Comparing the suitability of 6 implicit measures for individual use

*Pre-registration*

# 

# 

# Preamble

Given that some of the text in the preregistration will be used in the final manuscript, we opt to write the description below in the past tense for the sake of efficiency. Readers should also note that we may allude to specific figures in the text (e.g., “Figure XX”). Note that this text represents placeholder text which will be completed in the final study manuscript.

# Introduction

Implicit measures are widely used in psychological science and beyond as measures of attitudes, beliefs, and stereotypes. Many have used scores on these measures to make claims about the attitudes/beliefs/stereotypes of *individuals*, whether in the scientific literature (e.g., Nock et al. 2010) or in a public-facing context, such as feedback on your personal level of bias after completing an IAT on the Project Implicit website (<https://implicit.harvard.edu/implicit/>; i.e., effect size |*D*| < .15 = "little to no bias" > .15 = “a slight bias”, > .35 = “a moderate bias”, ≥ .65 = “a strong bias”). However, these point estimates of individuals’ scores necessarily have uncertainty around them. In other words, claims about individuals’ biases have generally failed to account for the uncertainty in these scores. As such, the practice of making inferences about individuals’ biases is out of step with the common practices of estimation and null hypothesis significance testing. For example, the claim that the IAT effect in a sample is >0 would need to be substantiated by an inference method, such as a p-value or confidence interval. In contrast, in both research and science communication contexts, individuals are typically given feedback about their level of bias based solely on their numerical *D* score, agnostic to any uncertainty around its estimation. If we consistently applied our otherwise ubiquitous analytic practices to inferences made about 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.

Surprisingly, this suggestion has not been made within the implicit measures literature until very recently (e.g., Hussey, 2021; Klein, 2021). While some work has estimated confidence intervals around IAT, no work has compared multiple implicit measures to assess their relative suitability for individual use. This study did this using a large open dataset of multiple implicit measures compared across multiple domains (Bar-Anan & Nosek, 2014).

As described below, we had three primary research questions. The first research question related to the relative ability of different measures to detect effects for 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., PI = 0.50). We investigated, for each measure and across domains, the proportion of participants who demonstrated a detectable effect (i.e., whose scores defected from the neutral point). On some measures this represented a reaction time bias, on others an evaluative bias. Comparisons between measures are useful here because measures that are more likely to show an effect at the individual level are therefore more useful at the individual level.

The second research question was a more general form of the first. Rather than comparing each individuals’ score against the neutral-point of zero effect (i.e., PI = 0.50), it can also be useful to know whether two individuals differ from one another, 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 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 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 ±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 ±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, i.e., the proportion of the observed range of confidence intervals covered by a given individual’s scores. 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, therefore making them more precise at the individual level.

# Method

## Sample

The sample used for these analyses was taken from Bar-Anan & Nosek’s (2014) data, collected via the Project Implicit website. Detailed information regarding the collection of these data can be found in Bar-Anan and Nosek (2014). The data consisted of 21060 participants in total who completed and met the screening criteria for at least one measure in the overall study. 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.

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). However, this task was not included in this 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.

## Prior knowledge of the dataset & method of developing the code

Code was adapted from an existing article (Hussey, 2020, <https://psyarxiv.com/w2ygr/>). In order to develop analysis code to implement our analyses seamlessly while minimizing our prior knowledge of the dataset, we reserved 10% of the total sample for exclusive use in the code development process (i.e., the “training” dataset). This allowed us to iteratively develop our code through trial-and-error in order to ensure the final version runs seamlessly, while also avoiding running any analyses on the final dataset (i.e., the 90% “testing” dataset). Results reported below are based on analyses for the 90% testing dataset only. Other than for the purposes of data processing (i.e., running the processing.Rmd script to produce the full processed dataset before splitting it 10%/90%), we did not look at or run any analyses on the testing dataset prior to preregistration (i.e., the analyses.Rmd has been developed and run only on the training dataset). We therefore have no prior knowledge of the testing dataset, following the suggestions of Baldwin, Pingault, Schoeler, Sallis & Munafò (2022; “Protecting against researcher bias in secondary data analysis: challenges and potential solutions”, <https://link.springer.com/article/10.1007/s10654-021-00839-0>).

