The role of attitude features in the reliability of IAT scores

Accepted Stage 2 Registered Report

Jamie Cummins, Ian Hussey, & Adriaan Spruyt

Ghent University, Belgium

Author note

JC, IH, Department of Experimental Clinical and Health Psychology, Ghent University and AS, Faculty of Economics and Business Administration, Ghent University. This research was conducted with the support of Grant BOF16/MET_V/002 (PI: Jan De Houwer), a postdoctoral fellowship from Ghent University awarded to IH, and Grants BOF.STG.2019.0063.01 and BOF.24Y.2019.0006.01 awarded to AS. Correspondence concerning this article should be sent to jamie.cummins@UGent.be

Abstract

Researchers commonly use the Implicit Association Test (IAT) to assess the automatic attitudes of individuals and groups. Although contended by some, the IAT is used in large part due to its psychometric properties, which are generally superior relative to most other measures of automatic cognition. Much focus has therefore been dedicated to the IAT's psychometric properties (particularly its internal consistency). However, this work has focused near-exclusively on moderators based on the procedural features of the IAT itself, and little on the varying properties of the construct under investigation within the measure. This is despite the fact that *attitude features* have already been demonstrated to influence explicit attitude measures. Here, we intend to investigate whether different attitude features can effectively predict the internal consistency of IAT scores using a large-scale IAT dataset (lowest N = 30161, highest N = 30502). We find that five of six attitude features (personal importance, degree of thinking, certainty, self-concept, and most strongly polarity) are positively related to the reliability of the IAT. Our findings have significant implications for the way in which the IAT's reliability has been conceived.

Keywords: Implicit Association Test; internal consistency; measurement; structural validity; attitude features

The role of attitude features in the reliability of IAT scores

Human beings are continuously evaluating stimuli (Fazio, 2001), and these evaluations help us to avoid threat and pursue desirable outcomes (Chen & Bargh, 1999). Given the ubiquity and importance of evaluation, a central focus of psychological science has been on how these evaluations are formed, maintained, measured, activated, and changed (De Houwer, 2009; De Houwer et al., 2001; Gawronski & Bodenhausen, 2006; Duckworth et al., 2002). In terms of the measurement of evaluations, this has typically been achieved using *direct measurement procedures* (i.e., directly asking participants about their evaluations towards relevant stimuli). Given their ease of implementation, such direct measurement procedures have been highly popular in areas of psychology which involve the study of evaluations, particularly within social psychology, and have provided psychological researchers with a great deal of utility (Robinson et al., 2013).

While direct measurement procedures have proven useful, they also come with certain limitations. In particular, the outcome of direct measurement procedures tends to reflect the operation of non-automatic processes (i.e., responses tend to be slow, controllable, intentional, and/or inefficient in terms of the cognitive resources employed; for more in-depth discussions, see De Houwer, 2006, and Moors & De Houwer, 2006). However, psychologists are often interested in evaluations because they are frequently formed, activated, and changed automatically (Ferguson & Zayas, 2009). Because of this, direct measures of evaluation are frequently poorly suited to address the questions which psychologists wish to investigate.

Fortunately, alternative methodologies have been developed which *can* capture evaluative processes under automaticity conditions. These methodologies come in the form of *indirect measurement procedures* (De Houwer et al., 2009). Indirect measurement procedures are assessment methods which seek to capture evaluations not through asking about them directly, but instead employing a putatively unrelated task from which evaluations can be inferred (Ranganath et al., 2008). For example, in the Implicit Association Test (IAT; Greenwald et al., 1998), participants are asked to simultaneously categorize some stimuli into one of two attribute categories (e.g., Pleasant or Unpleasant), and other stimuli into one of two target categories (e.g., Black or White faces). For each of these category pairs, the same two keys are involved for categorization responses (e.g., press ‘d’ for pleasant stimuli and White faces, press ‘k’ for unpleasant stimuli and Black faces). Participants are required to engage in these categorizations as quickly as possible. Critically, this configuration changes across blocks: in one block, the White and pleasant categories may share a response and the Black and unpleasant categories may share a response, while in the other block, the White and unpleasant categories may share a response and the Black and pleasant categories may share a response.

In this sense, participants are not directly asked about their evaluations: they are simply completing a categorization task. However, participants commonly show differences in response times across different blocks. For example, White participants tend to respond more quickly when the White (Black) and pleasant (unpleasant) categories share a response compared to the converse configuration (Greenwald et al., 1998). Researchers use such response time differences between blocks to make inferences about the evaluations of participants. Responding more quickly when White (Black) categories shared a response with pleasant (unpleasant) categories compared to the converse configuration, for example, is frequently taken as indicative of a tendency to evaluate White people more positively than Black people (Hussey & De Houwer, 2018; but see Schimmack, 2019). Generally, responses in the IAT are considered to reflect the operation of processes that are fast, uncontrollable, occurring without awareness, and/or unintentional (De Houwer & Moors, 2007), though counterclaims to these positions can also be made (see De Houwer et al., 2009; Hahn et al., 2014; 2019; Fiedler & Bluemke, 2005).

