Reviewer(s)' Comments to Author:  
Reviewer: 1  
  
Comments to the Author  
Olivier Corneille (I always sign my reviews)  
  
This research relied on bootstrapping to assess individual-level measurement precision in six implicit measures applied to three attitudinal domains. The analyses were applied to a very large database. Results suggest generally low levels of precision, although the authors acknowledge in the discussion that how precise a measurement should be depends on particular study objectives. No recommendation is made regarding how precision may be enhanced in these instruments, but the present research did not aim to address that question.  
  
I thought that the approach was interesting, the writing clear, and the discussion fair and cautious. Below, I provide three major and four minor comments, the consideration of which may help to strengthen the manuscript further. At the bottom of my comments, I provide the reference to three recent review articles, all of which discuss and refer to a large number of references supporting the three main points. I direct the authors to these articles should they seek any clarification.

Response: We thank the reviewer for their positive assessment and appreciate their thoughtful and useful comments, to which we respond below.  
  
1. Why should we even care?  
Even assuming measurement precision would be good or could be significantly improved, we would still not know (1) why we need measures of this sort nor (2) whether it makes sense to compare them :  
  
(1) The use of implicit measures like the IAT is based on a series of theoretical and empirical assumptions that have been proven wrong or remain unsubstantiated. The fact is: we still don’t know why we should be using them.  
  
(2) Implicit measures represent neither a meaningful theoretical nor empirical set. There is no theoretical relation between these tasks because the only distinct feature they share is to rely on indirect assessment, which has no theoretical implication. These measures also weakly converge with each other at the empirical level, which makes perfect sense considering that they widely differ on a structural level. Why comparing measurement precision in e.g. IAT versus GNAT performance when these tasks differ by so many features that divergences/convergences between them cannot be reliably interpreted anyway? In all likelihood, these tasks measure different things, thereby making precision comparison questions a perplexing objective. As a further and aggravating concern, we still don’t exactly know what is supposed to be distinctly measured by any of them (see also below)

Response: We agree with the reviewer; in fact, our personal opinions, based on having completed two PhDs on the topic, is that implicit measures are not particularly useful instruments. However, as the reviewer notes, they are based on empirical assumptions, many of which are wrong or substantiated. The goal of this paper was to test one of these empirical assumptions which has not yet been tested to date, and our evidence broadly suggests that this assumption (of individual-level utility and precision) is also wrong. In this sense, then, this paper adds to this emerging set of critiques regarding these measures. It also highlights yet another weakness in the field: the fact that a fundamental premise relating to their use and development has not been subjected to substantive empirical examination in the 25 years they have existed. Given that they are still widely-used, why we should care seems obvious: because this is how the scientific record gets corrected.

Regarding (2), we again agree with the reviewer that this does not represent an empirically meaningful set. However, we should also not confuse *ought* with *is*: it is a matter of fact that social psychologists choose between these different measures when designing their experiments, if nothing else because they want to choose to “best” indirect measure. Comparing the psychometric properties of tasks, even when they measure disparate constructs, is commonly done; e.g., we may compare the reliability of several different tasks used in a study even when they are meant to measure putatively distinct constructs. The question of “do these measures measure the same thing” is important, but so is “do these measures measure their construct(s) of interest precisely”.

With all of this said, we hope that the updates to our manuscript in light of the reviewer’s comments can better contextualise the goal and relevance of this research: to provide a method for estimating, and put data on, a feature of tasks that has been overlooked to date, despite its apparent theoretical and practical importance.   
  
2. Precision assessment should probably be applied to the latent process(es) level.  
Research shows that performance on these tasks is driven by independent latent processes that are conflated with each other unless advanced modelling is applied to the data (see for instance Quad modelling applied to IAT performance). It would probably make sense to apply precision questions to that process-level. For instance, measurement precision (assuming it could be estimated at the individual level) may be larger when applied to estimates of control abilities than when applied to representational influences underlying task performance.

Response: While we do not disagree that this may be interesting, we note that the modal use of these measures involves the empirical level, not the latent level (e.g., most studies which use the IAT tend to use the empirical scores). However, we now add a point in our discussion which highlights that researchers may also be interested in applying these analyses at the process-level.

