faces and bad words were now grouped on one key while White faces and good words were grouped on the other.

Evaluative Priming Task (EPT)

The EPT procedure followed the one described by Fazio and colleagues (1995). A total of 180 critical trials were administered within three blocks. An initial block instructed participants to categorize words as good or bad. The following three (critical) blocks also consisted of categorizing words into these two labels, but a prime item (i.e., attitude item) appeared before each word. For example, in the Age version of the EPT, critical trials consisted of either an old or a young face immediately preceding the good or bad words.

Single-Category Implicit Association Task (SC-IAT)

The SC-IAT is a modification of the IAT (Karpinski & Steinman, 2006). This measure consists of one practice block and four test blocks (192 critical trials total). In each test block (Blocks 2-5), participants were instructed to categorize randomly ordered attitude items (e.g., old faces) and attribute labels (i.e., good vs. bad words) into two categories. In Blocks 2 and 3, the two categories were, for example, old faces + good words and bad words alone. The two remaining blocks switched these labels, such that one key corresponded to old faces + bad words and the other key corresponded to good words alone. The following attitude items were selected for the SC-IAT: *Old People* for the Age condition, *Arab Muslims* for the Arab condition, *Disabled Persons* for the Disability condition, *Female* for the Gender-Science and Gender-Career conditions, *Black People* for the Race and Weapons conditions, *Judaism* for the Religion condition, *Gay People* for the Sexuality condition, and *Fat People* for the Weight condition (see Appendix A).

Sorting Paired Features (SPF)

As described by Bar-Anan and colleagues (2009), the SPF consists of sorting item pairs into category pairs that appear in each of the four screen corners. These category pairs include all four possible combinations of attitude items and attribute types (i.e., good or bad words). For instance, in the Race SPF, the four category pairs are Black faces + good words, White faces + good words, Black faces + bad words, and White faces + bad words. A total of 120 critical trials across three blocks were administered.

Analysis Strategy

Data Processing

Implicit measures were processed following Bar-Anan and Nosek's (2014) recommendations, who chose scoring algorithms that adhered to the current literature’s standards or that produced the best psychometric qualities. All measures were scored using a variation of the *D* algorithm, in which a participant’s average latency difference score was divided by the standard deviation of their response latencies across both critical conditions.

IAT. For all topics, the IAT was scored based on Greenwald and colleagues' (2003) method. We removed trials slower than 10000ms and faster than 400ms and excluded participants with more than 10% of trials faster than 300ms. Each participant’s IAT *D* score was the average between Blocks 3 and 6’s and Blocks 4 and 7’s *D* scores. Ranging between -2 and 2, more positive *D* scores meant stronger implicit associations towards the categories listed in Label 2 versus Label 1 in Appendix A.

EPT. The EPT scoring also algorithm followed Bar-Anan and Nosek's (2014). EPT sessions with more than 40% incorrect responses were excluded as well as trials two standard deviations away from average response latency in the trial's condition (e.g., *Fat-Bad*). For each block, we computed a single-category *D* score as the difference between each trial condition's average log transformed response latencies (e.g., *Fat-Bad* minus *Fat-Good*) divided by overall standard deviation. EPT preference scores were calculated using the difference between two single-category scores, averaged across three blocks. More positive *D* scores referred to stronger implicit associations towards the items listed in Stimuli 2 versus Stimuli 1 in Appendix A.

SC-IAT. Again, following Bar-Anan and Nosek (2014), data processing for the SC-IAT was similar to the IAT. For each attitude item (e.g., old faces + good words, old faces + bad words), we calculated the SC-IAT *D* scores, averaging across respective blocks. Then, we computed the preference score by taking the difference between the two single-category SC-IAT *D* scores, such that more positive *D* scores meant more positive implicit evaluations towards the categories listed in Label 1 (see Appendix A).

SPF. Exclusion criteria for single trials and participants were identical to those for the IAT. Following Bar-Anan and colleagues (2009), within each block, we computed the *D* score for each of the four trial types (e.g., *Fat-Good, Fat-Bad, Thin-Good, Thin-Bad*). Then, we calculated a preference score for each block using the difference between single-category *D* scores. Finally, a participant's SPF preference score corresponded to the average of scores across all three blocks. Similar to the IAT, more positive *D* scores referred to stronger implicit associations towards the categories listed in Label 2 versus Label 1 in Appendix A.

Data Analysis

All data analysis and visualization were completed using R version 3.6.0 (R Core Team, 2019) with the *tidyverse* (Wickham, 2019), *psych* (Revelle, 2019), and *metafor* (Viechtbauer, 2010) packages.

Test-Retest Reliability. To compute the test-retest reliability of an implicit measure, we first calculated the Pearson's correlation coefficient (r) for scores at Time 1 and Time 2. We repeated this procedure for each topic condition and across all topics for each implicit measure. Next, to better assess the overall reliability of an indirect test, we conducted a meta-analysis by fitting a random-effects model on all topic conditions per measure. Lastly, we computed Wald-type tests for each pair of implicit measures to test whether meta-analytic test-retest estimates obtained from the meta-analyses reliably differed from one another.

Convergent Validity. Convergent validity refers to correlations with other implicit measures. Similar to our procedure for test-retest reliability, we first calculated the Pearson's correlation coefficient (r) for each pair of implicit measure for each topic condition, as well as across all topics. Then, we again conducted a meta-analysis by fitting a random-effects model on all topic conditions per each pair of measures.

Results

Test-Retest Reliability

Table 2 shows test-retest correlations for each implicit measure by topic condition and averaged across topics. Collapsing across topics, the SC-IAT showed the strongest test-retest reliability ($r = .47$, 95% CI [.42, .53], $p < .001$), followed closely by the IAT ($r = .43$, 95% CI [.37, .49], $p < .001$) and the SPF ($r = .41$, 95% CI [.35, .47], $p < .001$). While all measures were statistically significant when averaged across topics, the EPT showed the weakest test-retest reliability (r = .19, 95% CI [.12, .26], p < .001). For the most part, this ranking held true when looking at individual topic conditions. In six out of ten topics (i.e., Arab, Disability, Gender-Career, Gender-Science, Sexuality, Weapons), the EPT showed the weakest test-retest reliability. Indeed, Time 1 and Time 2 EPT scores were even negatively correlated in the Disability and Sexuality conditions, although these negative correlations were not statistically significant. The SC-IAT showed the strongest test-retest reliability in five out of ten topics (i.e., Arab, Gender-Career, Gender-Science, Religion, Weight). Across all topics, the Disability-SPF showed the strongest test-retest correlations (r = .72, 95% CI [.58, .81], p < .001), while the Sexuality-EPT showed the weakest test-retest correlations (r = -.11, 95% CI [-.32, .11], p = .327).