One common suggestion for IAT researchers has been to engage in a renewed focus on the measurement properties of the procedure. This is because the structural validity of measurement procedures (i.e., the psychometric properties of a measurement procedure without reference to other measurement procedures) is seen as critical for procedures to subsequently demonstrate external validity (i.e., meaningful relationships with other measures of putatively related constructs; see Flake et al., 2017; Loevinger, 1957). As with the validation of most measurement procedures, the structural validity of indirect measurement procedures is most commonly investigated based on their reliability (see Hussey & Hughes, 2019). In general, indirect measures exhibit mixed degrees of reliability. In terms of internal consistency, sequential priming measures such as the Evaluative Priming Task tend to perform relatively poorly (Cronbach's alpha < .50; Bar-Anan & Nosek, 2014; Gawronski et al., 2009). By contrast, misattribution procedures such as the Affect Misattribution Procedure (Payne et al., 2005; Payne & Lundberg, 2014) and the Truth Misattribution Procedure (Cummins & De Houwer, 2019) tend to demonstrate high internal consistency (Cronbach's alpha typically above .90). However, much lower estimates for these coefficients are also seen (Bar-Anan & Nosek, 2014; see also Cummins et al., 2019). Indeed, Bar-Anan and Nosek (2014) demonstrated that, in a large-scale comparison of seven indirect measurement procedures across four different stimulus domains, only the Implicit Association Test (and its shortened equivalent, the Brief IAT) demonstrated satisfactory internal consistency in all measurement contexts. In general, the IAT and its variants demonstrate internal consistency which is comparable to direct measures of evaluations (around .85; Gawronski & De Houwer, 2014). Notably, many argue that this reliability coefficient is overestimated due to systematic error variance (e.g., Meissner et al., 2019). Nevertheless, the IAT is often selected for use over other indirect measures in part due to the likelihood that it will demonstrate higher internal consistency (i.e., better measurement properties) than alternatives measures.

Given the importance of the IAT's internal consistency to its users, determining the conditions under which the measure achieves the highest internal consistency is of interest. It is now clear that internal consistency of the IAT varies as a function of the procedure's length (Sriram & Greenwald, 2009), how scores are quantified (Greenwald et al., 2003), and a variety of other method variables (Nosek et al., 2005). Notably, all of this previous work has focused on the relationship between internal consistency and the features of the measurement procedure. However, an additional relationship has thus far failed to be considered: the one between the internal consistency and the features of the *attitude* being investigated. If the features of the attitude being investigated affect internal consistency estimates, then this could have drastic implications for when and how indirect measurement procedures are used. For example, just as it is difficult to generalize about Likert scales *in general* (i.e., agnostic from their items and response options), we caution against the tendency in the literature to attempt to estimate the internal consistency of the IAT *in general* (i.e., agnostic from the features of the stimuli employed in it). An IAT may exhibit very different measurement properties in the subset of participants for whom the attitude under investigation is highly salient compared to the subset of participants for whom the attitude is less salient, for example. Nonetheless, we acknowledge the need for researchers to be able to make heuristic judgments about whether a given measure is a good candidate for use in their research or not. As such, in order to reconcile these two points, we argue that there are likely to be properties of the person-stimulus relations (e.g., the personal importance of stimuli to individuals) that are predictive of the internal consistency that IAT will demonstrate.

Attitude features have already been explored in the context of predicting the *strength* of explicitly measured attitudes (in terms of magnitude and stability over time). In a recent review, Luttrell and Sawicki (2020) identified attitude certainty (the degree to which participants are certain of attitudes), personal importance (the degree to which the attitudes held by the participant are important to them), elaboration (degree of thinking about the attitude objects), and distinctiveness of attitude objects (among others) as key predictors of attitude strength. Others have also suggested that the degree to which attitudes are associated with self-concept influence attitude strength (Pomerantz et al., 1995; Eagly & Chaiken, 1995). In the context of implicit measures, less work on this front has been done. However, Spruyt et al. (2018) recently demonstrated that automatic stimulus evaluation in the absence of an explicit evaluative processing goal occurs only for attitude objects that are of personal importance to the observer (in line with findings related to personal importance in explicit attitudes).

Given that attitude features influence attitude strength, it seems reasonable to conjecture that these same attitude features may also likely play a role in predicting internal consistency of IAT scores (and scores in implicit measures more generally). By identifying which (if any) attitude features are most predictive of internal consistency in the IAT, this can provide IAT researchers with a means of identifying the participants (at the individual level) and domains (at the group-level) in which the IAT will likely perform best. In short, this would provide researchers with a toolkit for (i) identifying the subset of participants for whom the IAT will produce meaningful results, and (ii) knowing when and how to calibrate the IAT at the group-level to achieve satisfactory measurement properties.

We will pursue this question through the use of the large-scale Attitudes, Identity, and Individual Differences (AIIDs) dataset (*N* > 400,000), which was collected through the Project Implicit website as part of a larger study (Hussey et al., 2019). This study involved a planned missing-data design (for further information see Graham et al., 2006), with participants completing one IAT containing stimuli from one of 95 general content domains (for example, 50 Cent vs. Britney Spears, White American vs. African American, Individual vs. Collective), a subset of explicit measures, and one of fifteen commonly used individual differences measures. Notably, several explicit questions related directly to features of the attitude assessed within the IAT (e.g., personal importance, attitude polarity). Combined with the extremely large sample collected for the study, this provides an ideal context to inquire into the relationship (if any) between several attitude features and the internal consistency of IAT scores. We generally expect predictive relationships between the attitude features and IAT internal consistency.

Method

All data processing and analysis scripts for planned analyses in this paper can be found on the Open Science Framework (https://osf.io/2tpvg).

