The individual-level precision of implicit measures

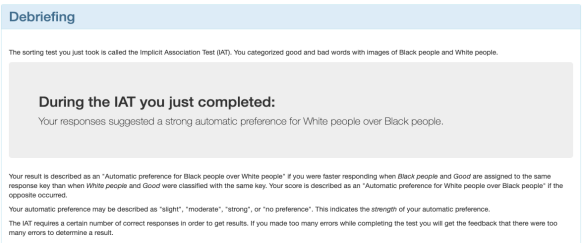
Jamie Cummins & Ian Hussey

Implicit attitude measures are widely used across many fields of psychological science. One core goal of these measures is to provide precise information which can be diagnostic of an individual person's attitude. To date, little progress has been made towards this goal. We argue that this is because psychologists have not yet even quantified individual-level precision in these tasks, much less been able to calibrate their measures towards it. We use bootstrapping to fit confidence intervals to individual-level implicit measure scores using a large dataset (N = ~23,413 individuals) of six different implicit measures assessing three different attitude domains (race, politics, and self-esteem). Our analyses focused on the ability of these measures to distinguish participants from neutral attitudes and each other, while also evaluating the width and coverage of confidence intervals. Despite some variation, all measures exhibited substantial room for improvement. We recommend that researchers in future should use metrics of individual-level precision to calibrate their tasks, both in the context of implicit measures and with tasks in psychological science more broadly.

Implicit measures and individual-level measurement

Implicit measures are widely used in psychological science and beyond as measures of attitudes, beliefs, and stereotypes (Greenwald et al., 2022; Kurdi, Mann, et al., 2019). An often repeated aspiration for these measures is that they may eventually allow us to make inferences about the attitudes/beliefs of individuals, beginning with the first paper detailing the Implicit Association Test (IAT: (Fiedler et al., 2006; Greenwald et al., 1998; Greenwald & Banaji, 1995) right up to present-day reviews of implicit measures (Greenwald & Lai, 2020). Some of the most exciting findings using implicit measures have also related to their prediction of individual behavior in domains such as suicidality (Nock et al., 2010) and voting intentions (Greenwald et al., 2009). These aspirations are also visible in the public face of these measures. Since 1998, the year the first IAT paper was published (Greenwald et al., 1998), the website Project Implicit (implicit.harvard.edu/implicit) has allowed individuals to complete an IAT online and receive individual feedback about their level of bias. For example, feedback at the end of a race IAT tells participants that "Your responses suggested [little to no/a slight/a moderate/a strong] automatic preference for [White people over Black people/Black people over White people]" (see Figure 1).

Figure 1. Screenshot of the feedback provided to a participant on the Project Implicit website in January 2023.



Based on the scientific record describing individual-level prediction as an important goal, and the fact that feedback about individual-level bias is actively given on the flagship website of the most popular implicit measure, it would be reasonable to assume that meaningful inferences about individual participant's implicit biases can be made using current methods. Surprisingly and concerningly, this is explicitly not the case. in their recent authoritative review of meta-analyses, Greenwald and Lai (2020) noted that there have not yet been any high-precision implicit measures developed which are capable of either (i) making diagnostic claims about the traits of individuals, or (ii) making diagnostic claims about precise differences in traits between individuals. In spite of the long-standing aspirations for individual-level precision the field has generally made little progress towards this goal. Although several studies have examined ways to improve the psychometric properties of these tasks such as through alternative scoring algorithms (Greenwald et al., 2003), no substantial increases in reliability have been obtained. Few advances have been attempted for improving individual-level precision. Indeed, it is easier to find examples of attempts to shorten these tasks than to lengthen them (the Brief IAT, Sriram & Greenwald, 2009; shortened Death IAT, Millner et al., 2018). This might make the tasks easier to administer to individuals, but it also makes individuals' scores less useful for individual predictions, given the inverse relationship between task length and reliability estimates (Streiner, 2003).

Despite the common suggestion that high-precision measures should be created, there has been little explanation as to how we might quantify such precision statistically (in other words, how would we know that a measure is precise or not, or quantify exactly how precise it is). Indeed, we argue that this dissociation between

the field's stated aims and its lackluster progress toward these goals is attributable in part to an absence of methods by which to quantify the individual level precision of implicit measures. While considering test-retest reliability might be a starting point, given that some have argued that precision can be improved via improving test-retest reliability (e.g., Greenwald & Lai, 2020), this does not in itself represent a metric of individual-level precision. Without "filling in the blanks" statistically, researchers may condition solely on improving test-retest performance (a metric defined conceptually and statistically at the group-level) without a direct measure of individual-level measurement precision (which is defined conceptually and statistically at the individual-level). Recently, Scheel (2022) has argued that many claims in psychological research are "not even wrong", insofar as they are so underspecified that to be wrong would be an improvement. Similarly, we argue that as currently used, implicit measures are "not even imprecise": collectively, the field does not yet even have tools to estimate the precision of implicit measures, let alone state that this precision is poor.

Unexplored existing solutions

To move this issue forward, it is useful to draw an analogy between our current practices here and our established statistical practices regarding the estimation of group level effect sizes (e.g., via 95% Confidence Intervals). A key limitation with the method of feedback on individual IAT performance used on Project Implicit, as discussed above, is that it fails to account for uncertainty around an individual's score. For example, suppose an individual registers a D score of 0.40 on the IAT. Based on the criteria above, they would be given the feedback that they demonstrated a "moderate" bias for White people over Black people. However, if one were to find that the 95% Confidence Intervals associated with this estimate vary between -0.10 and 0.90, then the interpretation becomes much more ambiguous: this score may represent anywhere between "little-to-no bias" and "a strong bias". Of course, we are not arguing that members of the public should be presented with confidence intervals that they may not be equipped to interpret; rather, our point is that we as researchers do not quantify uncertainty around performance on implicit measures in our research either. Importantly, this interpretation and use of point-estimate scores on implicit measures in the absence of any consideration of estimation precision is also out of step with the common practices of null hypothesis significance testing. For example, the assertion that a given sample demonstrated "moderate bias" would need to be substantiated not merely by the presentation of a mean score, but by an inference method such as a p-value or confidence interval. If we consistently applied our otherwise ubiquitous analytic practices to inferences made in individuals, we would only say that an individual demonstrated a bias on an IAT if we had reason to reject the null hypothesis that they did not. We argue that implicit measures may be more likely to achieve their aspirations as measures with individual level utility if researchers using them have methods of quantifying individual level precision. This is a reasonable first step for later work to then examine how individual level precision can be improved.