3. Whilst implicit measures are often considered measures of individual differences, researchers have also identified other goals - none of which have fared better. This may be highlighted, too. Here is a sample of such goals that can be identified in the implicit attitude measures literature:  
  
1. A measure of something that is indirectly measured (which is theoretically useless).  
2. A measure of mental constructs.  
3. A measure of unconscious mental constructs.  
4. A measure of associative processes.  
5. A measure of simple associative representations.  
6. A measure of automatic behavioral responses.  
7. A measure of how mental constructs are expressed in behavioral responses.  
8. A measure of distinct cognitive processes that independently contribute to overall task performance and that can be isolated through computational modeling.  
9. A measure of implicit biases, such as shooting decisions, in lab-analogs of real life situations.  
10. A measure of “gut reactions”.

Response: We agree; we have now added a paragraph in the discussion highlighting some of these issues.   
  
Minor:  
  
4. I’m not sure the screenshot is necessary.

Response: We believe that the screenshot has explanatory value. However, we update its description considering the information provided by Dr. Bar-Anan in his review below.

5. line 10 “beliefs and stereotypes”. Because stereotypes are defined as beliefs, I was not sure why these were opposed.

Response: Amended.

6. I could not find the preregistration after clicking on the osf link referred to in the manuscript. Please provide clearer information about how to access this file (and the code file).

Response: Amended.

7. Needless to say, I totally disagree with this the conclusion (p. 28, lines 17-18) that “Researchers may now have a sense of how exactly to interpret individual scores on implicit measures.»

Response: We have updated the sentence to be more accurate, namely: “Researchers may now have a sense of how precisely estimated individual scores on implicit measures are.”.

Summary assessment : It is important to assess and draw attention on the lack of precision of implicit measures. In my view, however, a high-order question is why we need to address such questions in the first place, considering that these measures generally have little added theoretical and practical value and cannot be compared to each other in any meaningful way in the absence of structural fit. In my view, it is unfortunate that these pressing questions were entirely overlooked in the present manuscript. Thankfully, the current review gave me the opportunity to raising them again.

Response: We hope that we have sufficiently addressed the concerns of the author, as well as conveying our agreement with them.

Way forward: If the authors have convincing answers to these questions, I believe they should be discussed in the manuscript. Alternatively, the concerns that these questions raise may be discussed in addition to the more specific (measurement precision) concerns evidenced in the present research. To be clear, I am perfectly fine with either approach, as long as the manuscript addresses the broader questions I am raising here, the nature of which I believe should take precedence over the more specific questions the authors focused on in the present research.

Response: While we respect the reviewer’s statement that they believe these questions should take precedence over the more specific questions of the manuscript, we respectfully also note that if these issues would take precedence then this would be a different paper than the one we have written. We hope that we have effectively (i) conveyed that we are in alignment with these points, (ii) implemented discussion of these points in the manuscript to highlight these facts, (iii) contextualized how our work here fits into such a critical lens, and (iv) adequately addressed the reviewer’s concerns.   
  
Potentially useful references to three review articles that have elaborated on the points above, each of which includes many additional useful references. I'm mentioning them for the sake of clarification - obviously NOT for the sake of citations.  
  
<https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1521%2Fsoco.2020.38.supp.s165&data=05%7C02%7Cjamie.cummins%40unibe.ch%7Cc4f852c694a34bea36ac08dda92c0a32%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638852729142462836%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=hv2HZuC%2F2VxaLGdB9i8DshgTV%2Flhbmcou5kTSziy9n8%3D&reserved=0>  
<https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1038%2Fs44159-024-00376-z&data=05%7C02%7Cjamie.cummins%40unibe.ch%7Cc4f852c694a34bea36ac08dda92c0a32%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638852729142485111%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=YAdnSIiGwBPi4UyqXvx%2FL9sYDTdOcwtCeOuUHnY0Bt0%3D&reserved=0>  
<https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1177%2F1088868320911325&data=05%7C02%7Cjamie.cummins%40unibe.ch%7Cc4f852c694a34bea36ac08dda92c0a32%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638852729142506388%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=hR1Vu4hBsldeAMQoP5SLXEEUau2WqVv4t%2FIBXX%2FntGA%3D&reserved=0>  
  
Reviewer: 2  
  
Comments to the Author  
Signed: Yoav Bar Anan  
  
1. The paper makes a valuable contribution. It raises awareness of the subject of precision of an individual score and provides some comparative estimates of that precision across popular measures in social psychology. Having little knowledge about individual-level precision, I cannot judge in confidence the quality of the theoretical and methodological content. Yet, the major arguments and analyses seem reasonable to me. I only have a few comments the authors might want to consider in a possible revision.

Response: We thank the reviewer for their positive assessment of our work.