Participants. The AIID dataset has been divided into exploratory and confirmatory subsets, with the confirmatory subset being roughly 5 times larger than the exploratory. Only the exploratory dataset was made publicly available prior to the submission of this Registered Report; the confirmatory dataset was provided to us after Stage 1 acceptance and used for confirmatory testing. We excluded participants who failed to meet any of the following criteria: age 18-65, fluent English, completed an evaluative IAT, performance inclusion criteria for IAT data (detailed below), and completed between 1 and 3 of the attitude feature questions included in our analyses[1]. If participants completed the same IAT more than once, we included only the first of those IATs which they completed. Analyses on the exploratory dataset consisted of 6440 and 6644 total experimental sessions (ranging from 6018 to 6249 individual participants; this range was because participants only ever completed one attitude features question, and there was variation in how many people completed each scale). The confirmatory dataset consisted of between 30161 and 30502 total experimental sessions (ranging from between 24406 to 24640 total individual participants).

Measures.

(Evaluative) Implicit Association Tests. In the IAT, participants were required to categorise attribute and target stimuli using the 'E' and 'I' keys on the computer keyboard. Attribute category labels varied between IATs, consisting of either positive/negative, good/bad, or pleasant/unpleasant labels[2]. Target category labels varied depending on the attitude domain being assessed. Each IAT began with an initial block of 20 trials, wherein participants categorized only the target stimuli. Participants next completed a second block of another 20 trials, this time only categorizing the attribute stimuli. Following this, participants completed two blocks (20 trials and 40 trials, respectively) in which they were required to categorise both the target and attribute stimuli simultaneously. In this block, one of the attribute labels shared a response key with one of the target labels, while the remaining attribute and target labels shared the other response key. The specific response arrangement required was varied across participants, such that some participants completed one arrangement first, and others completed

[1] This variation was because these questions were present as part of a pool of potential questions to be asked of participants, with some of these questions unrelated to our analyses.
[2] An additional IAT, measuring personal identification with the stimuli, was also employed in the design of the study (using Me/Others labels). However, this IAT was not of interest to our research question (this is discussed in further detail below).

the other arrangement first (i.e., blocks were counterbalanced). In a fifth block of 20 trials, participants then categorized only the target stimuli again. However, the response keys required for this categorization were switched relative to those in the first block. Participants then completed another two blocks with this new response key requirement (20 trials and 40 trials, respectively), again categorizing both the attribute and target stimuli.

On each trial of the IAT, a single stimulus was presented in the centre of the screen until participants emitted a response. If they responded correctly, then the stimulus was removed from the screen. On trials involving both targets and attributes, stimuli were randomised in such a way that every first trial consisted of a target stimulus, while every second trial consisted of a category stimulus. If the participant responded incorrectly then they were presented with a red X on-screen below the stimulus until they corrected their response, at which point both the red X and the stimulus were removed from the screen. Each trial was separated by a 500ms intertrial interval.

Measures of Attitude Features. For all the measures of attitude features, participants were required to answer based on a 6-point Likert scale for *both* target categories. For assessing *personal importance*, participants were asked “How personally important are your feelings towards [target category]?” with Likert anchors of 1 = not at all important and 6 = very important. For assessing *thinking*, participants were asked “How much do you think about your feelings towards [target category]?” with Likert anchors of 1 = not at all and 6 = a lot. For assessing *attitude certainty*, participants were asked “How certain are you about your feelings towards [target category]?” with Likert anchors of 1 = not at all certain and 6 = very certain. For assessing *attitude stability*, participants were asked “How much do you expect your feelings towards [target category] to change over time?” with Likert anchors of 1 = not at all and 6 = a lot. For assessing *attitude self-concept*, participants were asked “How much is [target category] part of your self-concept?” with Likert anchors of 1 = not at all and 6 = very much. Finally, for assessing *attitude polarity*, participants were asked to rate their agreement with the statement “Having positive feelings towards [target category 1] implies having negative feelings towards [target category 2]” with Likert anchors of 1 = Strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = slightly agree, 5 = agree, and 6 = strongly agree.

Following the suggestions of reviewers, we calculated Cronbach’s alpha for each of these scales in the confirmatory sample based on responses to each of the two stimulus categories for each scale. Importance (alpha = .78), thinking (alpha = .74), stability (alpha = .76), polarity (alpha = .91), and certainty (alpha = .77) exhibited values above .7, whereas self-concept (alpha = .46) did not.

Procedure. Prior to the completion of the study, participants created login details at the Project Implicit website and provided basic demographic information. Participants who were allocated to complete the AIIDs study next gave consent for their participation in the study. Then, the participant completed an IAT. IATs varied in terms of the attitude domain assessed (chosen randomly from one of 95 domains) and in terms of the type of IAT which was presented (three different IATs with different evaluative terms, and one self-identity IAT). Participants then completed self-report questions relating to the same attitude domain as assessed by the IAT. Self-report questions varied across participants such that each participant completed a subset of questions derived from an overall battery. Finally, after completing these self-report questions, participants completed one of twenty randomly assigned, commonly used individual difference measures. In this study, our analyses are focused solely on data from the evaluative IATs and the six questions relating to attitude features from the self-report measures.

Planned analyses.

Analytic strategy.

The AIIDs dataset was divided into an initial exploratory subset, and a much larger confirmatory subset. We used the exploratory subset to provide an initial general sense of whether each of the attitude features was related to internal consistency, and the confirmatory subset to gain precision in our effect size estimates. Given that *p*-value significance tests are not particularly meaningful with large samples, we focused primarily on the estimates of effect size for each of the attitude features and compared these effect size estimates by determining whether the point estimate of one effect size fell outside of the 95% confidence intervals of another effect size.