The Standard Error of Measurement

With all of this said, it is clear that the field would benefit from an inference method for individuals; one which directly quantifies the measurement error associated with an individual's implicit measure score, and which can fill in the blank which links test-retest reliability and individual scores. Fortunately, such a "precision" metric is well established in the psychological assessment literature, but surprisingly has not yet been employed within the implicit measures literature: the Standard Error of Measurement (SEm; Dudek, 1979). The SEm is defined as:

$$SEm = SD \times \sqrt{(1\text{-}r)}$$

where *SD* refers to the standard deviation, and *r* refers to the test-retest reliability of the measure. After the SEm has been computed, 95% Confidence Intervals around an individual's score can be computed as score ± (1.96 * SEm). The SEm therefore not only represents a metric of individual-level precision, but also clarifies the precise link between this precision and the group-level property of test-retest reliability.

To date, only two studies have used approximation methods similar to SEm to estimate individual-level precision in implicit measures, both of which assessed the IAT in the context of racial bias. Schimmack (2021) used a variant of the measure (substituting test-retest reliability with measure validity) and found that an individual with an IAT D score of .30 would have accompanying confidence intervals ranging from -0.51 to 1.11. Given the bounded nature of the IAT D score (from -2 to 2), this represented rather poor measurement precision. Klein (2020) estimated confidence intervals for participants in terms of their individual-level Cohen's d effect sizes (rather than IAT D score), and found that the median CI width s was 0.76. These results were slightly better than those found by Schimmack, but still a rather large range. Additionally, it is worth noting that the method employed by Klein is arguably inappropriate, given that his estimation method treated Cohen's d confidence intervals as equivalent to those of the IAT D score, despite these representing distinct effect size metrics.

The SEm itself is also not without its drawbacks. As noted above, the test-retest reliability of the measure is needed in order to estimate the SEm; however, implicit measures are not monoliths. The test-retest of implicit measures can vary as a function of the stimuli used within them, the length of the tasks, and a whole host of other features of stimuli and participants (Cummins et al., 2022). Individual participants are also not monoliths. The SEm assumes that the precision of individual scores on a measure will be identical for all individuals; however, it is almost always the case that some individuals' scores will be better-estimated than others (Cummins, 2023; Mollenkopf, 1949; Schmukle,

2023). As such, although the SEm represents a method to estimate individual-level precision, its reliance on the test-retest statistic leads to assumptions about generalizability at both the domain- and individual-levels that are often not met.

Bootstrapped confidence intervals for implicit measures

Fortunately, an alternative method can be used which does not rely on access to test-retest coefficients: namely, by bootstrapping confidence intervals around individuals' scores. Hussey (2020) proposed this method to estimate implicit measure confidence intervals, specifically around scores on the Implicit Relational Assessment Procedure (IRAP) across 18 different domains. Although the method of estimation was slightly different to the previous two studies, results were similarly poor, indicating that the IRAP does not provide precise individual-level estimates.

At this point, two facts should be clear: individual-level precision is an important feature of implicit measures which has not been examined in-depth to date, and the limited research which has been done has been limited by its methods (Klein, 2020; whose estimation method was arguably flawed), scope of measurement procedures (Schimmack, 2021; Klein, 2020; Hussey, 2022; each of which examined only a single implicit measure) and scope of domains examined (Schimmack, 2021; Klein, 2020; each of which examined the IAT's precision only in the context of racial bias). What is needed in order to move this longstanding goal forward is a more comprehensive investigation which examines multiple implicit measures across multiple domains using an estimation method that avoids the pitfalls of the SEm. This was precisely our aim.

We did this using a large open dataset consisting of data from 6 different implicit measures (Implicit Association Test: Greenwald et al., 1998; Brief Implicit Association Test: Sriram & Greenwald, 2009; Single-Target Implicit Association Test: Karpinski & Steinman, 2006; Affect Misattribution Procedure: Payne et al., 2005; Go/No-Go Association Test: Nosek & Banaji, 2001; and Evaluative Priming Task: Fazio et al., 1986) administered across three distinct domains (Bar-Anan & Nosek, 2014) estimating confidence intervals using the method employed by Hussey (2020). One challenge with this approach is the fact that different implicit measures are typically scored with different scoring methods (e.g., the D score for the IAT; categorisation difference scores in the AMP; raw response time differences in evaluative priming). To allow for direct comparison between confidence interval widths between measures, we therefore scored all measures using probabilistic index scores (PI scores; Acion et al., 2006). While these scores will be described in-depth below, it should be noted that PI scores can be more robust to outliers than other commonly used scoring methods (De Schryver & De Neve, 2019). Usefully, PI scores nonetheless correlate highly with D scores ($r = .88$; De Schryver et al., 2018), which many readers are likely more familiar with. PI scores also provide a standardized method of scoring data from tasks that are typically derived from different properties of participants' responses (e.g., accuracy, response times), providing an ideal scoring method to compare multiple measures (see also Cummins et al., 2021).

As described below, we had three primary research questions. Our first research question investigated the relative ability of different measures to detect non-zero effects at the level of individual participants. Specifically, researchers often wish to make claims about the presence or absence of bias in an individual with respect to the difference of their score from a particular point value (typically the neutral-point of zero effect, i.e., IAT D score = 0, or PI = 0.50). We investigated for each measure and across domains the proportion of participants for whom an effect was detectable (i.e., whose scores differed from the neutral point).