## Research Questions

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? I.e., instead of using the neutral point of zero effect (PI = 0.50), 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 cover? How do these proportions differ between measures?

# Analytic Plan

Below, we briefly describe our data processing and analysis steps. However, we also invite the reader to inspect our fully preregistered code for precise specifications on all aspects of the analyses. The implementations of these in R represent the precise and formal preregistration rather than our loose description here.

## 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, Nosek & Banaji, 2003); the AMP tends to use proportion of prime-consistent evaluative responses; the GNAT and EPT tend to be scored based on differential response latencies (although the GNAT can also be scored based on accuracy differentials). These different methods of scoring, and the corresponding differences in scales, score ranges, and error variances associated with them, make direct comparisons between the measures somewhat difficult. As such, we opted to instead score every measure using the same analytic method: namely, using probabilistic index (PI) scores. This metric has been referred to by many names, including Ruscio’s A (Ruscio, 2008), the probability of superiority, and the common language effect size (McGraw and Wong, 1992). PI scores 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 0.5 (i.e., equal probability). We refer to it as the PI on the basis that this is the term used in papers related to the current one (e.g., Hussey et al., 2020). PI scores have previously been used as a method of scoring several implicit measures (De Scryver et al., 2018; Cummins et al., 2021) and tend to exhibit superior psychometric properties to other methods of scoring, particularly in relation to their robustness to outliers, while remaining to be highly correlated with the IAT *D* score (De Scryver et al., 2018). In this manner, using a single robust scoring method allowed for direct comparisons between the measures.

### Confidence intervals around individuals’ scores

Confidence intervals around individuals’ scores were calculated using the strategy and implementation originally deployed by Hussey (2021; namely, bootstrapping confidence intervals using 2000 resamples).

## Analyses

### RQ1. Proportion of detectable effects

#### 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 XX.

#### Meta-analytic model

In order to compare the proportion of detectable effects between measures, the data from individuals was transformed and meta-analyzed. For each measure and domain, we calculated the proportion of detectable effects and its variance. We then logit transformed this proportion and entered it into a linear mixed-effects model using the R package lme4. The Wilkinson notation for the model was as follows:

proportion\_diff\_zero\_logit ~ 1 + measure + (1 | domain),

weights = 1/variance

That is, we entered measure as a fixed effect in order to estimate differences between these specific measures (i.e., measures are an exhaustive set for our purposes). Domain was entered as a random intercept in order 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. Results were back transformed from logits to proportions for plotting and reporting. 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 XX. Tables containing the numerical result can be found in the supplementary materials.

Results from the forest plot (i.e., the meta-analytic estimates) were interpreted with the aid of pairwise comparisons between the measures. These were calculated using the emmeans R package while also controlling error rates using Holm correction. Together, these were used to describe the detectible ordinal rankings among the measures from best to worst.

### 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 that each individuals score was detectibly different from. Pairwise comparisons between each participant and every other participant 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 present they were offset by 0.001.

#### Meta-analytic model

The individual level proportions were logit transformed and entered into a similar linear mixed-effects model as the previous one:

proportion\_discriminable\_logit ~ 1 + measure + (1 | domain),

weights = 1/variance

Results were back transformed from logits to proportions for plotting and reporting. 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 XX. Tables containing the numerical result can be found in the supplementary materials.

Similar to the previous analysis, results from the forest plot were interpreted with the aid of pairwise comparisons between the measures, again using Holm correction. Together, these were used to describe the detectible ordinal rankings among the measures from best to worst.

### RQ3. Coverage of Individuals’ Confidence Intervals

#### Calculation of scores

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

#### Meta-analytic model

The proportions were logit transformed and entered into a similar linear mixed-effects model as the previous two:

ci\_width\_proportion\_mean\_logit ~ 1 + measure + (1 | domain),

weights = 1/variance

Results were back transformed from logits to proportions for plotting and reporting. 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 XX. Tables containing the numerical result can be found in the supplementary materials.

Similar to the previous analysis, results from the forest plot were interpreted with the aid of pairwise comparisons between the measures, again using Holm correction. Together, these were used to describe the detectible ordinal rankings among the measures from best to worst.