We calculated internal consistency of the IAT using Cronbach's alpha (Cronbach, 1951) using a common parcelling method for IAT data (i.e., three IAT D scores, calculated based on the first, middle, and last 20 congruent and incongruent trials completed by participants). Although other parcelling strategies are used in the calculation of reliability (e.g., four parcels, or odd-even splits) we opted for this three-parcel strategy because (i) it is commonly used in the IAT literature, and (ii) the AIIDs dataset already had this three-parcel approach implemented within its processed data. For an in-depth discussion of these parcelling issues, see De Schryver et al. (2018). Although other, arguably superior metrics for internal consistency exist (e.g., omega, see McDonald, 1999), we opted for the use of Cronbach's alpha primarily because this is the most used metric of internal consistency in the IAT literature. As such, its use here too allowed for comparability to previous work which has also analysed the IAT's internal consistency, while the use of superior but less commonly used methods would have created greater ambiguity in terms of how comparable and applicable our results were to previous work on this topic. In calculating internal consistency, we employed bootstrapping to ensure a more robust estimate of the Cronbach's alpha value (see Mooney & Duval, 1993). Bootstrapping involves creating multiple sets of data ("bootstraps") from an initial set of data (i.e., random sampling with replacement). For our analyses, three IAT D scores (or Rusio's A score, see below) based on the three parcels were then calculated for each bootstrap. Internal consistency values were then calculated for each individual bootstrap by estimating Cronbach's alpha from the pairwise correlations between these parcels (using the *psych* package in R), with the median Cronbach's alpha value of these bootstraps taken as the estimate for Cronbach's alpha, and 95% confidence intervals computed using the percentile method (see Puth et al., 2015, for an extensive discussion of different methods for computing confidence intervals using bootstrapping).

Our research question relates to variation in the Cronbach's alpha of IAT scores as a function of the different attitude features. Given that Cronbach's alpha is a group-level statistic, a grouping of participants is required to produce such an estimate. As such, a key analytic question lies in how participants are stratified to produce these estimates. In this manuscript, we approached this in two different ways: by stratifying participants based on the mean score of each attitude feature, and stratifying based on the mean attitude feature score across the 95 different domains within the AIIDs dataset. We explain our rationale for both strategies below.

The most direct way of determining the relationship between reliability of IATs and attitude features of the target domains is to simply stratify participants based on the mean rating for each attitude feature and then calculate reliability coefficients for each of these groups. Mean scores on each attitude feature question were calculated by simply taking the mean of the two responses on the Likert scales for each attitude feature. For instance, for personal importance, the mean score was calculated by summing the responses for “How personally important are your feelings towards [Target category X]?” and “How personally important are feelings towards [Target category Y]?” and dividing by two. However, this approach comes with a drawback: if the mean of these scores is used as the basis for stratification, then only 11 data points are available for the calculation of the relevant correlation coefficient in our data. This limits the power of our model to detect the relevant correlation, with .8 power to detect a correlation of r = .74.

To overcome this, an alternative method was also used: a between-groups design using the mean of each attitude feature for each attitude domain. This represents a less direct approach to addressing our research question, but one which produces better estimates of the relationship between variables. This involved firstly stratifying participants by the domain of the target stimuli used in the IAT (which will yield 95 different groups, providing .8 power to detect a correlation of r = .29). Then, a reliability coefficient was estimated for each group, as well as a mean attitude feature score of each domain for each feature (i.e., by taking the mean of each attitude feature for each content domain). Reliability coefficients for each domain were then correlated with scores for each feature of those same domains. While this yields a better-estimated correlation coefficient, it only indirectly assesses attitude features (via global averages of each domain).

In the current work, we addressed our research question using both analytic strategies. Note, that, in addition to using bootstrapping for the robust estimation of the reliability estimates of each group, we also used bootstrapping in estimating the correlation coefficient itself for robustness. Our primary analysis of interest was the analysis which involved stratifying participants by mean ratings (for each of the six attitude features). However, the use of domain stratification served as an informative secondary analysis for the event in which the correlation effect sizes in our primary analysis would have proven difficult to interpret. Ultimately, our interpretation of whether internal consistency is related to specific attitude features was hinged most primarily on the analysis stratifying groups by mean attitude feature scores (i.e., the direct method with fewer data points).

Additionally, although IAT effects are commonly computed using the D1 scoring algorithm (Greenwald, Nosek, & Banaji, 2003), recent research has suggested that other effect sizes metrics have better psychometric properties (e.g., robustness to outliers, which are common in reaction time data). In particular, the Ruscio's A score (Ruscio, 2008; also referred to as the probability of superiority metric), which can be seen as a special case of probabilistic index models (Thas et al., 2012), has been shown to exhibit superior psychometric properties and is less sensitivity to outliers compared to the *D* score in the context of another implicit measure (De Schryver et al., 2018; for its use within the IAT see Cummins et al., 2021) Ruscio's A refers specifically to the probability that a randomly selected response time in one block will be larger than a randomly selected response time in another block. For example, an A score of .7 indicates the probability of incongruent trials having a longer RT than congruent trials (i.e., this probability value is .7). As its description implies, in the context of the IAT this score is calculated by randomly selecting a response time from the congruent block, and a response time from the incongruent block. If the response time from the incongruent block is larger than that of the congruent block, a value of '1' is recorded. If the opposite is true, a value of '0' is recorded. This procedure is repeated across several thousand combinations of congruent and incongruent trials. After this, the sum of the recorded values is divided by the total number of congruent-incongruent comparisons. Thus, a Ruscio's A score above .5 would suggest that incongruent trials were, in general, slower than congruent trials. An A below .5 would imply the converse. Readers should note that for a given domain, the block on which most participants tended to have shorter response times was designated the congruent block, while the block which participants were generally slower to respond on was designated the incongruent block (this designation is already present within the AIID dataset; see Hussey et al., 2019, for further information).