The second research question was a more general form of the first. The meaningfulness and interpretability of the neutral point of zero effect on implicit measures (i.e., PI = 0.50) has received much debate in the past (e.g., Blanton & Jaccard, 2006). In order to provide a more general test than the first research question, it can also be useful to quantify the degree to which differences in scores between two individuals can be detected, for example in order to make the inference that person A shows more or less bias than person B. We therefore assessed, for each individual, the proportion of other individuals' scores (within the same measure and domain) that differ significantly. That is, the discriminability of individuals from other individuals. Comparisons between measures are useful here because measures that are more able to discriminate scores between individuals are therefore more useful at the individual level.

The third research question was a related form of individual-level utility. A measure will demonstrate an observed range of scores in a given sample. If a measure has utility at the individual level, it will be capable of assigning individuals to a narrow part of that observed range. For example, imagine a depression scale with an observed range of 1 to 10 in a large sample, and an individual in that sample with a score of 3. A useful scale might be able to estimate an individual's scores within $\pm$ 2, so that the individual with a score of 3 can be more usefully said to be between 1 and 5. We could therefore infer that the individual's true score is "not high" (i.e., their interval excludes 6 to 10). In contrast, a less useful scale might have individual intervals of $\pm$ 5. The same individual could now only be said to have a score of 1 to 8, from which we can only infer that their true score is "not extremely high" (i.e., their interval only excludes 9 and 10). We can describe this property of a measure as individual coverage; that is, the typical proportion of the observed range of confidence intervals covered by a given individual's confidence interval. Comparisons between measures are useful here because measures with lower

individual coverage are more able to make inferences about where on the continuum individuals lie.

Method

Data source

This study uses openly available data collected on Project Implicit (https://implicit.harvard.edu), originally collected by Bar-Anan and Nosek (2014; data available from osf.io/qf9jx). The data, code, and preregistration for our analyses can be found on the Open Science Framework (osf.io/pq6nf).

Sample

The sample used for these analyses was taken from Bar-Anan & Nosek's (2014) data, collected via the Project Implicit website. A total of 23,413 unique individuals participated in this study (63% women, 36% men, 1% unknown; mean age = 29.1, SD = 12.0). Of this figure, 8.7% completed only one measure, 4.9% completed 2 measures, 7.7% completed three measures, and 31% completed four measures. 45.1% completed more than four measures, of which 10% completed more than ten measures. Detailed information regarding the collection of these data can be found in Bar-Anan and Nosek (2014). The data used in our analytic sample,composed of participants who completed at least one measure in the overall study and met common accuracy and latency performance exclusion criteria (full details in supplementary materials), leading to 21060 observations in total (i.e., some participants may have completed more than one of the measures). Within this, 6902 participants completed the Implicit Association Test (IAT), 7238 completed the Affect Misattribution Procedure (AMP), 6039 completed the Brief IAT (BIAT), 6795 completed the Evaluative Priming Task (EPT), 6529 completed the Go-No Go Association Test (GNAT), and 6626 completed the Single-Target IAT (ST-IAT). These completions were divided approximately evenly across the three domains of race, politics, and self-esteem, to which they were assigned randomly within the original study.

It is important to note that in their original study, Bar-Anan and Nosek (2014) also included a seventh implicit measure; namely, the sorting paired-features task (SPF, Bar-Anan et al., 2009). However, this task was not included in the current study on the basis that (a) it has seen much less use than the other tasks, and more importantly (b) effects on the task are typically quantified using more than one score for each individual. In contrast, the other 6 tasks are quantified using a single score. In order to compare like with like, only the other 6 tasks were included.

Measures

Herein we briefly describe each of the measurement procedures used within the study. For more detailed descriptions, see Bar-Anan and Nosek (2014) and the associated references provided under each measure.

Implicit Association Test (IAT)

The IAT used in this study followed the procedure outlined in Nosek et al. (2007). A single attitude-object-only practice block of 20 trials was followed by a second practice block of 20 trials involving only evaluative stimuli. The third (20 trials) and fourth (40 trials) blocks involved a combination of the required responses on the two previous blocks. Block 5 was identical to block 1 but with the required response directions switched, and the sixth (20 trials) and seventh (40 trials) blocks incorporated this new configuration in blocks otherwise identical to the third and fourth blocks. The order of required response configurations was randomised between participants.

Brief Implicit Association Test (BIAT)

The BIAT was developed to be a version of the IAT with a shorter administration time and slightly easier instructions for the participant. It requires only two (rather than four) responses on each critical block (Sriram & Greenwald, 2009).

Single-Target Implicit Association Test (ST-IAT)

The ST-IAT was identical to the IAT but with only one attitude-object (rather than two) investigated on each critical block (Karpinski & Steinman, 2006).

Affect Misattribution Procedure (AMP)

The AMP followed the procedure described by Payne et al. (2005).

Go-No Go Association Task (GNAT)

The GNAT here followed the procedure described by Nosek and Banaji (2001), with scores computed based on response latencies.

Evaluative Priming Task (EPT)

The EPT followed the procedure outlined by Fazio et al. (1995).

Procedure

For all participants, each session lasted approximately 15 minutes. Within each session, participants were presented with two "long-duration" and two "short-duration" measures (the implicit measures were divided across these two categories; see Bar-Anan and Nosek, 2014). There were no constraints on participants in terms of the measures they would receive beyond the fact that the same exact measure/domain combination could not be presented twice in one session.

Research Questions

As described earlier, we addressed three primary research questions in this study.

RQ1

For each measure, meta-analyzed across domains, what proportion of individual participants' scores were detectably different from the neutral point of zero effect (i.e., PI = 0.50)? How do these proportions differ between measures?

RQ2

For each measure, meta-analyzed across domains, what proportion of other participants' scores were individual participants' scores detectably different from? In contrast to RQ2, we compared each participant's score against all other participants' scores within the same measure and domain. How do these proportions differ between measures?