The results of our meta-analysis of overall test-retest reliability for each implicit measure are presented in Table 3 and Figure 1. Similar to the results mentioned above, the SC-IAT showed the greatest test-retest reliability (r =.39, 95% CI [.34, .45], p < .001). The I^2 statistic (I^2 = 0.00%) and the non-significant Cochran's test for heterogeneity (Q = 4.09, df = 9, p = .905) suggest that there was low heterogeneity among true effects. The IAT showed similar test-retest reliability, followed closely by SPF. However, both measures showed considerable heterogeneity among true effects, as demonstrated by their I^2 statistic and significant Cochran's test. As for the EPT, results of the meta-analysis reflected the lower test-retest reliability mentioned above (r = .13, 95% CI [.02, .24], p = .017). Here, the I^2 statistic (I^2 = 53.29%) was moderate-high and Cochran's test for heterogeneity was significant (Q = 19.27, df = 9, p = .023), indicating heterogeneity among true effects.

Lastly, the Wald-type tests for each pair of implicit tests are shown in Table 4. Across topics, we found that test-retest reliability for the EPT was significantly lower than all other measures. However, the other measures (i.e., IAT, SC-IAT, SPF) did not differ significantly from each other in test-retest reliability.

Table 2

Test-Retest Correlations for Implicit Measures (Overall and by Topic Condition)

	n	*r*	95% CI	*p*
Overall				
EPT	683	*.19*	[.12, .26]	<.001
IAT	704	.43	[.37, .49]	<.001
SC-IAT	808	**.47**	[.42, .53]	<.001
SPF	675	.41	[.35, .47]	<.001
Age				
EPT	26	.23	[-.17, .57]	.256
IAT	68	*.06*	[-.18, .29]	.631
SC-IAT	84	.38	[.18, .55]	<.001
SPF	36	**.39**	[.07, .64]	.020
Arab				
EPT	81	*.24*	[.02, .44]	.030
IAT	68	.48	[.27, .64]	<.001
SC-IAT	78	**.49**	[.30, .64]	<.001
SPF	73	.37	[.15, .55]	.001
Disability				
EPT	78	*-.09*	[-.31, .13]	.416
IAT	58	.53	[.32, .70]	<.001
SC-IAT	68	.34	[.11, .53]	.005
SPF	74	**.71**	[.58, .81]	<.001
Gender-Career				
EPT	75	*.06*	[-.17, .28]	.606
IAT	71	.22	[-.01, .43]	.063
SC-IAT	83	**.36**	[.16, .54]	<.001
SPF	83	.19	[-.02, .39]	.082
Gender-Science				
EPT	48	*.04*	[-.25, .32]	.785
IAT	70	.37	[.14, .55]	.002
SC-IAT	70	**.42**	[.21, .60]	<.001
SPF	61	.12	[-.14, .36]	.366
Race				
EPT	79	.35	[.14, .53]	.002

IAT	76	**.43**	[.23, .60]	<.001
SC-IAT	86	*.33*	[.12, .50]	.002
SPF	61	.37	[.13, .57]	.003
Religion				
EPT	70	.15	[-.09, .37]	.224
IAT	80	.14	[-.09, .35]	.230
SC-IAT	88	**.35**	[.15, .52]	<.001
SPF	65	*.02*	[-.22, .27]	.857
Sexuality				
EPT	81	*-.11*	[-.32, .11]	.327
IAT	70	**.56**	[.37, .70]	<.001
SC-IAT	95	.36	[.17, .53]	<.001
SPF	65	.44	[.22, .62]	<.001
Weapons				
EPT	68	*.10*	[-.14, .33]	.424
IAT	63	.24	[-.01, .46]	.059
SC-IAT	78	.36	[.15, .54]	.001
SPF	79	**.47**	[.28, .63]	<.001
Weight				
EPT	79	.32	[.11, .51]	.004
IAT	82	.36	[.16, .54]	<.001
SC-IAT	80	**.50**	[.31, .65]	<.001
SPF	80	*.18*	[-.05, .38]	.120

Note. **Bold** font = best test-retest reliability in topic; *underlined italic* font = worst test-reliability in topic.

Table 3

Overall Effect Sizes for Meta-Analyses of Test-Retest Correlations Among Implicit Measures

	Overall Effect Size Estimate			Heterogeneity		
	r	95% CI	*p*	*Q*	*p*	I^2*(%)*
EPT	.13	[.02, .24]	.017	19.27	.023	53.29
IAT	.35	[.25, .45]	<.001	23.70	.005	62.47
SC-IAT	.39	[.34, .45]	<.001	4.09	.905	0.00
SPF	.34	[.21, .47]	<.001	51.98	<.001	77.59

