Dear Corina,

We are happy to make these additional adjustments. Our changes are described in detail below.

Lindsay J. Alley
Jordan Axt
Jessica K. Flake

Response to recommender: Corina Logan

Thank you for your revision, which appropriately addressed the remaining comments.

In drafting my recommendation text, I went back through your answers to the questions at the submission page and determined that your answer to question 7 was incorrect. This is a quantitative study that tests a hypothesis (see research questions 1 and 2 in your introduction), therefore it needs a study design table. Please make a study design table according to the author guidelines (https://rr.peercommunityin.org/help/guide_for_authors#h_27513965735331613309625021), reference it in your main text, and include the table in your manuscript document either in the main text or as supplementary material. Note that research question 2 tests a hypothesis because it uses a significance test that tests for the existence of something, not just the amount of something. Research question 1 is an estimation problem, but would also benefit from being included in the study design table.

We have completed the study design table and uploaded it to the osf page for our study. Additionally, we have added the following reference to the table on page 13 of our manuscript:

"For an overview of the design of our study to answer each of our research questions, see the Study Design Table in our supplementary materials (osf.io/ht48z)." p. 13

For your convenience, here is the table in full:

Question	**Sampling plan**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**	**Interpretation given different outcomes**	**Theory that could be shown wrong by the outcomes**
RQ1. To what extent do measures function equivalently across different convenience samples in the Many Labs projects?	Using the previously collected open data from the Many Labs projects, we will examine every measure that meets our criteria for baseline model fit. We will use only data from participants collected in English.	We will test the equivalence of loadings (metric equivalence) and intercepts (scalar equivalence) using likelihood ratio tests for each measure and sample group pair examined at $\alpha = .05$. If the equivalence of all loadings or intercepts is rejected, we will test the equivalence of parameters at the item level using univariate score tests at $\alpha = .05$ / the number of items. We will also calculate and report dMACS effect sizes at the item level.	According to our review of the simulation literature on the likelihood ratio test for detecting measurement non-equivalence, we most likely have power of 80% or greater for tests involving only the 9 largest samples we are examining. Tests involving the 5 smaller samples may be underpowered and results will be discussed with caution.	If all measures are equivalent across all convenience samples: these samples are likely to display measurement equivalence. The pooling of samples in the ML was justified, and pooling or comparing measurements using others samples from these sources without correcting for non-equivalence is likely to be justified in future cases, though not guaranteed. If some measures are equivalent across convenience samples but others are not: measurement equivalence for convenience samples is dependent upon the construct and/or the specific measure. It should be tested or accounted for if measures from these data sources will be pooled or compared. If some crowdsourced samples are equivalent with student samples and others are not: measurement equivalence across convenience samples is dependent on the specific source, rather than being generalizable across crowdsourced and student samples more broadly. Interpretation will depend on the pattern of results. Given the sample from India, language and culture may be a more reliable source of non-equivalence than convenience sample type.	The theory that measurement properties are equivalent across convenience sample sources (student and crowdsourced). This theory is assumed by the pooling of these data sources using uncorrected sum scores in the ML projects.

				If all measures are non-equivalent across all convenience samples: data from these sample sources should not be pooled or compared without considering potential measurement differences, as they are likely to be a reliable source of non-equivalence. Pooling these samples was not justified in the ML and may have impacted results.	
RQ2. When measures are non-equivalent, does correcting for this change the statistical significance or effect sizes of the replications?	Based upon the analyses conducted for RQ1, we will examine for RQ2 only the measures and samples which demonstrate configural equivalence but display statistically significant metric or scalar non-equivalence.	We will develop a partial equivalence model for each measure and sample pair on the basis of the results of the univariate score tests from RQ1. This model will restrict parameters found to be equivalent so they are equal across groups and free parameters that display statistically significant non-equivalence. We will generate factor scores from this multiple group model, which will correct for the non-equivalent parameters. We will reproduce the replication effects using these factor scores and compare these results to the effects estimated using original scoring methods. To determine whether effect sizes are different, we will calculate 95% confidence intervals.	Answering this research question will itself constitute a sensitivity analysis. We are not attempting to make inferences to other cases with these analyses; rather, we are aiming to describe whether the presence of measurement non-equivalence has had an impact on the estimation of effects in the ML replications.	If the results of the replications are not changed by correcting for non-equivalence, then, while the pooling of the samples was not justified in the cases where they displayed non-equivalence, the results were robust to this. If the results of the replications are changed by correcting for non-equivalence, then these findings are not robust to the presence of non-equivalence. This may serve as a cautionary note and impetus for changing research practices of researchers pooling or comparing samples from these sources, although the results will not necessarily generalize to other cases, as the robustness of findings depend on particular features of the data in each case.	This analysis is not attempting to disprove any theory, but rather explore the robustness of the ML findings to the presence of measurement non-equivalence.

Also, please update your answer to question 7 in the report survey to choose the first option: "YES - THE RESEARCH INVOLVES AT LEAST SOME QUANTITATIVE HYPOTHESIS-TESTING AND THE REPORT INCLUDES A STUDY DESIGN TEMPLATE". If you need help with this, or would like me to change your answer for you, just let me know.

We have updated our response to this question.

All my best,

Corina