RQ3

For each measure, meta-analyzed across domains, what proportion of the observed range of scores did individuals' 95% Confidence Interval typically cover? How do these proportions differ between measures?

Results

Data processing

Scoring algorithm

The implicit measures we compared typically use different methods and metrics for scoring. The IAT, ST-IAT, and B-IAT tend to use a D score based on response times (Greenwald et al., 2003); the AMP tends to use proportion of prime-consistent evaluative responses (Payne et al., 2005); the GNAT and EPT tend to be scored based on differential response latencies (although the GNAT can also be scored based on accuracy differentials; Fazio et al., 1995; Gomez et al., 2007; Nosek & Banaji, 2001). These different methods of scoring, and the corresponding differences in scales, score ranges, and error variances associated with them, would limit direct comparisons between the measures. As such, we opted to instead score every measure using the same analytic method: namely, using probabilistic index (PI) scores (De Schryver et al., 2018). This metric has been referred to by many names, including Ruscio's A (2008) and the common language effect size (McGraw & Wong, 1992). We refer to it here as the PI on the basis that this is the term used in papers related to the current one and when scoring data from implicit measures (e.g., Hussey, 2020; De Schryver et al., 2018). PI scores are highly interpretable: they are the probability of a randomly selected response in one block type being superior (e.g., a longer reaction time, or more positive evaluation) to a randomly selected response in the other block type. As a probability value, PIs can range from 0 to 1, with the neutral point of zero effect being 0.50 (i.e., equal probability). In this manner, using a single robust and interpretable scoring method allowed for direct comparisons between the measures.

Confidence intervals around individuals' scores

Confidence intervals around individuals' scores were calculated by bootstrapping confidence intervals using the basic method and 2000 resamples. This was implemented in R using the boot package (Canty & Ripley, 2021).

Analyses

Descriptive statistics

PI Scores

We first aimed to gauge the modal CI width for each measure across each domain using maximum a posteriori estimation (i.e., computing the mode of the posterior distribution of CI width values). These results are presented in Table 1.

Table 1. Maximum a posteriori values for each measure across each domain.

Measure	Domain		
	Politics	Race	Self
IAT	0.21	0.21	0.21
B-IAT	0.20	0.20	0.20
ST-IAT	0.17	0.17	0.16
AMP	0.28	0.28	0.28
GNAT	0.19	0.19	0.19
EPT	0.17	0.17	0.17

IAT D Scores

Although we focus on PI scores in the measures here to make comparisons on the same scale across measures, the inspiration for this work came in part from the criteria associated with the IAT D score on Project Implicit, as described in the Introduction. Therefore, as an additional descriptive analysis, we also estimated confidence intervals around the D score of the IAT in the context of implicit racial attitudes at each of the cut-offs given by Project Implicit (0, 0.15, 0.35, and 0.65 respectively for no bias, weak bias, moderate bias, and strong bias). We also provide updated interpretations of these cut-offs in line with the values covered by the associated confidence intervals. These results are presented in Table 2.

Table 2. Project Implicit cut-off values for each of the three IATs in the context of racial attitudes, their corresponding confidence intervals, and the updated interpretations based on these confidence intervals.

Project Implicit cut-off	Interpretation of cut-off	Associated Confidence Intervals	Appropriate new interpretation
0	No bias	-0.38, 0.38	Moderately negative to moderately positive bias
0.15	Weak bias	-0.21, 0.51	Weak negative to moderate positive bias
0.35	Moderate bias	0.02, 0.68	No bias to strong positive bias
0.65	Strong bias	0.36, 0.94	Moderate positive bias to strong positive bias

RQ1. Proportion of effects detectable from zero effect

Calculation of scores

95% CIs on individuals' scores were used to assess whether each individual excluded the neutral point of zero effect on the task (i.e., PI = 0.50). Intervals that

excluded the neutral point (PI = 0.50) were scored as a detectable effect. A caterpillar plot of individual participants' scores and their CIs, split by measure and domain, can be found in Figure 1.

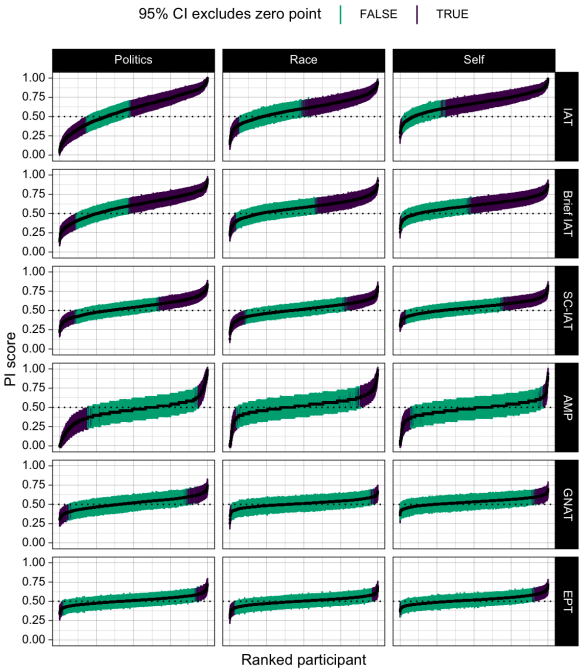


Figure 1. Caterpillar plot of the distribution of PI scores, and their associated confidence intervals, for each participant across each measure and domain.

Meta-analytic model

To compare the proportion of detectable effects between measures, the data from individuals was meta-analyzed. For each measure and domain, we calculated the proportion of detectable effects and its variance. We then entered the proportions into a linear mixed-effects model using the R package lme4 (Bates et al., 2015). The Wilkinson notation for the model was as follows:

$$\text{proportion\_diff\_zero} \sim 1 + \text{measure} + (1 \mid \text{domain}),$$
$$\text{weights} = 1/\text{variance}$$

That is, we entered measure as a fixed effect to estimate the proportions for each measure and make inferences about differences between them (i.e., measures are an exhaustive set for our purposes). Domain was entered as a random intercept to acknowledge the non-independence of attitudes within each domain, and the fact that there are other domains to be generalized to in principle (i.e., domain is non-exhaustive, and attitude domain is the data generating signal). We weighted by inverse variance, as is common in meta-analytic models (Viechtbauer, 2005). A forest plot of the individual effect sizes for each domain and the meta-analyzed effect size for each measure can be found in Figure 2A. Tables containing full results from this and all subsequent models, along with the data presented in the figures in table format, can be found in the online supplementary materials.