Figure 1

Forest Plots of Meta-Analyses for Each Implicit Measure

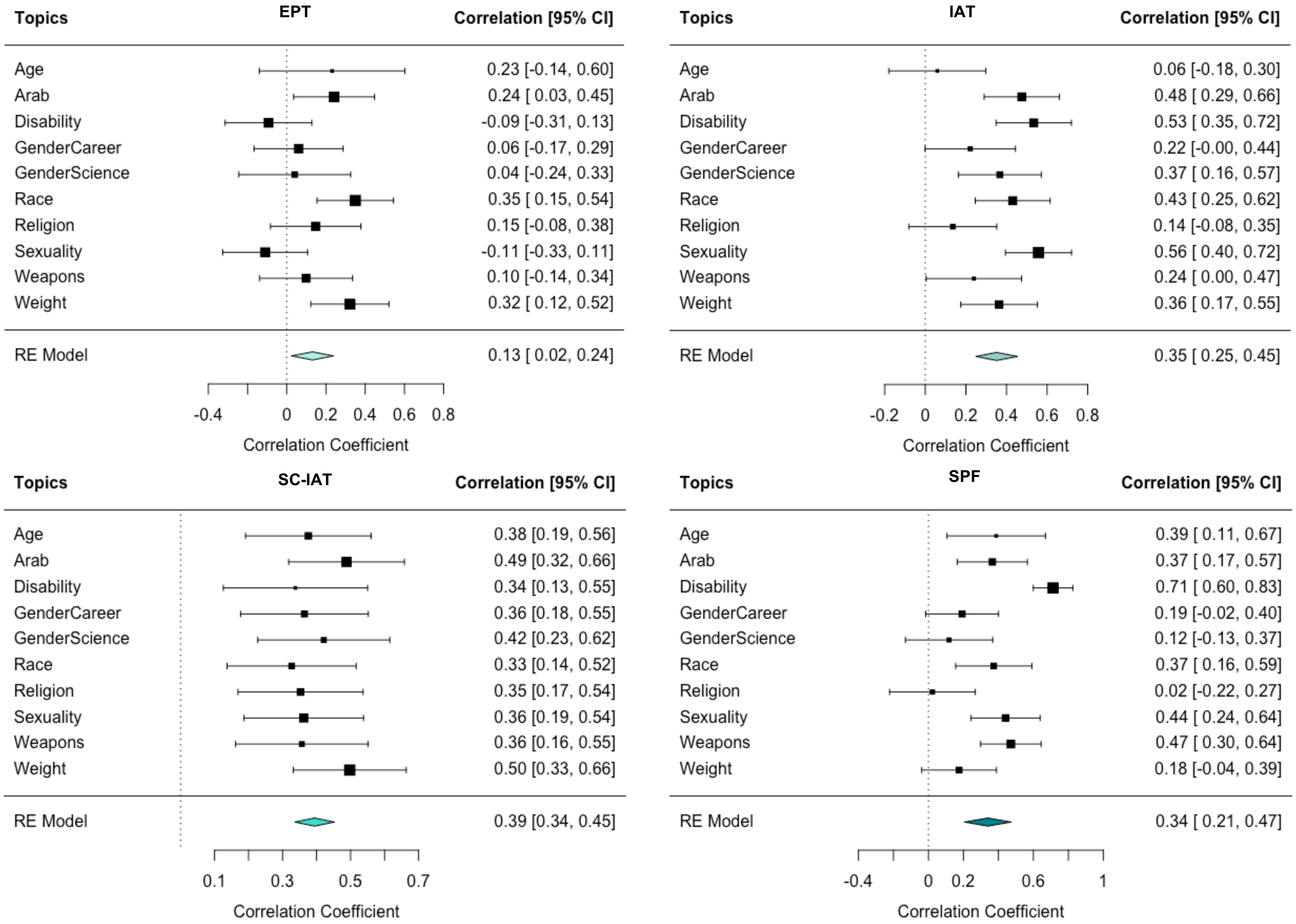

Table 4

Wald-Type Tests Among Meta-Analyses of Implicit Association Measures

	IAT				SC-IAT				SPF			
	b_1	*SE*	*z*	*p*	*b_1*	*SE*	*z*	*p*	*b_1*	*SE*	*z*	*p*
EPT	0.22	0.08	2.91	.004	0.26	0.06	4.23	<.001	0.21	0.09	2.39	.017
IAT					0.04	0.06	0.70	.482	-0.01	0.09	-0.14	.887
SC-IAT									-0.06	0.07	-0.74	.460

Note. Difference between two estimates is calculated as the measure listed in the column versus the measure listed in the row.

Convergent Validity

Table 5 presents overall correlations among all implicit measures and correlations broken down by topic condition. Collapsing across topics, the IAT was the most highly correlated with other implicit measures, especially with the SPF (*r* = .33, 95% CI [.28, .38], *p* < .001). On the other hand, the EPT was the least strongly correlated with other implicit measures, with correlations between scores on the EPT and SC-IAT even being weakly negative (*r* = -.03, 95% CI [-.08, .03], *p* = .323). This pattern held across topic conditions – within all ten topics, the strongest correlations between pairs of implicit measures were either IAT with SC-IAT (i.e., Gender-Career, Gender-Science, Religion, Weapons, Weight) or IAT with SPF (i.e., Age, Arab, Disability, Race, Sexuality), while the weakest correlations were either EPT with IAT (i.e., Gender-Career), EPT with SC-IAT (i.e., Age, Arab, Disability, Race, Sexuality, Weight), or EPT with SPF (i.e., Gender-Science, Religion, Weapons). Across all topics, the pair of implicit measures with the strongest correlation was the Disability-IAT and Disability-SPF (*r* =.45, 95% CI [.28, .60], *p* < .001), while the pair with the weakest correlation was the Religion-EPT and Religion-SPF (*r* = -.12, 95% CI [-.28, .05], *p* = .153).

Similar to test-retest reliability, we conducted a meta-analysis to better assess the overall associations among pairs of implicit measures (see Figure 2 and Table 6). We found that, as with results described above, the IAT and SPF were most strongly correlated to one another (*r* = .31, 95% CI [.24, .37], *p* < .001), with the I^2 statistic and Cochran's test showing low heterogeneity among true effects (I^2 = 35.44%, *Q* = 13.91, *df* = 9, *p* = .126). All other pairs of implicit measures were moderately to highly significantly positively correlated with each other, with the exception of the EPT and SC-IAT (*r* = .00, 95% CI [-.06, .05], *p* = .913). Assumptions of homogeneity for all pairs of implicit tests were held, except for the EPT and SPF (see Table 6).