For the purposes of this research, our primary analyses of interest related to reliability as calculated using the IAT D1 score (and this was the primary score of interest to addressing our research question). However, we also reported these same analyses using the Ruscio’s A score for those interested in this alternative scoring algorithm. Doing so also functioned as a robustness analysis for our primary analyses using an alternative algorithm. Note however that our interpretation of results is based primarily on analyses of the D1 scores.

As mentioned previously, we used bootstrapping with 1000 iterations for the robust estimation of both Cronbach’s alpha and the correlations between Cronbach’s alpha and the different attitude features. To do this, within each bootstrap of the data, we calculated bootstrapped Cronbach’s alpha values for each level of each attitude feature, and then correlated these Cronbach’s alpha values with each level of each attitude feature. Because we opted for this bootstrapping procedure, we interpreted “significance” within each subset of the data (i.e., exploratory and confirmatory) on the basis of confidence interval estimates. Specifically, we considered correlations “significant” if the lower-bound 95% confidence interval excluded (i.e., is greater than) zero.

Data preparation. We conducted our analyses only on those participants who completed an evaluative IAT. While an argument might be made that these effects might also extend to self-identity as well as stimulus evaluation, this was not the primary process of interest in this analysis. In line with previous recommendations (Nosek et al., 2007) and with the strict exclusion strategy recommended for the AIIDs dataset, participants were excluded when they met the following criteria in the IAT: (i) > 35% of responses < 300ms in any practice block, (ii) > 25% of responses <300ms in any critical block, (iii) >10% of responses <300ms across all critical blocks, (iv) > 50% error rate in any given practice block, (v) >40% error rate across all practice blocks, (vi) >40% error rate in any given critical block, (vii) >30% error rate across all critical blocks, and (viii) >10% of responses >10000ms in any given critical block (this final criterion was added by Hussey et al., 2019, for stringency).

Exploratory subset.

In line with the submission guidelines for Registered Reports relating to the AIIDs dataset, we firstly conducted our proposed analyses on the exploratory dataset provided by Hussey et al. (2019) prior to Stage 1 acceptance. These analyses are reported in the supplementary materials.

Confirmatory subset.

For both analyses we predicted that the lower-bound confidence interval for the correlation between each attitude feature and internal consistency scores will be greater than zero.

Stratifying based on attitude feature scores. To determine whether Cronbach's alpha values varied as a function of the different attitude features of the evaluated stimuli, we computed bootstrapped correlations between each attitude feature and the relevant internal consistency values. We also used an identical analytic strategy to the primary analysis, but now estimating internal consistency based on the A score (rather than the D1 score). Plots for this analytic approach based on internal consistency computed with D scores can be seen in Figure 1.

Stratifying based on domain. Here we firstly computed mean scores for each attitude feature and for each of the target domains. We then used bootstrapping to calculate the correlation between internal consistency scores and attitude feature scores of each of the 95 domains. We also conducted an identical procedure to the primary analysis with stratification based on domain, but again computing reliability using Ruscio's A rather than the D score. Results from the confirmatory dataset for both analytic strategies are detailed in Table 1.