Results of the meta-analysis were interpreted with the aid of pairwise comparisons between the measures. These were calculated using the emmeans R package (Lenth, 2022) while also controlling error rates using Holm correction. Results from these pairwise comparisons are presented in Table 3.

Table 3. Pairwise comparisons of the estimated marginal means of the proportions of participants discriminable from 0.50 for each measure.

Measure 1	Measure 2	Estimated marginal mean difference	95% CIs	p-value
IAT	B-IAT	0.12	0.03, 0.22	$< .001$
IAT	ST-IAT	0.30	0.21, 0.39	$< .001$
IAT	AMP	0.49	0.41, 0.57	$< .001$
IAT	GNAT	0.57	0.49, 0.64	$< .001$
IAT	EPT	0.58	0.50, 0.65	$< .001$
B-IAT	ST-IAT	0.18	0.08, 0.28	$< .001$
B-IAT	AMP	0.37	0.28, 0.45	$< .001$
B-IAT	GNAT	0.44	0.36, 0.53	$< .001$
B-IAT	EPT	0.45	0.37, 0.53	$< .001$
ST-IAT	AMP	0.19	0.10, 0.27	$< .001$
ST-IAT	GNAT	0.26	0.18, 0.34	$< .001$
ST-IAT	EPT	0.27	0.20, 0.35	$< .001$
AMP	GNAT	0.08	0.01, 0.14	.022
AMP	EPT	0.09	0.02, 0.15	.007
GNAT	EPT	0.01	-0.05, 0.07	.713

RQ2. Proportion of scores discriminable from other scores

Calculation of scores

95% CIs on individuals' scores were also used to assess the proportion of other participants' scores from which each individual's score was detectibly different. Pairwise comparisons between each participant and every other participant (separately for each measure and domain) were calculated using the 95% Confidence Interval on the difference scores between them via bootstrapping, to create one proportion for each participant and its variance. For this and all subsequent analyses, if proportions of 0 or 1 or variances of 0 were obtained, these values were offset by 0.001 in order to allow for meta-analysis.

Meta-analytic model

The individual level proportions were entered into a similar linear mixed-effects model to the previous one:

$$\text{proportion\_discriminable} \sim 1 + \text{measure} + (1 \mid \text{domain}),$$
$$\text{weights} = 1/\text{variance}$$

A forest plot of the individual effect sizes for each domain and the meta-analyzed effect size for each measure can be found in Figure 2B. Like the previous analysis, results from the forest plot were interpreted with the aid of pairwise comparisons between the measures, again using Holm correction. These pairwise comparisons are presented in Table 4.

Table 4. Pairwise comparisons of the estimated marginal means of participants who could be discriminated from one another for each measure.

Measure 1	Measure 2	Estimated marginal mean difference	95% CIs	p-value
IAT	B-IAT	0.14	0.06, 0.22	< .001
IAT	ST-IAT	0.19	0.11, 0.26	< .001
IAT	AMP	0.30	0.23, 0.38	< .001
IAT	GNAT	0.43	0.38, 0.49	< .001
IAT	EPT	0.42	0.36, 0.48	< .001
B-IAT	ST-IAT	0.05	-0.04, 0.13	.261
B-IAT	AMP	0.17	0.08, 0.25	< .001
B-IAT	GNAT	0.29	0.23, 0.36	< .001
B-IAT	EPT	0.28	0.21, 0.35	< .001
ST-IAT	AMP	0.12	0.18, 0.31	.005
ST-IAT	GNAT	0.25	0.18, 0.31	< .001
ST-IAT	EPT	0.24	0.17, 0.30	< .001
AMP	GNAT	0.13	0.06, 0.20	< .001
AMP	EPT	0.12	0.05, 0.19	< .001
GNAT	EPT	-0.01	-0.06, 0.04	.646

Table 5. Pairwise comparisons of the estimated marginal means of the coverage of participants' CIs for each measure.

Measure 1	Measure 2	Estimated marginal mean difference	95% CIs	p-value
IAT	B-IAT	-0.03	-0.05, 0.00	.019
IAT	ST-IAT	-0.02	-0.04, 0.00	.073
IAT	AMP	-0.03	-0.06, 0.00	.06
IAT	GNAT	-0.13	-0.15, -0.10	< .001
IAT	EPT	-0.11	-0.13, -0.09	< .001
B-IAT	ST-IAT	0.01	-0.01, 0.02	.274
B-IAT	AMP	0.00	-0.03, 0.02	.807
B-IAT	GNAT	-0.10	-0.12, -0.08	< .001
B-IAT	EPT	-0.08	-0.10, -0.07	< .001
ST-IAT	AMP	-0.01	-0.04, 0.01	.381
ST-IAT	GNAT	-0.11	-0.13, -0.09	< .001
ST-IAT	EPT	-0.09	-0.10, -0.08	< .001
AMP	GNAT	-0.10	-0.13, -0.07	< .001
AMP	EPT	-0.08	-0.11, -0.05	< .001
GNAT	EPT	0.02	0.00, 0.03	.058

RQ3. Coverage of Individuals' Confidence Intervals

Calculation of scores

95% CIs on individuals' scores were also used to assess the typical proportion of the observed range covered by an individual interval. First, the observed range of intervals was calculated for each domain and measure. Then, each interval was divided by this observed range to calculate a proportion. To meta-analyze these proportions, their mean and variance were then calculated.

