Dear Corina,

Thank you for the input on our manuscript. We appreciate receiving such helpful and constructive reviews. We have addressed all reviewer comments in bold below and adjusted the manuscript at the indicated page number. Thank you for your time and consideration. Please let us know if there are any additional changes you would like to see before moving on to the next stage of the registered report process.

Lindsay J. Alley
Jordan Axt
Jessica K. Flake

Response to Reviewer 1: Benjamin Farrar

In this Stage 1 Registered Report submission, Alley et al. propose to test for measurement non-equivalence between online and student samples across ManyLabs samples, including student samples, MTurk samples and a sample from Project Implicit. The analysis to decide which scales to include for equivalence testing has already been performed, and this is appropriately outlined in the level of bias control section of the RR. I was able to fully reproduce the inclusion analysis from the provided code.

We are happy to hear that you were able to reproduce the analysis, thank you for checking this.

Overall, I consider this to be a well thought-out and very well written submission that will be widely read, and it is great that the authors have chosen to submit this as a RR. The code annotation, in particular, was very clear - which is appreciated. The research question is valid and there is a strong rationale for testing measurement equivalence between convenience samples. While I am not an expert in testing measurement equivalence, the analysis pipeline was clear and, combined with the annotated code, there is sufficient methodological detail to closely replicate the proposed study procedures and to prevent undisclosed flexibility in the procedures and analyses.

Thank you for the kind comments, we are very glad to hear that you feel the research questions and rationale are valid and clear.

I could not find the individual analysis code mentioned in the paper (pg. 15 “Code for the following analyses can be found in the supplementary materials. A separate R file for each measure, and the files are named for the measure analyzed) but this may be a placeholder for when each individual analysis has been performed, and I consider the “Planned Analysis Code” to sufficiently outline the analysis pipeline for all measures, which I was able to fully reproduce.

You are correct that this text is meant as a placeholder to indicate how the code will be organized once analyses are completed at stage 2. We have added the following clarifying text:

"[Note: only the Planned Analysis Code is included at stage 1. The other files described will be added when full analyses are completed for stage 2.]" p. 16

Below I list minor comments and questions for the authors, but I am confident this is a scientifically valid, thorough and very valuable submission that I look forward to reading once it is complete.

General:
Throughout the report, there was not much discussion of the power of the planned analyses to detect measurement inequivalence. This may be a moot point as the sample sizes in the current project are quite large, and the exact analyses to be conducted are contingent on the hierarchical analysis plan. However, it would be a good addition to have a power curve to make explicit the range of effect sizes that would be detected in a "typical" analysis, and their theoretical meaning – although to some extent this is already achieved in the sensitivity analysis, which is great.

We agree that the consideration of power is important and that we neglected the issue. Unfortunately, developing a power curve for the analyses that we are undertaking is complex. The power of the test of measurement equivalence is impacted by many features of the specific data and model, including the number of items, the strength of the inter-item relationships, and, in addition to the degree of non-equivalence (the effect size), the type and structure of the non-equivalence (Is it the loadings, intercepts, or both? Are they consistently higher in one group, or is it mixed? How many items display some type of non-equivalence?).

In order to estimate power with any accuracy, we would need to run a series of simulations for each of the nine measures we plan to examine. In lieu of this, we have added a paragraph discussing the issue of power, as we agree on its importance and regret our initial neglect of the issue. The following paragraph has been added to page 19/20:

> **"In addition to unbalanced sample sizes, it is important to consider the impact of sample size on power, as results of statistical tests should be interpreted with caution in situations where the power to detect a meaningful effect is insufficient. Power for the χ^2-difference test of the equivalence of loadings and intercepts across groups is complex, as it is influenced not only by sample size and the amount and degree of non-equivalence, but also by many other features of the data and model, including: the strength of the loadings for non-equivalent items (Meade & Bauer, 2007), whether the direction of the non-equivalence is uniform or mixed (i.e. some loadings higher and some lower in the focal group, versus all loadings lower in the focal group; Meade & Bauer, 2007), the number of factors (French & Finch, 2006; Meade & Bauer, 2007), and the number of items per factor (Finch & French, 2018; French & Finch, 2006).**
>
> **Simulation research on the χ^2-difference test of the equivalence of loadings has found that, for sample sizes of 150 to 200 per group, power varies substantially based on these features (as low as .29 or as high as .95; French & Finch, 2016, 2006; Koziol & Bovaird, 2018; Meade & Bauer, 2007). For sample sizes of 400 to**