Table 5

Correlations Among Implicit Measures (Overall and by Topic Condition)

	IAT				SC-IAT				SPF			
	n	*r*	95% CI	*p*	*n*	*r*	95% CI	*p*	*n*	*r*	95% CI	*p*
Overall												
EPT	1112	.16	[.10, .22]	<.001	1202	*-.03*	[-.08, .03]	.323	1309	.15	[.09, .20]	<.001
IAT					1001	.18	[.12, .24]	<.001	1121	**.33**	[.28, .38]	<.001
SC-IAT									1078	.11	[.05, .16]	.005
Age												
EPT	97	.18	[-.02, .37]	.071	112	*-.04*	[-.22, .15]	.699	66	.09	[-.16, .32]	.473
IAT					115	.15	[-.04, .32]	.114	121	**.35**	[.18, .50]	<.001
SC-IAT									102	.17	[-.02, .35]	.085
Arab												
EPT	128	.22	[.05, .38]	.014	108	*-.05*	[-.24, .14]	.599	141	.19	[.03, .35]	.021
IAT					90	.09	[-.12, .29]	.412	110	**.37**	[.20, .52]	<.001
SC-IAT									91	.02	[-.19, .22]	.864
Disability												
EPT	100	.21	[.02, .39]	.033	133	*-.11*	[-.28, .06]	.200	132	.13	[-.04, .29]	.144
IAT					99	.25	[.05, .42]	.014	98	**.45**	[.28, .60]	<.001
SC-IAT									111	.19	[.00, .36]	.047
Gender-Career												
EPT	93	*-.07*	[-.27, .13]	.498	123	*-.07*	[-.25, .11]	.425	131	.15	[-.02, .32]	.083
IAT					116	**.26**	[.09, .43]	.004	110	.24	[.05, .40]	.013
SC-IAT									124	.15	[-.03, .31]	.105
Gender-Science												
EPT	95	.16	[-.05, .35]	.128	108	.15	[-.04, .33]	.115	122	*-.02*	[-.20, .16]	.809
IAT					92	**.37**	[.18, .53]	<.001	118	.20	[.02, .36]	.033
SC-IAT									94	.12	[-.08, .32]	.240

Race												
EPT	121	.25	[.07, .41]	.006	126	*.04*	[-.13, .22]	.625	141	.29	[.14, .44]	<.001
IAT					109	.26	[.08, .43]	.006	128	**.33**	[.17, .48]	<.001
SC-IAT									102	.17	[-.02, .35]	.083
Religion												
EPT	120	.03	[-.15, .21]	.713	116	-.07	[-.24, .12]	.485	141	*-.12*	[-.28, .05]	.153
IAT					94	**.21**	[.00, .39]	.047	92	.17	[-.03, .36]	.101
SC-IAT									118	.20	[.02, .37]	.030
Sexuality												
EPT	135	.21	[.04, .37]	.014	158	*.03*	[-.13, .18]	.713	154	.17	[.01, .32]	.035
IAT					114	.36	[.19, .51]	<.001	112	**.44**	[.28, .58]	<.001
SC-IAT									111	.27	[.09, .43]	.005
Weapons												
EPT	117	.23	[.05, .39]	.014	118	.10	[-.08, .27]	.291	140	*.04*	[-.12, .21]	.619
IAT					90	**.37**	[.18, .54]	<.001	117	.19	[.00, .36]	.045
SC-IAT									120	.13	[-.05, .30]	.166
Weight												
EPT	108	.03	[-.16, .22]	.759	101	*-.01*	[-.21, .18]	.887	143	.17	[.01, .32]	.043
IAT					84	**.25**	[.04, .44]	.023	117	.23	[.05, .40]	.012
SC-IAT									107	.08	[-.11, .26]	.426

Note. **Bold** font = best test-retest reliability in topic; *underlined italic* font = worst test-reliability in topic.

Figure 2

Forest Plots of Meta-Analyses of Correlations Among Implicit Measures

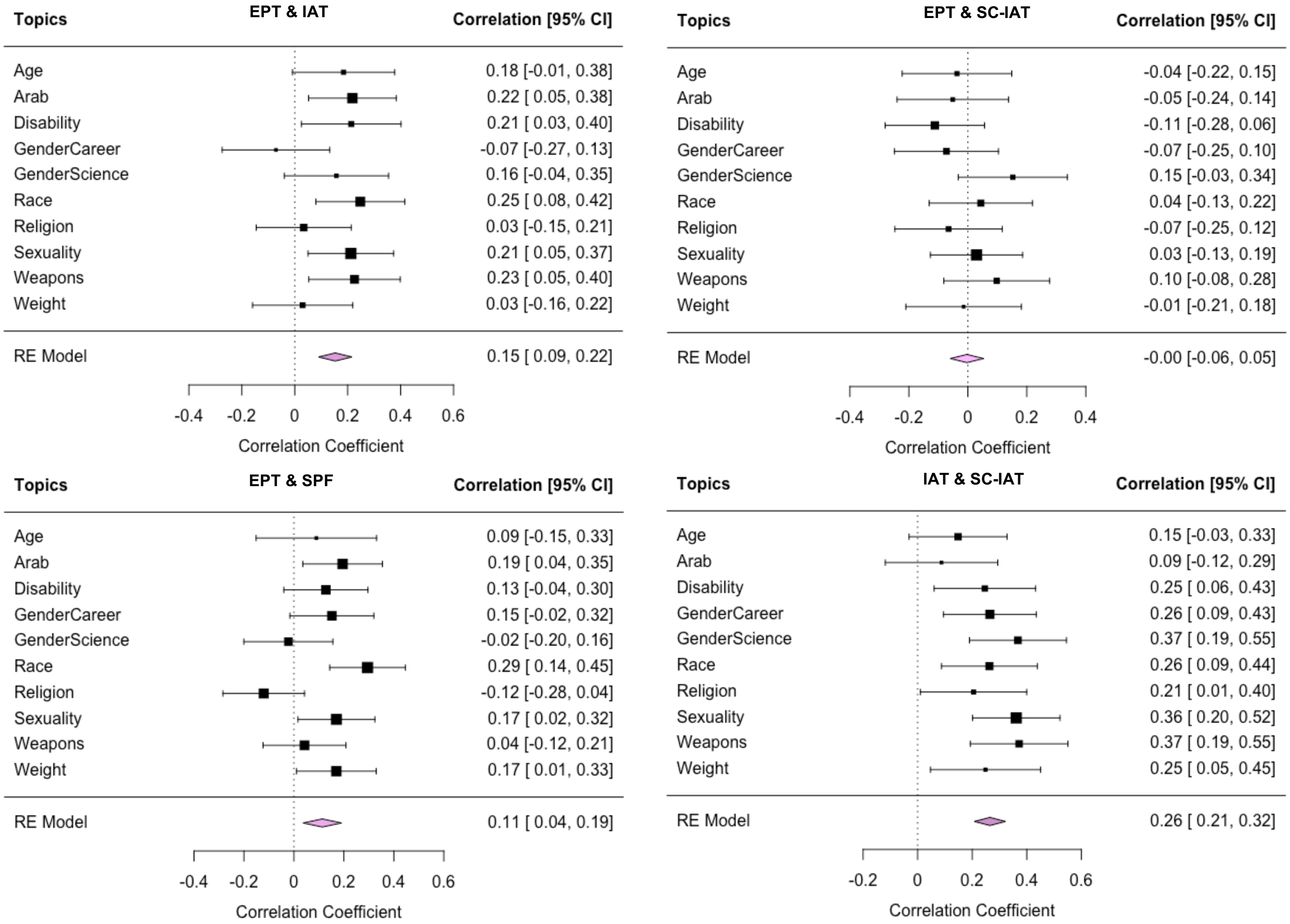

Figure 2 (continued)