Meta-analytic model

The proportions were entered into a similar linear mixed-effects model to the previous two:

$$\text{ci\_width\_proportion\_mean} \sim 1 + \text{measure} + (1 \mid \text{domain}),$$
$$\text{weights} = 1/\text{variance}$$

A forest plot of the individual effect sizes for each domain and the meta-analyzed effect size for each measure can be found in Figure 2C. Tables containing the numerical result can be found in the supplementary materials. Results were again interpreted with the aid of pairwise comparisons between the measures using Holm corrections, which can be found in Table 5.

Summary of results

Across all three analyses, the IAT was descriptively the best performer. It demonstrated the best discriminability of scores from both zero (approximately 70%), and from other scores (approximately 50%). After this, the IAT derivatives (namely, the BIAT and ST-IAT) performed relatively well. Notably, the BIAT (0.20 modal CI width; 55% discriminability from zero; 38% discriminability from others) was superior to the ST-IAT in our meta-analytic models (0.17 modal CI width; 38% discriminability from zero; 32% discriminability from others). The AMP followed next, with a modal CI width of 0.28, around 20% discriminability from zero, and around 23% discriminability from others.

The worst performers were the EPT (0.17 modal CI width; 10% discriminability from zero; 10% discriminability from others) and the GNAT (0.19 modal CI width; 10% discriminability from zero; 6% discriminability from others). In terms of their CI coverage, the IAT, BIAT, ST-IAT, and AMP all performed relatively similarly (and did not significantly differ from one another in the pairwise comparisons, except for the IAT and B-IAT). The EPT and GNAT had significantly poorer coverage than these 4 measures, and they did not significantly differ from one another in this regard. Descriptively, we noted that precision across measures was generally higher in the politics domain relative to race and self-esteem (particularly for RQ1 and RQ2).

Discussion

A central aim of the implicit measures field has been to use these measures to make predictions or inferences about the implicit biases of individual participants. Researchers using implicit measures have been acutely aware that these measures are currently insufficient to do so (Greenwald & Lai, 2020). Previous research using these measures has not focused on estimating the precision of these individual-level scores. Consequently, we have little sense of exactly how precise these measures are, and no sense of how one measure compares to another in this regard. In this study, we attempted to unpack this by estimating and comparing the precision of six different implicit measures across three different domains. A relatively clear picture emerged: while none of the implicit measures are currently suitable for individual level inferences, some measures are superior to others. The IAT demonstrated the best performance across the metrics we examined, followed by the IAT variants (BIAT and ST-IAT), and then the AMP. The GNAT and EPT demonstrated poor performance in both relative and absolute terms.

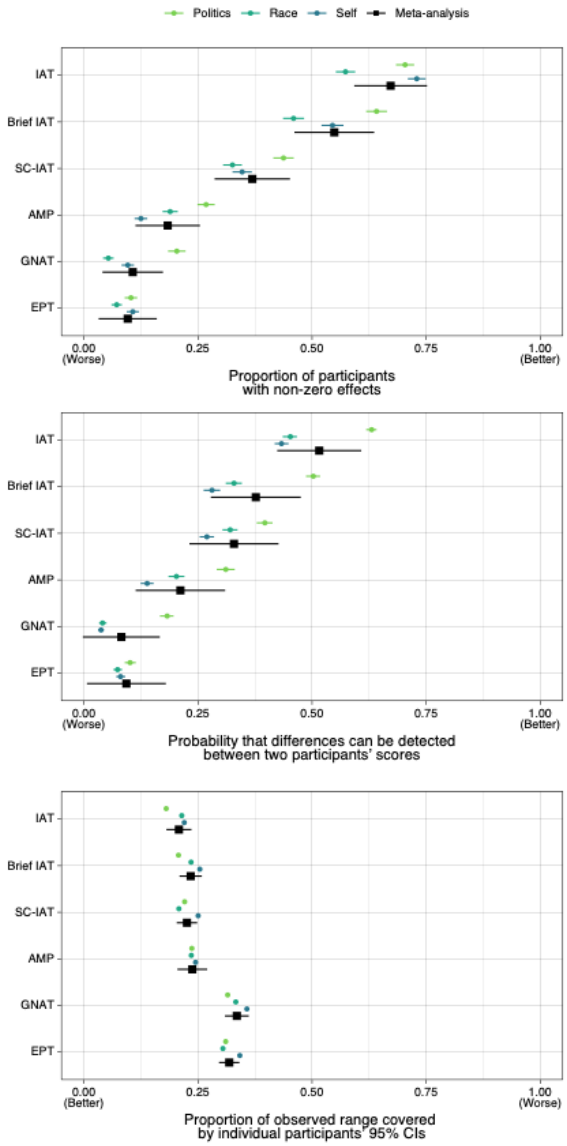


Figure 2. Forest plot for the meta-analytic models associated with the three research questions. The upper third of the plot shows the meta-analytic model for the proportion of participants whose scores differed detectably from zero; the middle third of the plot shows the meta-analytic model for the probability of detectable difference between two participants; and the lower third shows the meta-analytic model for the coverage of the confidence intervals.

Implicit measures should be calibrated for individual-level precision

While these comparative assessments of the individual utility of six common implicit measures are useful in and of themselves, we feel that the most important aspect of our work here is that it provides researchers with a framework with which to assess the precision of their implicit measures. This has until now been sorely lacking in the implicit social cognition literature. Previously, researchers have alluded to individual utility based on test-retest reliability (TRR; Greenwald & Lai, 2020). However, TRR does not directly quantify the (im)precision of scores in measures; the method we employed here does. TRR also typically requires additional data collection (i.e., to obtain a TRR estimate) or requires the researcher to rely on estimates from published work (which may not be generalisable; cf. Cummins et al., 2022). Bootstrapping does not suffer from this limitation, instead allowing researchers to obtain estimates of individual precision from the data they have at hand. This opens many avenues for future research. Researchers can also have a better sense of how exactly to interpret individual scores on implicit measures and can update their beliefs accordingly. For example, in the context of D scores in the IAT as highlighted in Table 2, scores of 0 in the IAT can indicate anywhere between moderate negative and moderate positive bias, rather than no bias (as is stated on the Project Implicit website), and it is only at a score of around 0.35 that one can reliably conclude that that individual has a non-zero bias (and even then, this bias may barely differ from zero). These results strongly suggest that the interpretation of scores provided to participants on Project Implicit should be accordingly updated with these more accurate interpretations.