2. Project Implicit does no longer provide the kind of feedback that authors state it does. It only reports the participants results (e.g., “You were much faster at sorting 'Black people' with 'Bad' and 'White people' with 'Good' than 'White people' with 'Bad' and 'Black people' with 'Good'”), the distribution of results across many participants, and what meaning researchers typically attach to results on the group level. (in full disclosure: I am a member of Project Implicit's scientific advisory board, and I am the one who implemented this change in the code of Project Implicit’s IAT).

Response: We greatly appreciate the reviewer’s comment regarding this, and have updated the description of our Figure to reflect this fact. We retain the figure (caveating that its content is now outdated) in order to reflect the fact that individual-level description is demonstrably a goal of interest to the implicit measures research community.

3. It would greatly enhance the informative of the research, if the authors compare the indirect measures with direct measures. In the dataset the authors analyzed for this research. We (Bar-Anan & Nosek) also had speeded and non-speeded rating of the items that were used in the indirect measures (at least for political and race attitudes). These measures had only a few trials (perhaps 10), but maybe precision parameters can be computed from them. The three questionnaires (Rosenberg, MRS, and RWA) may also provide some comparison regarding what precision can be achieved with more reliable measures.

Response: Although it would be interesting to examine the precision of the speeded response measures, we noted three issues (i) the limited number of trials renders comparison with the implicit measures not particularly meaningful; we would not be comparing like-with-like, and the implicit measures’ scores would be more precise, (ii) researchers in general have not talked about speeded response measures as individually-diagnostic in the same manner as implicit measures, and (iii) the individual-item responses appear not to be present in the data. However, we did compute these estimates for the RSE, MRS, and RWA (although given the limited number of items for the MRS in particular, its estimates will by definition be imprecise). These results are presented in the supplementary materials.

4. The authors mentioned that they did not include the SPF in their analysis because ”(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.”   (a) is true. I think that about 90% of published SPF studies were conducted by me. Regarding (b), the SPF includes the same conditions as the EPT (e.g., white-good, white-bad, black-good, black-bad), so the scoring of the EPT can also apply to the SPF.

Response: We have updated this sentence to reflect more accurately the description of the SPF and to reflect the fact that we, the authors, were just generally not that familiar with the SPF.

5. In a supplement, it may be good to show the same results for the scores commonly used for each task. The authors argued that it is better to score all the tasks similarly. I am not sure that it is true, unless the authors’ scoring is optimal for all measures. If not, it seems better to score each task in its most recommended scoring algorithm. That includes the most recommended practices for trial-exclusion, error-response treatment, and participant-exclusion. It will be unfortunate if researchers in the field will choose to ignore this paper due to uncertainty regarding the preparation and scoring of the data.

Response: We now add these analyses in the supplementary materials, and highlight in the main manuscript that our results are robust to the scoring approach used.

6. In the same context, it might be useful to know that we (Segal-Gordon & Bar-Anan, 2024) published an improved scoring algorithm for the EPT.

Response: We have now added reference to this paper as a suggested alternative score.

7. Please consider adding a readme file to explain the processing and analysis files.

Response: We have now added a README to explain the files, as well as why one would run them. We have additionally added further comments within the files to give further context to readers.

8. It was not entirely clear what (if anything) the authors recommend researchers do to improve individual-level precision. Are there any clear steps for improvement in the context of indirect evaluation measures?

Response: We have now added additional paragraphs to discuss this issue; namely, to suggest some approaches to improving precision, as well as to identify a smallest effect size of interest to identify the criterion of precision required.   
  
9. In the figures, the authors used “SC-IAT” (Single-Category IAT) and in the table they used “ST-IAT” (Single-Target IAT). It would be less confusing if the authors choose one name for this measure.

Response: Amended – we’ll stick with ST-IAT.

Reviewer 3:  
  
This study aims to evaluate the individual-level precision of commonly used implicit

measures by analyzing a large-scale dataset using a unified scoring approach

(Probabilistic Index (PI) scores) and bootstrapped confidence intervals. The findings

suggest that current implicit tasks lack the precision needed to make meaningful

individual-level inferences. I strongly agree with the authors that improving individual-

level precision is essential, and that increasing task lengths and benchmarking

methodological precision are important steps not only for implicit measures but for

psychological science more broadly. I offer the following suggestions to help

strengthen the manuscript:

Response: We thank the reviewer for their positive assessment of our work.