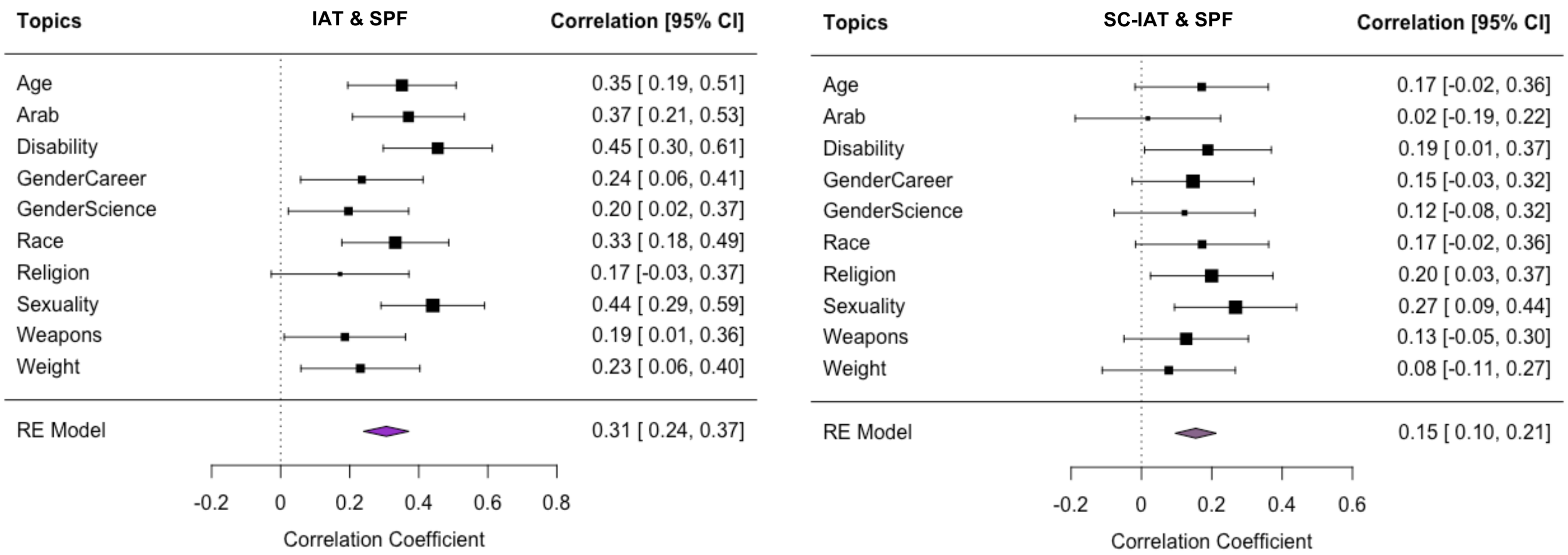

Table 6

Overall Effect Sizes for Meta-Analyses of Correlations Among Implicit Measures

	Overall Effect Size Estimate			Heterogeneity		
	r	95% CI	*p*	*Q*	*p*	I^2*(%)*
EPT & IAT	.15	[.09, .22]	<.001	11.49	.244	18.28
EPT & SC-IAT	.00	[-.06, .05]	.913	7.42	.594	0.00
EPT & SPF	.11	[.04, .19]	.004	18.60	.029	51.88
IAT & SC-IAT	.26	[.21, .32]	<.001	8.95	.442	0.00
IAT & SPF	.31	[.24, .37]	<.001	13.91	.126	35.44
SC-IAT & SPF	.15	[.10, .21]	<.001	4.60	.868	0.00

Discussion

The present study assessed the test-retest reliability and convergent validity among four implicit measures (i.e., EPT, IAT, SC-IAT, SPF) across ten attitudinal domains. Given our sample size, this work represents the largest comparative analysis of multiple implicit measures and domains. The findings of our study bolster existing knowledge about the psychometric properties of these measures and their variability across topics, which should inform theoretical perspectives on implicit social cognition and also guide future use of implicit measures in both laboratory and applied settings.

Test-Retest Reliability

When averaged across topics (see Table 2 and Table 3), we found that all implicit measures had acceptable test-retest reliability, such that all scores of a measure on Time 1 were reliably associated with scores on Time 2. However, we also found substantial differences in the strength of these associations. In particular, we found that the SC-IAT had the strongest test-retest reliability (r = .47), followed closely by the IAT (r = .43) and the SPF (r = .41). Although we hypothesized that the IAT would have the best test-retest reliability among implicit measures based on previous literature (Bar-Anan & Nosek, 2014; Greenwald & Lai, 2020), these three measures had very comparable r coefficients, which was further supported by their non-significant Wald-type tests (see Table 4). This finding can be explained in part by the fact that the SC-IAT and SPF are variants of the IAT, which is reflected in their procedural similarities involving the categorization of target items in attribute and attitude category labels.

Conversely, in line with our hypotheses and prior comparative analyses (Bar-Anan & Nosek, 2014), the EPT had significantly lower test-retest reliability (r = .19) when compared to other implicit measures. Indeed, this pattern of comparatively weaker test-retest reliability for the

EPT was found in six out of ten domains (Arab, Disability, Gender-Career, Gender-Science, Sexuality, Weapons), with the Disability and Sexuality domains even showing *negative* correlations between scores on Time 1 and Time 2. Again, we can attribute a part of the explanation to the procedural differences of the EPT with the other three implicit measures assessed in our study (i.e., IAT, SC-IAT, SPF), which do not involve a priming component. Relatedly, according to Bar-Anan and Nosek (2014), the EPT may also have a noisier psychometric performance because attitude categories (e.g., "Black People", "White People") are not explicitly mentioned to participants. The EPT method may then lead to better measurement of spontaneous evaluations of attitude items, as well as better concealment of the EPT's purpose from participants, but also weaken key aspects of psychometrics performance, such as test-retest reliability. As mentioned previously, test-retest reliability plays an important role in reducing measurement error, which may lead to potential theoretical misunderstandings concerning measures of implicit associations (Connor & Evers, 2020).

