The reviewers and I appreciated the methodological rigor of the work.   
  
As you will see, Reviewers 1 and 3 concluded that the paper’s primary finding, although based on a novel method allowing assessing individual-level precision, does not force a change in concepts or assumptions pertaining to implicitly assessed attitudes as to warrant publication in a journal with a broad, non-specialized readership.   
  
Reviewer 2 found the paper’s bootstrapping approach to be potentially quite interesting in its own rite but to be better suited to publication in an outlet that affords longer-format papers, which would allow more thoroughly fleshing out the assumptions and benefits of that approach. 

Reviewer: 1

My main reason for this reaction is that the arguments presented here are highly likely to be already well-known among researchers in the field. That is, there are many prior studies highlighting the poor psychometric properties and high levels of measurement error in the indirect measures used here. Granted, the authors are making this case with a novel analysis, but **I am not convinced that such analyses truly tell us something substantively more than what has already presented from prior analyses**.

Response: We now make it clear that what is novel is (i) providing an inference method for differences between participants, which was not previously present, and (ii) a novel method for quantifying individual precision with personalized CIs.

For instance, the paper that originally used these data (Bar-Anan & Nosek, 2014) already does a clear job of illustrating how these measures suffer from generally low levels of test-retest reliability and construct validity (in terms of correlating with other indirect measures meant to assess the same construct). Similar estimates of low measurement reliability have been presented elsewhere (e.g., Connor & Evers, 2020; Greenwald & Lai, 2020), as have poor estimates of predictive validity (Buttrick et al., 2020; Axt, Buttrick & Feng, 2024).   
  
Is it really possible that the above findings could emerge in a world where these measures showed high levels of individual precision? I do not think so, and if it \*is\* possible then the authors have not articulated that position well. Rather it seems that the only conclusion to draw from the great amount of prior work highlighting the poor measurement properties of indirect measures of implicit associations is that they must suffer from low levels of individual precision. Even if that specific point has never been shown before, the consequences of that phenomenon (in terms of weak test-retest correlations or predictive validity) has been documented repeatedly. Granted, the authors would argue that there is value in quantifying just how imprecise these measures are at the individual level and I largely agree with that claim. Indeed, I hope and expect this work will be published in a journal more focused on these issues, but I do not believe Psych Science is that journal.

Response: Yes, but this does not tell us directly about precision of individuals in terms of their actual scores, nor give us a direct sense of what improvement we might want to see in TRR (e.g., would .9 be good enough? What about .8? – these questions can only be answered when translated into how they impact individual precision, and currently we don’t have any intuition on this). Add to this the fact that TRR is only a group-level statistic, whereas bootstrapped CIs are individual-specific. Fair enough reviewer doesn’t think PS is the right venue for this, but their point is kind of silly.

**This variability then calls into question the generalizability of the findings presented here, which are focused on only a small number of domains.**

Response: Fair enough, but there currently does not exist any better dataset with multiple implicit measures.   
  
**Alternatively, the authors could have identified a possible solution to this issue; for instance, how many IATs would a participant need to complete in order to achieve even moderate levels of individual precision (see Carpenter, Goedderz & Lai, 2022 for a similar approach in terms of predictive validity)?**

Response: Separate question and depends on many factors – too computationally intense to go through here and would instead require a simulation study.

**- More information on what are acceptable levels of individual-level precision would be useful to readers that may not be familiar with this metric.**

Response: We don’t know; this is subjective to the goals of the researcher – but we give the implied precision of the Project Implicit criteria in the discussion as an example now.