Researchers can also now more readily and directly examine the way in which procedural or experimental manipulations impact the precision of their implicit measures. This, in turn, will allow researchers to pursue an agenda which can be oriented on improving the precision of their measures, as advocated by Greenwald and Lai (2020) and others. What could such a research agenda look like? Fortunately, there are several known methods for improving the precision of measurement procedures. First, increasing the number of trials used in the task can improve precision of measurement (although issues such as fatigue-based effects would need to be considered; Langner et al., 2010). Interestingly, this is in stark contrast to the typical recommendations that researchers using these measures for diagnosis have made, wherein the measures are typically encouraged to be shortened in order to increase their practicality (Millner et al., 2018). Whereas such previous work has focused on the ease with which implicit measures can be completed (e.g., lowering task duration and participant burden), this is likely at the expense of the precision of the estimates obtained.

Second, the nature of the stimuli used within the task, and the method by which they are selected, can also be examined. Recent research suggests that a variety of features of the stimuli used in the test can impact the psychometric properties of implicit measures at the group-level, such as the oppositionality between concept categories, their salience to participants, and their relation to the to-be-predicted behavior (Cummins et al., 2022; Greenwald et al., 2022; Irving & Smith, 2020). These stimulus features are promising candidates for the improvement of individual-level precision. Other features of these tasks which are known to affect group-level reliability, such as the method scoring of the task,

will also likely impact individual-level precision. Although previous research has demonstrated that various methods of scoring IAT data tend to produce very highly correlated scores, making differences in scoring methods relatively unimportant for group level analyses, it is important to note that research elsewhere emphasizes that even very highly correlated scores (e.g., $r = .99$) can produce very different individual level predictions (McNeish, 2022). Indeed, recent computational approaches to the scoring of implicit measures, such as the GSR model (Kvam et al., 2023) produce superior estimates of test-retest reliability; by implication, individual precision will also likely be superior compared to modal scoring practices. However, no one of these factors is likely to be a silver bullet for the issue of individual-level precision; the solution will be found in the careful consideration of the combination of these and other factors (cf. Kurdi, Seitchik, et al., 2019).

Our work here highlights the need for improved measurement practices within the field of implicit measures; a need which is currently echoing throughout psychological science at large (Flake & Fried, 2020; Hussey & Hughes, 2020). Whereas the goal of individual-level prediction has been present in the field of implicit measures for 25 years, there has been little attention paid directly to *how precisely* we can quantify this precision. Indeed, it is somewhat alarming that we have collectively vocalized the importance of precise individual-level measurement with implicit measures, yet never developed any criterion to quantify precision. We have not simply missed the target, we have failed to identify it as a target altogether: our measures are not even imprecise, their precision has simply never been quantified (cf. Scheel, 2022). The link between theory and measurement becomes clear when considering this: in the absence of good measurement of phenomena of interest, our theoretical programmes regarding these phenomena will necessarily fail.

Individual-level precision beyond implicit measures

Beyond the field of implicit measures, many other research areas across psychological science aim to make claims about individuals in the absence of estimating individual-level effects. In one recent paper McManus et al. (2023) noted that the majority of psychological researchers wish to make claims about at least a majority of individuals when conducting psychological experiments. Others still have proposed that the presence of absence of effects within individual participants represents a more meaningful effect size metric than many other group-level approaches (Grice et al., 2020). The method of individual-level estimation that we have used here can likely be applied to many of these areas. Indeed, this method is arguably superior in some ways to arithmetic methods such as the SEm, given that the bootstrapping procedure does require knowing any task parameters (e.g., test-retest reliability) in advance. This method can thus be used with essentially any performance-based task which consists of response times and/or accuracy scores, even if it is the first time this task has been employed. This can allow researchers to focus very early on the precision of their measure during its development, as well as probing the precision of extant measures. If the goal of a measurement procedure is to make individual-level inferences, then those developing that measure should strongly consider quantifying individual-level precision explicitly as early as possible in the development process of their measure.

Limitations

A critical limit of the generalisability of our findings here relates to our selection of measures. The implicit measures subjected to testing here may not be representative of all implicit measures, particularly more recently developed "relational implicit measures" (Cummins & De Houwer, 2022). Even within the measures we tested, our results may not generalize to the measures at large; we investigated the properties of the measures across three domains, but there are countless others which may come with idiosyncratic differences resulting in idiosyncratic differences in the properties of the tasks in those contexts. Further still, although confidence intervals can be fitted to scores in many psychological measures, it remains to be seen how these various other measures may perform. We advocate strongly for tests of the generalizability of these results across other implicit measures, domains, and psychological tasks at large in future research.

One further constraint on generalizability, and a more general issue with our work, relates to the use of bootstrapping to estimate confidence intervals. Although a powerful approach to estimation, bootstrapping is not without its limitations. Most critically, for procedures with a limited number of trials, many bootstrapping approaches may produce biased estimates (Mostofian & Zuckerman, 2019). Although other bootstrapping methods exist which can correct for bias due to small samples (e.g., bias-corrected and accelerated bootstrapping; Puth et al., 2015), these methods can suffer from convergence issues, or may produce scores in some bootstrap samples which fall outside of the possible bounds of the scale (e.g., outside of 0 and 1 in the PI). Although our findings are relatively robust across different bootstrapping methods, it is critical to carefully consider the method of choice when using this approach.