Across topics, we found significant heterogeneity among true effects (i.e., variability among topics) for the test-retest reliability of the EPT, IAT, and SPF. In other words, these measures behaved differently across their administrations over time, depending on the topic they were assessing. For instance, the test-retest reliability of the SPF was considerably stronger in the Disability topic condition than in the Religion topic condition ($r = .71$ and $r = .02$, respectively). Our findings bolster previous single-domain studies and comparative analyses showing domain-related differences for the psychometric properties of implicit measures (e.g., Bar-Anan & Nosek, 2014; Cunningham et al., 2001), and our results support the broader conclusion that variability in test-retest reliability across domains is a more widespread issue than in the limited number of topics used in prior work. However, because our analysis is the first to investigate the test-retest reliability of implicit measures assessing several topics (e.g., disability, religion), further investigation is needed to both replicate and explain our specific domain-related findings; that is, it is currently unclear why the test-retest reliability of the SPF is significantly stronger in the Disability topic condition compared to the Religion topic condition.

Finally, it is also unclear why the SC-IAT was the only measure to show homogeneity among true effects (i.e., low variability among topics). While it is possible that this homogeneity is related to the SC-IAT's strong average test-retest reliability, it is also difficult to disentangle the reasons why other measures with comparably strong test-retest reliability (i.e., IAT, SPF) showed high heterogeneity instead. One possible explanation for this finding is that the SC-IAT deals with a single attitude item (e.g., "Black People"), while all other implicit measures used here include two attitude items (e.g., "Black People" and "White People"). Since the SC-IAT used in this study focused on the more salient attitude item, it may allow for greater consistency in attitudes across administrations, resulting in lower variability among topics and higher test-retest reliability. Future research can investigate this issue directly by including SC-IATs with the less salient attitude item.

Convergent Validity

Averaged across topics (see Table 5 and Table 6), all pairs of implicit measures were significantly correlated to each other, with the exception of the EPT and the SC-IAT. The IAT was the most highly correlated with other implicit measures (mean r = .22), which was consistent with our second hypothesis and with previous literature (Bar-Anan & Nosek, 2014). Specifically, the most highly correlated pair of measures was the IAT and SPF (r = .33), with the Disability-IAT and Disability-SPF and the Sexuality-IAT and Sexuality-SPF being the strongest correlated pairs of measures among all topic conditions (r = .45 and r = .44, respectively). One explanation for the high convergent validity in the disability and sexuality topic conditions comes from past theorizing that attitude domains with stronger elements personal experience and clearer bases of comparisons (e.g., disabled versus abled) may produce greater implicit-explicit correlations (Nosek, 2007), and in turn greater associations between implicit measures. At the same time, this can be considered only a tentative explanation, given that several other domains used here matched these same criteria (e.g., weight) but failed to show greater evidence of convergent validity. Future research could make progress on this issue by identifying *a priori* factors that are believed to moderate the strength of convergent validity across domains (e.g., by directly asking participants about the salience of a category or its level of personal relevance).

Additionally, the IAT and SC-IAT were on average the pair of implicit measures with the second-highest correlation ($r = .18$). The higher convergent validity of the IAT with the SPF and SC-IAT relative to the EPT can again be attributed to the fact that the SPF and SC-IAT are variants of the IAT, such that their procedures all involve categorization of target items in attitude and attribute labels without a priming component.

Although the EPT's relatively lower average convergent validity was consistent with our second hypothesis and with existing literature (Bar-Anan & Nosek, 2014), we did not expect scores on the EPT to be *negatively* associated with scores on the SC-IAT ($r = -.03$), albeit not reliably. While we can refer to our previous discussion about the procedural differences surrounding the EPT, including the priming component and the non-explicit mention of category labels, it is also important to highlight that previous comparative analyses have found that other measures exhibiting these characteristics (i.e., the AMP) have stronger convergent validity with the IAT and its variants than the EPT (Bar-Anan & Nosek, 2014). Therefore, considering the psychometric challenges surrounding the EPT mentioned previously (e.g., low internal consistency, low test-retest reliability), the AMP might be a better alternative for researchers seeking to assess attitudinal domains without explicit mention of category labels. However, further comparative investigations of the psychometric properties of the AMP are necessary, as recent evidence also raised concerns about the measure's validity and ability to assess implicit attitudes independent from awareness of the primes' influence (Cummins et al., 2019).

The psychometric challenges surrounding the EPT has been highlighted by several other studies. Indeed, in addition to our findings about lower overall test-retest reliability and convergent validity when compared to other implicit measures, the EPT has also been shown to have lower internal consistency and weaker associations to implicit and explicit measures of the same constructs (Bar-Anan & Nosek, 2014). In a recent study, Koppehele-Gossel and colleagues (2020) found that the EPT exhibited low and unsatisfactory internal consistency (median $\alpha = .24$) throughout all ten outlier-treatment algorithms, which may reflect inherent methodological issues about the implicit measure itself rather than a feature of a specific algorithm. Although these psychometric challenges should be investigated further, future uses of the EPT within research or applied settings – especially for correlational studies that are highly dependent on the internal consistency of measures – require additional caution.

