Dear Dr. Cummins:  
  
Thank you for submitting "The individual-level precision of implicit measures" [PSCI-24-0602] to Psychological Science. I sent your manuscript to three experts. Their comments appear below. I thank the reviewers for their thoughtful work. I also have had the opportunity to review the submission myself, and the Statistics, Transparency, and Rigor (STAR) Editor team has done a light transparency check.  
  
The reviewers and I appreciated the methodological rigor of the work. At the same time, each of the three reviewers raised significant issues with the submission.  The nature of the issues is such that I am declining your manuscript for publication in Psychological Science. I am sorry to bear this disappointing news.  
  
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.   
  
These concerns do not take away from the strengths of this research, or its importance. As you know, Psychological Science must be extraordinarily selective due to the large number of manuscripts that are submitted to the journal (around 1,500 new submissions are expected this year). As a result, we must reject even some very good submissions.  
  
I am keenly aware of the disappointment that a letter of this nature brings. I am sorry that the outcome was not positive, and that I cannot invite a revision. The reviewers and I are hopeful you will find the comments in this action letter and appended reviews useful as you consider the next steps in this research program.  
  
Thank you again for considering Psychological Science as an outlet for your work. I hope you will do so again in the future.  
  
Sincerely,  
  
Antonio Freitas  
Associate Editor, Psychological Science  
[psci@psychologicalscience.org](mailto:psci@psychologicalscience.org)

Reviewer: 1  
  
Comments to the Author  
This was a technically-sound and intriguing paper that I imagine would be of some interest to researchers working in implicit social cognition and quantitative methods. The conclusions offered by the authors are mostly well-justified (though there is a lot of extrapolation from an analysis that only included three attitude domains). However, to quote from my own most recent rejection at Psych Science, I do not believe that this work represents the type of “trailblazing discovery” that has become the expectation for publication in this journal.   
  
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.   
  
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.   
  
I’d also like to defend myself against some of the authors potential cynicism to my reasoning, which seems to create an incentive structure where only surprising or shocking research can be published in high-impact outlets, and this push for surprising results leads to weak, unreliable research and a biased literature. I do actually think there is some version of this paper that would be publishable in Psych Science to overcome this issue of novelty relative to the existing literature, but that paper cannot be written using a dataset that included only three attitude domains (race, politics and self-esteem). The meta-analyses themselves show significant variation across domains, so it seems unlikely that the point estimates generated here are all that reflective of implicit associations more generally, and many papers in this area have presented psychometric tests on a wide range of topics (e.g., Nosek, 2005; Cummins et al., 2022). Indeed, each of these papers also shows significant variability across domains in terms of implicit-explicit correlations or internal reliability. This variability then calls into question the generalizability of the findings presented here, which are focused on only a small number of domains.   
  
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)?   
  
Below are some other points that I hope the authors consider in future versions of this manuscript:  
  
- I thought the manuscript did an uneven job of presenting the feedback given to participants on Project Implicit. The manuscript only shows the feedback given to participants in terms of the “weak, moderate or strong” preference shown on the task, but neglects to mention any of the text given to participants to contextualize this score and discuss issues of reliability. For instance, here is what is currently shown on the PI debriefing page right below the feedback:  
  
Disclaimer: These IAT results are provided for educational purposes only. The results may fluctuate and should not be used to make important decisions. The results are influenced by variables related to the test (e.g., the words or images used to represent categories) and the person (e.g., being tired, what you were thinking about before the IAT).  
  
The authors still have the right to disagree with giving any feedback at all to participants given the measurement error present in these tests, but I do think the discussion of this issue could at least be more accurate.   
  
- I think the manuscript would benefit from more discussion about controversies surrounding the AMP; for instance, whether the AMP functions as a truly “implicit” measure is of much discussion. Bar-Anan and Vianello (2018), using the same dataset as the current work, found that the AMP was the only measure that did not load more strongly onto an implicit than explicit construct in structural equation models. I would be curious whether the present analyses speak at all to other work regarding the AMP.  
  
- More information on what are acceptable levels of individual-level precision would be useful to readers that may not be familiar with this metric.   
  
- I understood the reasoning behind using the Probability Index as a single scoring procedure for all measures. But is this not overlooking prior work that sought to identify the best scoring algorithm for each measure (one of the goals of the original Bar-Anan & Nosek paper)? It does not then seem that the analyses are setting these measures up for success if they are overlooking prior work that identified the best way to score them (i.e., the best way to minimize measurement error). This is probably a minor point given that Probability Index scoring and traditional scoring procedures are well-correlated in the IRAP, but still I think the best approach would be to analyze individual-level precision using both the Probability Index and whatever is the most widely used scoring algorithm for each measure. I imagine results would be largely similar across the two approaches but that remains to be seen and demonstrating this would illustrate the robustness of your results.   
  