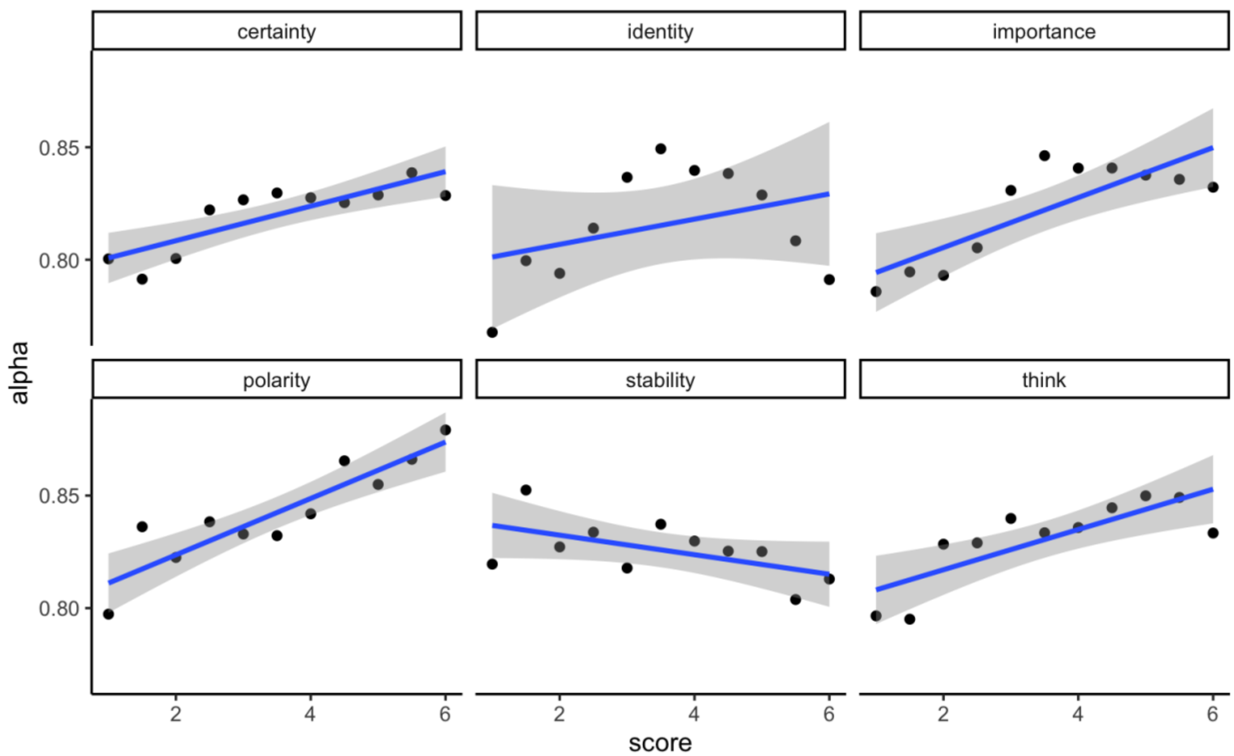

Figure 1. Scatterplots for the relationship between levels of each attitude feature and corresponding Cohen's d-based Cronbach's alpha values (i.e., stratifying based on attitude feature scores).

Table 1. Results using the D1 score and the Ruscio's A score when stratifying based on (i) attitude features and (ii) domains in the confirmatory dataset.

Attitude feature	Stratifying based on attitude features		Stratifying based on domain	
	r	95% CIs	r	95% CIs
	D1			
Personal Importance	.78	[.63, .87]	.21	[.16, .26]
Thinking	.76	[.56, .88]	.17	[.22, .28]
Certainty	.68	[.27, .86]	.07	[.02, .13]
Stability	-.46	[-.70, .05]	.21	[.16, .26]
Self-Concept	.35	[.10, .57]	.23	[.17, .28]
Polarity	.87	[.73, .96]	.54	[.50, .58]
	Ruscio's A			
Personal Importance	.83	[.72, .90]	.31	[.27, .36]
Thinking	.78	[.60, .89]	.31	[.16, .35]
Certainty	.71	[.43, .88]	.07	[.02, .12]
Stability	-.33	[-.66, .22]	.28	[.24, .33]
Self-Concept	.52	[.30, .68]	.33	[.27, .37]
Polarity	.89	[.75, .96]	.56	[.52, .59]

Correlation between attitude features. Following the Stage 2 submission of our manuscript, one reviewer requested that we examine the interrelations between the different attitude features to provide greater context to the pattern of results above. Since participants only ever completed 3 of the 6 attitude feature measures, we instead opted to analyse this question by examining the correlation between attitude features when stratified by domain. These correlations are detailed in Table 2.

Table 2. Correlations between the mean domain-stratified scores of the attitude features.

	Certainty	Self-Concept	Importance	Polarity	Stability	Thinking
Certainty	1					
Self-Concept	.10	1				
Importance	.22*	.89*	1			
Polarity	-.07	.09	.26*	1		
Stability	-.23*	.65*	.64*	.14	1	
Thinking	.19	.83*	.96*	.26*	68*	1

* p < .05

Discussion

The aim of this registered report was to assess the degree to which the reliability of the IAT, as indexed by the Cronbach's alpha, is related to six different attitude features. Overall, we found robust correlations between different attitude features and IAT reliability which were consistent across both different analytic strategies, as well as across two different IAT scoring methods. The specifics of these correlations, as well as their implications and utility, are discussed below.

Summary and interpretation of results

Across both analytic strategies and scoring methods, the polarity of the attitude domain was the one attitude feature whose (positive) correlations with reliability were consistently above a conventionally "strong" correlation effect size (i.e., $r > .5$). Notably, the strength of this correlation dropped substantially (from .87 to .54) from stratifying based on scores on attitude features vs. stratifying based on domains. Indeed, this was a relatively consistent pattern: correlations in general were dramatically smaller for the second analytic strategy compared to the first. However, this is not surprising given that the second strategy represented a noisier means of addressing our research question (as discussed above). Notably, attitude stability's estimate when using the first analytic strategy included zero, indicating that the correlation was not significantly different from zero. Estimates for all attitude features, except for attitude stability, excluded zero, indicating that they were positively correlated with reliability. The attitude feature most weakly (but positively and significantly) related to reliability was self-concept, followed by attitude certainty, personal importance, thinking, and polarity (whose lower-bound CIs all exceeded the point estimate of self-concept). These four features contained one another's point estimates within their CIs with exception of polarity; the point estimate of polarity exceeded the upper-bound CI of attitude certainty but was included in the CIs of importance and thinking.

In the context of the second analytic strategy, polarity was starkly superior; its lower-bound CI drastically exceeded not only the point estimates of all other attitude features, but also their upper-bound CIs, positively correlating with reliability. Personal importance, thinking, self-concept, and stability again all contained one another's point estimates within their CIs, whereas the point estimate for certainty was below the lower-bound CIs of these other features. Ironically, attitude stability was the only feature whose lower-bound estimate was not stably above zero in both analytic strategies.