Reviewer: 2  
  
**In my view, the biggest issue with this manuscript is that it does not provide clear standards against which to evaluate the results of their analyses.** -> we can use a “by your logic” approach where small/medium/large effects minimally should be discriminable from one another

**They label the right side of their x-axis as “Better”, but based on theory alone (as illustrated in my example) we shouldn’t expect any measure to converge on 1.00.** -> we need to clarify that the precision of the instrument is always better for the instrument, and doesn’t make claims about true effects  
  
Reviewer: 3  
  
However, I do not agree with the authors in their assessment that the question of individual-level measurement precision has not yet been investigated (see, for example, the claim in the abstract: “We argue this is because psychologists have not yet even quantified the individual-level precision of these tasks”). For more than two decades, researchers have been concerned with the question of the reliability of implicit measures, calculated either via internal consistency (usually via split-half reliability) or via test-retest reliability. From these reliabilities, the standard error of measurement (SEM) can be calculated very easily using the formula shown in the manuscript on p. 6 (which, however, does not go back to Dudek, as referenced in the manuscript, but can be easily derived from the definition of reliability in classical test theory).  -> but this is not done EVER!  
  
While the calculation of the SEM based on the test-retest-reliability is at least acknowledged by the authors on page 6 (but rejected for reasons I do not quite understand), **the calculation of the SEM based on the split-half-reliability is not even mentioned in the manuscript (this should be changed during a revision).** Instead, the authors use a method for measuring confidence intervals that is not based on overall reliability but on a person-specific bootstrap method.

This bootstrap method for individual confidence interval calculation was new to me and was seemingly first presented in a pre-print by Hussey (2020). I therefore also took a closer look at this pre-print from 2020. In principle, I find this approach interesting and I think it should work. But even in this pre-print, the calculation approach has not been introduced in methodical detail (e.g. using simulation studies), but the approach was applied here to the example of the Implicit Relational Assessment Procedure. In that respect, this pre-print was not as informative as I had hoped, and I am still not sure I got the methodological approach completely right, but it was still helpful in understanding the approach. **In any case, more relevant methodological aspects should be included in the revised manuscript, because I find it unfortunate if you can only understand the central method of a manuscript if you look at a pre-print beforehand.**  
  
Table 2 is an exception here, in which results for the traditional D-score of the race IAT are reported. I used this table to examine the extent to which the authors' bootstrap approach leads to different or the same results as the classical calculation using split-half reliabilities. According to Table 2, the individual confidence intervals based on the bootstrap approach are x +- 0.38 (slightly smaller for higher D-scores). Alternatively, I calculated the confidence intervals using the split-half reliability. Race-IAT split-half reliability in the Project Implicit is .69 (Charlesworth et al., 2023, BRM). The standard deviation for Race-IAT is 0.36443 (as I calculated from the attached data in OSF). This gives SEM = 0.36443\*sqrt(1-0.69) = 0.20. The 95% confidence interval is then x +- 1.96\*0.20 = x +- 0.39**. Interestingly, as I expected, the results are very similar.** However, it would be interesting to systematically investigate whether and under what conditions there are differences or not. -> We need to get ahead of this; say we would expect similarity, but bootstrapping enables this without need of test-retest, or while allowing for individualization of CIs.  
  
I am therefore of the opinion that the following statement by the authors on p. 5 is not correct: “Although some argue that precision can be improved by enhancing test-retest reliability (Greenwald & Lai, 2020), this alone does not quantify individual-level precision. Scheel (2022) recently argued that many claims in psychological research are “not even wrong”, as they are so underspecified that to be wrong would be an improvement. We would similarly argue that implicit measures are currently ‘not even imprecise’; the field lacks tools to even estimate their precision.” I disagree. If the implicit tests had a higher test-retest reliability, this would lead to an enhanced individual-level measurement precision. Good intelligence tests, for example, have a test-retest reliability of >.90, and this is precisely why they can be used for individual measurement with high precision. IATs, on the other hand, only have a stability of .50, which is precisely why they only provide individual measurements with very low precision. I also did not understand the justification with the reference to Scheel (2022) in the quote above. Here, too, I think that implicit measures are by no means “not even imprecise”, but rather have been proven to be imprecise for years because the reliability is insufficient.  