Across attitudinal domains, we found considerably less heterogeneity for convergent validity than we did for test-retest reliability. Indeed, for all pairs of implicit measures, we found homogeneity among true effects (i.e., low variability among topics), with the exception of the EPT and SPF. While these results are discordant from previous literature showing differences in measures' convergent validity depending on the topic assessed, it is also important to note that our study was restricted to topics related to intergroup attitudes and stereotypes. In the comparative analysis by Bar-Anan and Nosek (2014), the researchers selected three rather distinct constructs (i.e., race, politics, self-esteem), which relate to both intergroup attitudes and concepts related to the self. Given other single-domain studies showing low convergent validity for implicit measures of self-esteem (e.g., Bosson et al., 2000), self-related constructs may simply be more complex and multifaceted than constructs related to intergroup attitudes like race or politics. Similarly, explicit measures of self-esteem have been shown to outperform implicit measures for both individual and cross-cultural comparisons, again suggesting that implicit measures may not be a valid method to assess self-esteem (Falk et al., 2015). As such, the relatively higher homogeneity found in the present study may be accounted for by the absence of self-related domains, which might have driven the variability observed in existing comparative analyses. Nevertheless, our findings demonstrate the convergent validity of the IAT, SC-IAT, and SPF for intergroup domains, results that bolster the psychometric validity of these implicit measures for broader applications within research and applied settings.

Limitations and Future Directions

There are several limitations we would like to acknowledge in the present work. To start, we only included four implicit measures, and three of them were procedurally very similar to one another (i.e., IAT, SC-IAT, SPF). Comparatively assessing the test-retest reliability and convergent validity of other measures, such as behavioural methods like mouse-tracking (Freeman & Ambady, 2010), would allow us to get a broader understanding of the psychometric properties of implicit measures as a whole. Relatedly, although we included a wide range of attitudinal domains, we only dealt with intergroup attitudes and stereotypes. In addition to self-related domains as mentioned previously, future work could examine more clinically-oriented topics such as implicit anxiety (e.g., Egloff & Schmukle, 2002; Stieger et al., 2010) or physical/sedentary activity (e.g., Chevance et al., 2017).

Despite our large sample size, there are a few limitations to the external validity of our findings. Indeed, the majority of participants were White, female, and US citizens. These characteristics are especially important to consider, given that attitudes and stereotypes can be susceptible to environmental influences (Nosek et al., 2007). Moreover, the present study was conducted on Project Implicit, a website known for measuring implicit social cognition. Therefore, participants may already have background knowledge or familiarity with some of the implicit measures that were assessed, which could limit the generalizability of our findings, though existing research has shown that implicit measures are very difficult to fake without specific instructions to do so (Steffens, 2004; Stieger et al., 2011).

Conclusions and Implications

In the present work, we compared the test-retest reliability and convergent validity of four implicit measures (i.e., EPT, IAT, SC-IAT, SPF) across ten attitudinal domains. Averaged across topics, the EPT had significantly weaker test-retest reliability and convergent validity when compared to the other measures. Furthermore, when taking into account attitudinal domains, all measures, with the exception of the SC-IAT, showed substantial heterogeneity among true effects in meta-analytic estimates of test-retest reliability. However, all measures except the EPT retained assumptions of homogeneity in meta-analytic estimates of convergent validity.

Taken together, given our sample size, the current study is the largest comparative analysis of multiple implicit measures and domains. Our findings will not only bolster current literature regarding the comparative strengths and weaknesses of each measure, but also provide knowledge about the variability of psychometric properties across different attitudinal domains. Better understanding of the psychometric validity of these measures is crucial for theoretical advances in the field of implicit social cognition, as it can help avoid measurement error and conceptual misunderstandings. Moreover, measurement precision is also important for the increased application of implicit measures in clinical settings, such as assessing for suicidal ideation (Nock et al., 2010) or discriminatory behavior (Glover et al., 2017). Lastly, given the frequent use of implicit measures in public discourse about intergroup biases and disparities, a greater understanding of the validity of these measures will also benefit the general population.

References

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*(8), 896–906. https://doi.org/10.1177/1948550617728995

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The Sorting Paired Features Task. *Experimental Psychology*, *56*(5), 329–343. https://doi.org/10.1027/1618-3169.56.5.329

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264–1272. https://doi.org/10.1037/xge0000383

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*(6), 977–997. https://doi.org/10.1002/ejsp.487

Bogart, L. M., Bird, S. T., Walt, L. C., Delahanty, D. L., & Figler, J. L. (2004). Association of stereotypes about physicians to health care satisfaction, help-seeking behavior, and adherence to treatment. *Social Science & Medicine*, *58*(6), 1049–1058. https://doi.org/10.1016/S0277-9536(03)00277-6

Bosson, J. K., Swann Jr., 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*(4), 631–643. https://doi.org/10.1037/0022-3514.79.4.631

Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of Implicit Association Tests. *Journal of Experimental Social Psychology*, *88*, 103941. https://doi.org/10.1016/j.jesp.2019.103941

Chevance, G., Héraud, N., Guerrieri, A., Rebar, A., & Boiché, J. (2017). Measuring implicit attitudes toward physical activity and sedentary behaviors: Test-retest reliability of three scoring algorithms of the Implicit Association Test and Single Category-Implicit Association Test. *Psychology of Sport and Exercise*, *31*, 70–78. https://doi.org/10.1016/j.psychsport.2017.04.007

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (In crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345. https://doi.org/10.1177/1745691620931492

Connor, P., Sarafidis, V., Zyphur, M. J., Keltner, D., & Chen, S. (2019). Income inequality and White-on-Black racial bias in the United States: Evidence from Project Implicit and Google Trends. *Psychological Science*, *30*(2), 205–222. https://doi.org/10.1177/0956797618815441

Coulacoglou, C., & Saklofske, D. H. (2017). Chapter 3—Validity. In C. Coulacoglou & D. H. Saklofske (Eds.), *Psychometrics and Psychological Assessment* (pp. 45–66). Academic Press. https://doi.org/10.1016/B978-0-12-802219-1.00003-1

Cummins, J., Hussey, I., & Hughes, S. (2019). *The AMPeror's new clothes: Performance on the Affect Misattribution Procedure is mainly driven by awareness of influence of the primes*. PsyArXiv. https://doi.org/10.31234/osf.io/d5zn8

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., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347. https://doi.org/10.1037/a0014211

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology*, *83*(6), 1441–1455. https://doi.org/10.1037/0022-3514.83.6.1441

Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X. J., & Hsu, C.-W. (2015). Are implicit self-esteem measures valid for assessing individual and cultural differences? *Journal of Personality*, *83*(1), 56–68. https://doi.org/10.1111/jopy.12082

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*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229. https://doi.org/10.1037/0022-3514.50.2.229

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522. https://doi.org/10.1037/pspa0000160