Reviewer: 2  
  
Comments to the Author  
This manuscript estimates the individual-level precision of 6 implicit measures across 3 content domains. There is a lot to like about this manuscript. It investigates an important question that, in my opinion, has received insufficient in the implicit social cognition literature but is critical for the kinds of claims routinely made about research based on implicit measures. It is methodologically and statistically rigorous, relying on very large participant samples and novel analytic methods. That said, this manuscript falls short of its goals in several ways, which I elaborate upon below.  
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. Consequently, the strong conclusions that they draw in the Abstract (“…implicit measures were extremely imprecise as measures of individual attitudes.”) and Discussion (“…none of the implicit measures appeared suitable for precise individual level inferences…”) come across as unsupported. We as a field have agreed-upon guidelines / heuristics for what constitutes acceptable measurement reliability (alpha > .7), multicollinearity (VIF <4 or <10, depending on who you ask), etc. For a measure to be reliable, are there established guidelines for interpreting what proportion of participants need to have non-zero effects, what proportion of differences can be detected between participants, or what proportion of the observed range is covered by individual CIs? If I take the Results section of this paper at face value, I can clearly see that the IAT has a much higher proportion of participants with non-zero effects, and higher proportion of participants whose scores can be distinguished from one another, than do either the AMP, GNAT, and EPT. That’s an important and interesting finding. But nothing in this manuscript as written provides clear support for the conclusion that they’re all unsuitable. The authors’ conclusions don’t follow clearly from the data. Crucial information or argument connecting the data to the conclusions seems to be missing.  
Similarly, the authors do not clearly articulate their rationale for the 3 research questions they test. I mean, I can kind of intuitively understand the first one, why we would want a measure of individual differences to consistently produce scores that can reliably be distinguished from zero. Distinguishing scores from zero conceptually underpins null hypothesis significance testing, so that general approach makes sense. The authors’ specific approach to testing this research question seems to assume that \*all\* scores should be different from zero (in one direction or the other). However, in the context of implicit attitude measures, we should not actually expect that. In fact, there’s a lot of theory & empirical evidence that some participant groups should have scores close to zero. For example, Black people completing a Black-White racial IAT might be expected to have pro-Black evaluations because of positive ingroup experiences, ingroup-enhancing motivations, etc, but at the same time have pro-White evaluations because of positive outgroup cultural exemplars, system-justifying motivations, etc. Consequently, the net product of these competing evaluations should produce an IAT score close to zero, reflecting ambivalence (but not indifference, as some researchers incorrectly assume). Again, the problem here is that the authors don’t provide an argument for what the threshold of acceptable non-zero effects should be. 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.   
I also don’t understand their rationale for their second research question. They cite Greenwald and Lai (2020) as noting that our field lacks high-precision implicit measures of trait differences between individuals. I’ve been publishing in this field for a little while now, and I have to confess that I don’t understand why we would want a measure to be able to distinguish among individuals in the way the authors of this manuscript are testing. To be sure, we definitely want to distinguish between \*groups\* of people (e.g., treatment versus control group), and I myself use the exact same analytic method the authors use here (i.e., observing whether the 95% CI of the difference of distributions includes 0) in my own research. But why would we expect each individual to be a unique flower? On the contrary, theory and empirical evidence would seem to suggest that we should expect a degree of implicit attitude clustering. As the field of regional/geographic psychology argues, people in close geographic proximity share common cultural influences, and lots of theories argue that responses on implicit measures reflect culture to varying degrees. Thus, some people’s implicit biases should be pretty similar to other people’s biases (i.e., people with whom they share culture/geographic proximity). Moreover, the example feedback from Project Implicit that the authors summarize in Figure 1 aligns with their first research question (“Are you biased, and if so how much?”) but doesn’t dovetail with their second research question (“How biased are you compared to other people?”). The authors are going to need to do more to unpack the logic underpinning this second research question, ideally offering some sort of threshold for judging an acceptable amount of detectable differences between individuals. Surely, 1.00 can’t be that standard.  
Their third research question suffers from the same limitation. It’s intuitive (to me, at least) that smaller CIs are better. But, again, given that the authors want to draw absolute conclusions about whether any of these measures are suitable for individual-level inference, they will need to offer some sort of standard against which they can judge each measure.     
The Measures and Methods sections of this manuscript seem to be light on important information. I think that the single-sentence descriptions of most of the measures will make this a challenging read for anyone who isn’t already an expert in the implicit social cognitive literature. I was not familiar with the Probability Index before reading this manuscript, and even after reading it I still feel like I have only the most cursory understanding of it. For example, I get that PI=.5 reflects an equal probability that a randomly-selected data point in one block type is different from a randomly-selected data point in another block. However, I seem to be missing the logic behind how to interpret PI>.5 and PI<.5. Given that the authors report that PI correlates highly with the IAT D-score, I assume that values > versus <.5 correspond to compatibility effects somehow, but any deeper understanding is lost on me. And on a very minor but related point, the authors should define acronyms before using them. They refer to “PI=0.50” on p.11 but don’t define PI until p.12.  
This manuscript includes a number of tables and figures, which are great for summarizing information. However, the authors devote almost no text at all to interpreting the information summarized in the tables and figures. I have no idea at all what to make of the maximum a posteriori CI distribution values reported in Table 1. The AMP seems to have wider CIs than the other measures, and I assume this is a case where values closer to 0.00 are better than values closer to 1.00. But again, without any standard of judgment, I can only draw relative conclusions (AMP is worst) but not absolute ones (whether any are acceptable). The same goes for Tables 3, 4, and 5. The pairwise comparisons provide very strong evidence for relative conclusions – which in my view could be a real strength of this manuscript – but the authors don’t interpret or discuss them at all.   
Even Table 2, which is seemingly straightforward, could be improved. As written, Table 2 would seem to suggest that the appropriate way to describe an IAT D score = .15 is as “weak negative to moderate positive bias”. However, I don’t read the big-picture takeaway of this manuscript to be to assign new labels to the Project Implicit cutoffs, but that’s the message that Table 2 conveys. Instead, I think an important message of this manuscript is that the level of measurement precision in IAT D-scores means that D=.15 \*could\* reflect bias that ranges from weakly negative to moderately positive. In its current form, Table 2 conveys a different message – a message I think is incorrect.    
Taken together, this manuscript seems to be missing a considerable amount of information necessary to support the conclusions the authors draw. As someone who periodically submits my work to Psychological Science myself, I suspect that a lot of that important information ended up on the cutting room floor in pursuit of this journal’s tight word limits. I just don’t think that the story behind these really interesting data can be told in a short report format – or, at least, the authors have not presented strong evidence for their claims in this draft of the manuscript. Again, I want to reiterate that I think that these are really interesting data and have the potential to make an important contribution to, and perhaps beyond, the implicit social cognitive literature. The aim of this work – to quantify and evaluate the individual-level precision of implicit measures – is crucial, and I look forward to seeing it published in some form in some journal somewhere. But given the tension between Psychological Science’s tight word limit and it’s aim to be a general-interest journal, I fear that this manuscript in its current form is too light on supporting information to effectively achieve the goals the authors have set for themselves.    
  