Implications of results

In the first instance, our results show that single self-report items assessing different attitude features are positively related to the reliability of IAT scores, even at the level of domains. This observation has implications at the level of the group as well as the level of the individual. At the level of the group, these results have interesting implications for the optimisation of measurement properties. The IAT in general exhibits relatively mixed results in terms of reliability across various studies (Kurdi et al., 2019). This can be problematic, particularly from a psychometric perspective where reliability is a critical first step towards the development of a valid measurement instrument (Flake et al., 2017; Van Dessel et al., 2020). Our results offer a means of determining in advance whether a newly developed IAT will exhibit satisfactory reliability, the knowledge of which may then be factored into the subsequent design of the IAT. For example, suppose a researcher seeks to develop a reliable IAT measuring automatic preferences in a novel context (e.g., assessing preferences between country musician Daniel O'Donnell and pop singer Justin Bieber). Without our findings, the researcher may invest time and resources into running their study, only to find that the reliability of their IAT was poor: a finding that may have been knowable in advance of running the study by simply assessing a small number of attitude features for this population regarding this domain of interest.

A potentially intuitive extension of this point would involve trying to prospectively adjust features of the to-be-used IAT to ameliorate these reliability issues (e.g., increasing the number of trials in the task, etc.). However, it is important to note that there is a second element to consider: the reliability of the task at the level of the individual. That is, the properties of the procedure which may in principle enhance reliability (e.g., increasing the number of trials) are likely eclipsed by the inter-individual variation in attitude features. Put another way, our results illustrate that the reliability of the IAT may be best understood as an indicator of the features of the attitude under examination, rather than as an indicator of the properties of the task itself. These features in themselves can vary within individual members of the population. As such, claims regarding reliability may not even be made to IATs assessing specific domains, because this reliability coefficient will be determined by *the specific individuals* being assessed. This will be particularly problematic for domains within which there is substantial inter-individual variation in attitude features.

This also serves to highlight a related, often-overlooked point: discussing the reliability of “The IAT”, or any measure, is meaningless in the absence of knowledge about (i) the stimuli being used within the measure, and (ii) the context in which the measure is being utilised. Although most researchers are declaratively aware of this point, it is easy to forget in practice. Indeed, as Brick et al. (2021) point out, psychologists tend to *essentialise* within their scientific practices, and this tendency to essentialism also includes the measures that we use. In other words: our findings echo the critical importance of avoiding treating “The IAT” as a monolith, and to instead bear in mind that the validity/reliability of the measure is best understood in terms of the specific contexts within which it is used.

Limitations and future directions

Our study has the benefit of having examined (i) multiple attitude features across (ii) almost 100 different domains, assessing robustness across (iii) different analytic strategies and (iv) different methods of scoring. However, our study was limited to examining responding only in the context of a single measure (i.e., the IAT). Although the IAT is very commonly used to measure automatic preferences, other measures (such as the AMP; Payne et al., 2005) are also staples of such research. In particular, the AMP differs from the IAT in the sense that the categories of stimuli being assessed are not explicitly treated as relative to one another within the procedure (Znanewitz et al., 2018). In the case where less relativistic responses are required to complete the task, it is easy to imagine that the polarity of the attitude objects may be less salient, and that other features (e.g., personal importance, self-concept) may instead come to the fore. Indeed, *relational implicit measures* (such as the Propositional Evaluation Paradigm and Truth Misattribution Procedure) may be affected by attitude features even differently still (Cummins & De Houwer, 2019; 2020; Müller & Rothermund, 2019). As such, future research should focus on systematically investigating these issues within the context of these other measures.

One further issue worth pursuing relates to the relative contribution of procedural features vs. attitude features in informing reliability. As formerly mentioned, researchers may in principle try to change procedural parameters of the IAT to prospectively ameliorate potential reliability issues. However, the extent to which this may be effective is unclear. It may be the case that most of the variation in reliability is attributable to (interindividual) attitude features, and not procedural features. Future research should seek to investigate this, which could also shed further light on the extent of the role which attitude features play in informing the reliability of these procedures.

Conclusion

Using a large dataset consisting of IAT data from 95 different attitude domains, and measures of 6 different features of attitudes (personal importance, polarity, self-concept, thinking, stability, and certainty), we investigated the degree to which each of these attitude features was related to the reliability of scores in the IAT. Across two analytic strategies and two different scoring methods, attitude polarity was most strongly and robustly related to the reliability of the IAT. Personal importance, self-concept, thinking, and certainty, although at times yielding smaller effect sizes than polarity, exhibited consistent non-zero correlations with reliability. Attitude stability showed mixed results across the analyses. Our results provide IAT researchers with a potential means of prospectively assessing the measurement properties of the IAT in a novel domain, and most critically demonstrate that the reliability of the IAT is reflective of the features of the attitude under investigation, rather than an objective property of the task itself.

Open Practices

The data, processing, and analysis scripts, as well as the Stage 1 submission relating to this manuscript, can be found on the Open Science Framework here: https://osf.io/uga8j/.