Conclusion

This work represents the first comparison of multiple implicit measures, across multiple domains, in terms of their individual-level measurement precision. Although we hope that our results will be informative and interesting to those researchers who have used implicit measures across various studies, our ultimate hope is that researchers will *use* metrics of individual level precision as benchmarks to improve these measures in the future. After all, theoretical advances regarding individual-level processes will only ever be made if we can precisely and validly measure those processes. Psychological science cannot be a science of persons without the precise measurement of persons.

Author note

JC, Department of Experimental Clinical and Health Psychology, Ghent University and Institute of Psychology, University of Bern, & IH, Institute of Psychology, University of Bern. JC was supported by FWO grant 1202624N. Correspondence concerning this article should be sent to jamie.cummins@ugent.be or ian.hussey@unibe.ch.

References

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. Statistics in Medicine, 25(4), 591–602. https://doi.org/10.1002/sim.2256

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: A measure of association strengths. Experimental Psychology, 56(5), 329–343. https://doi.org/10.1027/1618-3169.56.5.329

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. American Psychologist, 61(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Canty, A., & Ripley, B. (2021). boot: Bootstrap R (S-Plus) Functions (1.3-28) [R].

Cummins, J. (2023). On the measurement of relational responding. Journal of Contextual Behavioral Science. https://doi.org/10.1016/j.jcbs.2023.10.003

Cummins, J., Hussey, I., & Spruyt, A. (2022). The role of attitude features in the reliability of IAT scores. Journal of Experimental Social Psychology, 101, 104330. https://doi.org/10.1016/j.jesp.2022.104330

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 : Journal of the Society of Psychologists in Addictive Behaviors, 35(4), 458–471. https://doi.org/10.1037/adb0000643

De Schryver, M., & De Neve, J. (2019). A tutorial on probabilistic index models: Regression models for the effect size P(Y1. Psychological Methods, 24(4). https://doi.org/10.1037/met0000194

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, D. (2018). The PIIRAP: An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure. Journal of Contextual Behavioral Science, 7, 97–103. https://doi.org/10.1016/j.jcbs.2018.01.001

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. Psychological Bulletin, 86(2), 335–337. https://doi.org/10.1037/0033-2909.86.2.335

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), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). European Review of Social Psychology, 17(1), 74–147. https://doi.org/10.1080/10463280600681248

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Advances in Methods and Practices in Psychological Science, 3(4), 456–465. https://doi.org/10.1177/2515245920952393

Gomez, P., Ratcliff, R., & Perea, M. (2007). A Model of the Go/No-Go Task. Journal of Experimental Psychology. General, 136(3), 389–413. https://doi.org/10.1037/0096-3445.136.3.389

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., ... Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. Behavior Research Methods, 54(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

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

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential

Election. Analyses of Social Issues and Public Policy, 9(1), 241–253. https://doi.org/10.1111/j.1530-2415.2009.01195.x

Hussey, I. (2020). The Implicit Relational Assessment Procedure is not suitable for individual use. PsyArXiv. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. Advances in Methods and Practices in Psychological Science, 3(2), 166–184. https://doi.org/10.1177/2515245919882903

Irving, L. H., & Smith, C. T. (2020). Measure what you are trying to predict: Applying the correspondence principle to the Implicit Association Test. Journal of Experimental Social Psychology, 86, 103898. https://doi.org/10.1016/j.jesp.2019.103898

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

Klein, C. (2020). Confidence Intervals on Implicit Association Test Scores Are Really Rather Large. PsyArXiv. https://doi.org/10.31234/osf.io/5djkh

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. Proceedings of the National Academy of Sciences, 116(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

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

Kvam, P. D., Irving, L. H., Sokratous, K., & Smith, C. T. (2023). Improving the reliability and validity of the IAT with a dynamic model driven by similarity. Behavior Research Methods. https://doi.org/10.3758/s13428-023-02141-1

Langner, R., Steinborn, M. B., Chatterjee, A., Sturm, W., & Willmes, K. (2010). Mental fatigue and temporal preparation in simple reaction-time performance. Acta Psychologica, 133(1), 64–72. https://doi.org/10.1016/j.actpsy.2009.10.001

Lenth, R. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means (1.8.2) [R].

McGraw, K. O., & Wong, S. (1992). A common language effect size statistic. Psychological Bulletin, 111(2), 361.

McManus, R., Young, L., & Sweetman, J. (2023). Psychology is a Property of Persons, Not Averages or Distributions: Confronting the Group-to-Person Generalizability Problem in Experimental Psychology. Advances in Methods and Practices in Psychological Science.

McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. Behavior Research Methods. https://doi.org/10.3758/s13428-022-02016-x

Millner, A. J., Coppersmith, D. D. L., Teachman, B. A., & Nock, M. K. (2018). The Brief Death Implicit Association Test: Scoring recommendations, reliability, validity, and comparisons with the Death Implicit Association Test. Psychological Assessment, 30(10), 1356–1366. https://doi.org/10.1037/pas0000580

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. Psychometrika, 14(3), 189–229. https://doi.org/10.1007/BF02289153

Mostofian, B., & Zuckerman, D. M. (2019). Statistical uncertainty analysis for small-sample, high log-variance data: Cautions for bootstrapping and Bayesian bootstrapping. Journal of Chemical Theory and Computation, 15(6), 3499–3509. https://doi.org/10.1021/acs.jctc.9b00015

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., & Banaji, M. R. (2001). The Go/No-go Association Task. Social Cognition, 19(6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886

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, Automatic processes in social thinking and behavior (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

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

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

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. Infant and Child Development, 31(1), e2295. https://doi.org/10.1002/icd.2295

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. Perspectives on Psychological Science, 16(2), 396–414. https://doi.org/10.1177/1745691619863798

Schmukle, S. C. (2023). Unbiased Confidence Intervals for Individual Scores in Psychological Testing: The Rescaled Estimated True Score (RETS) Approach. unpublished manuscript.

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

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. Journal of Personality Assessment, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. Journal of Educational and Behavioral Statistics, 30(3), 261–293.