1. One methodological strength of this work is the use of PI scores to

standardize the scoring of diverse implicit tasks, enabling direct comparisons

of individual-level measurement precision. However, this approach may come

with a trade-off in construct fidelity. While PI is psychometrically valid, it may

not be the most construct-valid scoring method for all tasks. Each task was

originally designed with a scoring method aligned with its theoretical

underpinnings (e.g., D scores for the IAT, prime-consistent response rates for

the AMP). Replacing these with PI scores may obscure important task-

specific properties. This, for me, seems akin to comparing Stroop and stop-

signal tasks using raw reaction times rather than construct-sensitive metrics

like the congruency effect or stop-signal reaction time. While PI facilitates

comparisons, the authors might consider whether retaining task-specific

scoring could provide more construct-valid insights. Perhaps it may be helpful

to report bootstrapped confidence intervals for each task’s original scoring

metric as supplementary materials.

Response: We now add these analyses in the supplementary materials, and note that our results are robust to the choice of scoring approach used.

2. This study shows that none of the six tasks achieves sufficient individual-level

precision. It would be useful to specify what sufficient might look like. Without

a benchmark or “gold standard,” it's difficult to assess how far short current

measures fall, or what threshold future measures should aim for. I believe the

point the paper is making surely is defining such a benchmark is currently

challenging due to limited prior data. However, offering preliminary criteria or

guidance for establishing one would enhance the paper’s utility. For example,

how narrow should a confidence interval be, or what proportion of participants

should show detectable effects, for a task to be considered precise enough for

individual-level inference?

Response: In our opinion, this question is a little like asking “what should an effect size be?” for a group-level study design: we simply cannot answer it as it is so heavily dependent on the researchers’ questions and inference goals. However, we have now added an additional paragraph sketching out *what factors* a researcher might need to consider when motivating their study for individual-level comparisons, and broadly outlining *how* one could use this information to make informed design decisions. We hope this is satisfactory to the reviewer – we do appreciate the desire for specificity and recommendations on these types of criteria, but just do not feel comfortable nor qualified to make blanket recommendations on an inherently study-specific topic.

3. If the key contribution of this work is methodological—as the authors suggest

in the discussion (“the most important aspect is providing a framework to

assess the precision of implicit measures”)—then a visual or graphical

workflow could help readers understand how to apply the framework to their

own data. For example, what would an “ideal” caterpillar plot look like, and

how narrow should the CIs be to indicate sufficient precision? This could help

researchers calibrate and benchmark their own tasks moving forward.

Response: We hope that the newly-added paragraph in the discussion aids with this.

4. Related to the former point, I think the legend for Figure 1 could be more

descriptive. For example, clarify that the dashed line represents the “neutral”

score (PI = 0.50), what it means when a participant’s confidence interval

crosses that line, and which measures exhibit many such cases. Since some

readers may focus on the figure without reading the full results section, a

more self-contained legend would improve clarity.

Response: (1) we realized this Figure was erroneously labelled 1 rather than 2, and have corrected this; (2) we have now updated the description in line with the reviewer’s recommendations.

5. The IAT consistently outperformed other measures in terms of precision, but

the manuscript does not fully explain why. Were there specific features—e.g.,

task length, design, scoring assumptions—that contributed to this superior

performance? Understanding these differences could provide practical

guidance for improving future implicit tasks.

Response: We now include a paragraph which highlights some of the methodological aspects of the IAT which may have given rise to its relatively superior precision, as well as recommending that researchers may compare different methodological features of these different tasks in order to gain inspiration in how to improve precision in the future.

6. The data used in this study were collected online, which is often associated

with greater measurement error compared to in-lab data. Is online data

collection representative of current practices in the field? If so, it would be

valuable to discuss how this may have influenced the findings in the literature.

Could the relatively low precision observed be partly due to the data collection

mode? Would precision estimates look different in controlled lab settings?

Response: We have now added a paragraph discussing this question.

7. Given that 10% of participants completed more than ten measures, it seems

feasible to examine rank-order consistency across tasks measuring the same

domain (e.g., multiple measures of implicit racial attitudes). While not the main

focus of this paper, such an analysis could provide additional support for the

claim that individual-level precision matters—more precise measures should

show higher consistency in ranking individuals. If not feasible here, this could

be a valuable direction for future research.

Response: We appreciate the reviewer’s thoughtful suggestion to examine rank-order consistency across measures targeting the same domain. While we agree that such an analysis could provide a useful external benchmark for the importance of individual-level precision, this was beyond the scope of our current study, which focused on quantifying precision within each task. Additionally, the practical number of participants who completed a sufficient number of overlapping tasks within the same domain is not very large in the sample, limited the usefulness of this dataset for this purpose. Therefore, we determined that it would be better that such an investigation be conducted with data which are more appropriate to this question, or that better yet have been collected with this explicit purpose in mind. Nonetheless, we have added a paragraph to the Discussion section highlighting this as a valuable direction for future research.