References

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

Brick, C., Hood, B., Ekroll, V., & de-Wit, L. (2021). Illusory essences: A bias holding back theorizing in psychological science. *Perspectives on Psychological Science.* 1-16. https://doi.org/10.1177/1745691621991838

Chen, M., & Bargh, J. A. (1999). Consequences of Automatic Evaluation: Immediate Behavioral Predispositions to Approach or Avoid the Stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cummins, J., & De Houwer, J. (2019). An inkblot for beliefs: The Truth Misattribution Procedure. *PLOS ONE*, *14*(6), 0218661. https://doi.org/10.1371/journal.pone.0218661

Cummins, J, & De Houwer, J. (2020). The Shape of Belief: Developing a Mousetracking-Based Relational Implicit Measure. *Social Psychological and Personality Science*, 1948550620978019. https://doi.org/10.1177/1948550620978019

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. *Preprint*. https://doi.org/10.31234/osf.io/d5zn8

Cummins, J, Lindgren, K. P., & De Houwer, J. (2021). On the role of (implicit) drinking self-identity in alcohol use and problematic drinking: A comparison of five measures. *Psychology of Addictive Behaviors*, *35*(4), 458–471. https://doi.org/10.1037/adb0000643

De Houwer, J. (2006). What Are Implicit Measures and Why Are We Using Them? In *Handbook of implicit cognition and addiction* (pp. 11–28). https://doi.org/10.4135/9781412976237.n2

De Houwer, J. (2009). How do People Evaluate Objects? A Brief Review. *Social and Personality Psychology Compass*, *3*(1), 36–48. https://doi.org/10.1111/j.1751-9004.2008.00162.x

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink (Ed.), *Implicit Measures of Attitudes* (pp. 179–194). Press.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

De Houwer, J., Thomas, S. P., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853–869. https://doi.org/10.1037/0033-2909.127.6.853

De Schryver, M., & De Neve, J. (2018). A tutorial on probabilistic index models: Regression models for the effect size P(Y1 < Y2. *Psychological Methods*. https://doi.org/10.1037/met0000194

De Schryver, M., Hughes, S., De Houwer, J., & Rosseel, Y. (2018). On the Reliability of Implicit Measures. In *Current Practices and Novel Perspectives* (Preprint). https://doi.org/10.31234/osf.io/w7j86

Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The Automatic Evaluation of Novel Stimuli. *Psychological Science*, *13*(6), 513–519. https://doi.org/10.1111/1467-9280.00490

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude Strength: Antecedents and Consequences*. Psychology Press.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An. *Cognition and Emotion*, *15*(2), 115–141.

Ferguson, M. J., & Zayas, V. (2009). Automatic Evaluation. *Current Directions in Psychological Science*, *18*(6), 362–366.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and Unaided Response Control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*(4), 307–316. https://doi.org/10.1207/s15324834basp2704_3

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2009). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition and Emotion*, *24*(6), 1008–1025. https://doi.org/10.1080/02699930903112712

Gawronski, B., & De Houwer, J. (2014). Implicit Measures in Social and Personality Psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). Cambridge University Press.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

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.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hussey, I., & De Houwer, J. (2018). *Completing a Race IAT increases implicit racial bias*. https://doi.org/10.31234/osf.io/vxsj7

Hussey, I., & Hughes, S. (2019). Hidden invalidity among fifteen commonly used measures in social and personality psychology. In *Advances in Methods and Practices in Psychological Science*. https://psyarxiv.com/7rbfp/

Hussey, I., Hughes, S., & Nosek, B. A. (2019). *Attitudes, Identities and Individual Differences: A large dataset for investigating relations among implicit and explicit attitudes and identity*. Open Science Framework. https://osf.io/pcjwf

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Luttrell, A., & Sawicki, V. (2020). Attitude strength: Distinguishing predictors versus defining features. *Social and Personality Psychology Compass*, *14*(8), e12555. https://doi.org/10.1111/spc3.12555

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting Behavior With Implicit Measures: Disillusioning Findings, Reasonable Explanations, and Sophisticated Solutions. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02483

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. SAGE.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Müller, F., & Rothermund, K. (2019). The Propositional Evaluation Paradigm: Indirect Assessment of Personal Beliefs and Attitudes. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02385

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

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). Psychology Press.

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, 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. https://doi.org/10.1111/spc3.12148

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology*, *69*(3), 408–419. https://doi.org/10.1037//0022-3514.69.3.408

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology*, *84*(4), 892–897. https://doi.org/10.1111/1365-2656.12382

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*(2), 386–396. https://doi.org/10.1016/j.jesp.2006.12.008

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (2013). *Measures of Personality and Social Psychological Attitudes: Measures of Social Psychological Attitudes*. Academic Press.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Schimmack, U. (2019). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, 1745691619863798. https://doi.org/10.1177/1745691619863798

Spruyt, A., Tibboel, H., De Schryver, M., & De Houwer, J. (2018). Automatic stimulus evaluation depends on goal relevance. *Emotion*, *18*(3), 332–341. https://doi.org/10.1037/emo0000361

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

Thas, O., Neve, J. D., Clement, L., & Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, *74*(4), 623–671. https://doi.org/10.1111/j.1467-9868.2011.01020.x

Van Dessel, P., Cummins, J., Hughes, S., Kasran, S., Cathelyn, F., & Moran, T. (2020). Reflecting on 25 Years of Research Using Implicit Measures: Recommendations for Their Future Use. *Social Cognition*, *38*, s223–s242. https://doi.org/10.1521/soco.2020.38.supp.s223

Znanewitz, J., Braun, L., Hensel, D., Altobelli, C. F., & Hattke, F. (2018). A critical comparison of selected implicit measurement methods. *Journal of Neuroscience, Psychology, and Economics*, *11*(4), 249–266. https://doi.org/10.1037/npe0000086