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241. https://doi.org/10.3758/BRM.42.1.226

Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin*, *38*(5), 559–569. https://doi.org/10.1177/0146167211435981

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219–1260. https://doi.org/10.1093/qje/qjx006

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*(1), 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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. https://doi.org/10.1037/0022-3514.85.2.197

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*(1), 16–32. https://doi.org/10.1037/0022-3514.91.1.16

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology*, *87*, 103905. https://doi.org/10.1016/j.jesp.2019.103905

Krabbe, P. F. M. (2017). Chapter 7—Validity. In P. F. M. Krabbe (Ed.), *The Measurement of Health and Health Status* (pp. 113–134). Academic Press. https://doi.org/10.1016/B978-0-12-801504-9.00007-6

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. https://doi.org/10.1002/per.792

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2018). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569. https://doi.org/10.1037/amp0000364

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive validity of the implicit association test in studies of brands, consumer attitudes, and behavior. *Journal of Consumer Psychology*, *14*(4), 405–415. https://doi.org/10.1207/s15327663jcp1404_9

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the Suicidal Mind: Implicit Cognition Predicts Suicidal Behavior. *Psychological Science*, *21*(4), 511–517. https://doi.org/10.1177/0956797610364762

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

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. https://doi.org/10.1016/j.tics.2011.01.005

Nosek, B., Ranganath, K., Smith, C., Chugh, D., Olson, K., Lindner, N., Greenwald, A., Devos, T., Banaji, M., Smyth, F., & Hansen, J. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*. https://doi.org/10.1080/10463280701489053

Olson, M. A., Fazio, R. H., & Hermann, A. D. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science*, *18*(4), 287–291. https://doi.org/10.1111/j.1467-9280.2007.01890.x

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rae, J. R., & Olson, K. R. (2017). Test–retest reliability and predictive validity of the Implicit Association Test in children. *Developmental Psychology*, *54*(2), 308. https://doi.org/10.1037/dev0000437

Reinecke, A., Becker, E. S., & Rinck, M. (2010). Three indirect tasks assessing implicit threat associations and behavioral response tendencies: Test-retest reliability and validity.

Zeitschrift Für Psychologie / Journal of Psychology, *218*(1), 4–11. https://doi.org/10.1027/0044-3409/a000002

Revelle, W. (2019). *psych: Procedures for personality and psychological research*. Northwestern University, Evanston, USA. https://CRAN.R-project.org/package=psych

Rosenbusch, H., Evans, A. M., & Zeelenberg, M. (2020). Interregional and intraregional variability of intergroup attitudes predict online hostility. *European Journal of Personality*, *34*(5), 859–872. https://doi.org/10.1002/per.2301

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? *European Journal of Psychological Assessment*, *24*(4), 273–281. https://doi.org/10.1027/1015-5759.24.4.273

Spruyt, A., Lemaigre, V., Salhi, B., Van Gucht, D., Tibboel, H., Van Bockstaele, B., De Houwer, J., Van Meerbeeck, J., & Nackaerts, K. (2015). Implicit attitudes towards smoking predict long-term relapse in abstinent smokers. *PSYCHOPHARMACOLOGY*, *232*(14), 2551–2561. https://doi.org/10.1007/s00213-015-3893-2

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*(3), 165–179. https://doi.org/10.1027/1618-3169.51.3.165

Stieger, S., Göritz, A. S., & Burger, C. (2010). Personalizing the IAT and the SC-IAT: Impact of idiographic stimulus selection in the measurement of implicit anxiety. *Personality and Individual Differences*, *48*(8), 940–944.

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 Reports*, *109*(1), 219–230. https://doi.org/10.2466/03.09.22.28.PR0.109.4.219-230

Viechtbauer, W. (2010). *Conducting meta-analyses in R with the metafor package.* Journal of Statistical Software, 36(3), 1-48. https://www.jstatsoft.org/v36/i03/

Wickham et al., (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data*, *2*(1), e3. https://doi.org/10.5334/jopd.ac

Zenko, Z., & Ekkekakis, P. (2019). Internal consistency and validity of measures of automatic exercise associations. *Psychology of Sport and Exercise*, *43*, 4–15. https://doi.org/10.1016/j.psychsport.2018.12.005

Appendix A

Table A1

Attitude Labels and Item Stimuli by Topic Condition

	Attitude Category Labels (for IAT, SC-IAT, SPF)		Attitude Item Stimuli (Exemplars in IAT, SC-IAT, SPF; Primes in EPT)	
	Label 1	Label 2	Stimuli 1	Stimuli 2
Age	Old People	Young People	Images of old people (3 males, 3 females)	Images of young people (3 males, 3 females)
Arab	Arab Muslims	Other People	Words (*Hakim, Sharif, Yousef, Wahib, Akbar, Muhsin, Salim, Karim, Habib, Ashraf*)	Words (*Ernesto, Matthais, Maarten, Philippe, Guillaume, Benoit, Takuya, Kazuki, Chaiyo, Marcelo*)
Disability	Disabled Persons	Abled Persons	Images related to disabled persons (crutches, wheelchair, guide dog, blind person with cane)	Images related to abled persons (walking, running, walking on a road, skiing)
Gender-Career	Female	Male	Words (*Rebecca, Michelle, Emily, Julia, Anna*)	Words (*Ben, Paul, Daniel, John, Jeffrey*)
Gender-Science	Female	Male	Words (*Mother, Wife, Aunt, Woman, Girl, Female, Grandma, Daughter*)	Words (*Man, Son, Father, Boy, Uncle, Grandpa, Husband, Male*)
Race	Black People	White People	Images of Black people (3 males, 3 females)	Images of White people (3 males, 3 females)

Religion	Judaism	Other Religions	Images related to Judaism (menorah, star of David, Dreidel, Shabbat)	Images related to other religions (Buddha, totem, New Testament, Hindu statue)
Sexuality	Gay People	Straight People	Images (two men, two women) and words (*Gay, Homosexual, Gay People*)	Image (man and woman) and words (*Straight, Heterosexual, Straight People*)
Weapons	Black People	White People	Images of Black people (3 males, 3 females)	Images of White people (3 males, 3 females)
Weight	Fat People	Thin People	Images of fat people (4 males, 4 females)	Images of thin people (4 males, 4 females)