  
Reviewer: 3  
  
Comments to the Author  
The manuscript deals with the question of the precision of individual measurements of implicit tests. I see this as a very important question as to whether these measurements can be used for individual assessment, because this is a prerequisite for feedback such as that given in the Project Implicit (see Figure 1 in the present manuscript), but also, for example, for individual assessments in the context of training on implicit bias. I agree with the authors' general evaluation that implicit measures do not have sufficient measurement precision for individual testing.   
  
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).    
  
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.    
  
Conceptually, the bootstrap approach should in my view be similar to the calculation of the SEM based on internal consistencies (since both are based on the same information), but with the difference and potential advantage that person-specific confidence intervals can be calculated that vary between participants. This may be an advantage, especially when using probabilistic index (PI) scores (DeSchryver et al., 2018), which are bounded at 0 and 1. While I can understand why the authors opted for these PI scores in their analysis, I also see it as a problem that this makes it more difficult to connect this paper to previous literature. This could lead some supporters of implicit measures to consider the present work irrelevant because the “correct” scores were not used. At this point, I would therefore like to recommend that the authors consider reporting the results for the traditional measures as well.  
  
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.  
  
From a diagnostic point of view, however, I would personally advocate the use of test-retest reliability instead of split-half reliability (or the bootstrap approach), because I think that at least with the IAT the split-half reliability might be artificially increased for methodological reasons. With a test-retest correlation of .50 for the race IAT, I calculate a confidence interval of  x  +- 0.51, which is significantly larger than the intervals stated by the authors in this manuscript.  
  
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.    
  
I liked Figures 1 and 2 (from the results section; there is also a Figure 1 at the front, so the numbering is not quite correct) because they nicely illustrate the significance of large confidence intervals. However, in my view, the IAT performs quite well here, particularly in comparison to the other measures, reflecting the comparatively high internal consistencies (i.e. split-half reliabilities) of IATs in the range of .60 to .80, while the split-half reliabilities for AMP, GNAT and EPT are much, much lower. However, if confidence intervals based on test-retest reliability had been used here, the results would certainly have been far more sobering for the IAT.   
  
My overall conclusion is that the results reported so far in the manuscript are not really new, because it is already known from the well-known low split-half and test-retest reliabilities that individual-level precision of implicit measures is low (although I see that this cannot be said often enough). The bootstrap approach, on the other hand, is new, but in my view it is not yet sufficiently introduced and explained in the manuscript for a method that is not yet established. A reference to the pre-print by Hussey (2020) is not sufficient here, but a revision of the manuscript should, in my view, introduce the bootstrapping method in more detail so that reading the pre-print is no longer necessary to understand the method. It should then also be investigated (in addition to the preregistered analyses) to what extent the results of the bootstrap approach differ empirically from those of the SEM calculation from classical test theory. Then the potential of the bootstraps approach to generate findings that go beyond those already known could be better recognized. In particular, the point of confidence intervals varying between individuals could be potentially interesting here. For example, it could be that implicit bias can be precisely recorded in the IAT for some people, while this is not the case for others. Such an analysis could not be investigated using the SEM approach from classical test theory because this approach assumes a constant confidence interval for all individuals.