500 per group, power is generally high: while one study reported power of .57 in a condition with 500 per group (French & Finch, 2006), this was an anomaly, and every other study reported values of .89 or greater (French & Finch, 2016; Koziol & Bovaird, 2018; Meade & Bauer, 2007). Of the 14 sample groups that we plan to examine, 5 of them have a sample size less than 400, and one of these is below 300 (the online student sample in ML2 slate 1). As such, we expect that results involving these sample groups should be interpreted with caution."

Additionally, we have added this as a limitation to discuss in the point form draft of our discussion section on page 26:

"Power: tests involving the 5 smaller samples may have had low power to detect meaningful non-equivalence. We will highlight which effects these are and discuss implications a lack of power might have had. We will also discuss in detail any other factors we observe that are known to decrease power, such as low loadings and poor model fit."

Ordered minor comments

Page 6: This contains a very clear and accessible description of measurement equivalence. Please provide a more explicit formal definition also be provided after the first sentence, with any appropriate references.

We have added a more detailed definition to the beginning of this paragraph at the top of page 6. The first few sentences now read:

"Also called measurement invariance, measurement equivalence is concerned with whether a particular scale is measuring the same thing in the same way across different groups. Formally, this means that, for a given level of the latent trait, the conditional distribution of the items of the measure is the same across subpopulations (Meredith & Millsap, 1992). Thus, within a latent variable modelling framework, "measuring something in the same way" means that the items of the scale are related to the latent variable in the same manner across groups."

Page 8: "It would also be ideal for replication researchers to test for ME between original and replication studies (Fabrigar & Wegener, 2016), but this is difficult in practice: for the most part, the original studies that are replicated do not have publicly available data." Presumably the small sample size of many original studies would also mean it would be difficult to detect meaningful measurement invariance between samples.

We have amended this sentence as follows to include sample size considerations:

"It would also be ideal for replication researchers to test for ME between original and replication studies (Fabrigar & Wegener, 2016), but this is difficult in practice: for the most part, the original studies that are replicated do not have publicly available data and have small sample sizes (Fraley & Vazire, 2014). This is a barrier to detecting measurement non-equivalence, as sample sizes of approximately 400 per group are recommended to detect meaningful effects (French & Finch, 2016; Koziol & Bovaird, 2018; Meade & Bauer, 2007)." p. 8

Page 9: If the authors want a reference on between-platform differences in participant responses: Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 54(4), 1643-1662.

Thank you for providing this reference, we have added it to page 10:

> **"MTurk is the most studied online crowdsourcing platform, but research on differences from student samples may not generalize to other, similar data-collection platforms. Peer et al. (2022) found that data from Prolific, CloudResearch, Qualtrics, and Dynata differed from MTurk in terms of demographics and data quality."**

Page 9/10: "While mean differences on a measure of a construct do not necessarily indicate non-equivalence, any of these differences in sample characteristics could potentially contribute to non-equivalent measurement." Could this be unpacked? If there are substantial differences in some sample characteristics, is absolute equivalence plausible, or would you expect some non-equivalence but perhaps with very small effect sizes?

It is still possible for measurement to be equivalent across groups even if they differ substantially in terms of sample characteristics. We view differences on a latent trait as only being valid to interpret if there is measurement equivalence. However, when data are collected from heterogenous groups, it is important to consider if measurement equivalence holds, as people from groups that differ from each other are more likely to understand items differently as compared to those from more similar groups.

To clarify this point, we have added the following:

> **"While differences across samples do not necessarily indicate non-equivalence, differences in sample characteristics could potentially contribute to non-equivalent measurement for particular constructs, as respondents from groups that differ from each other may understand items differently. However, it is also possible that very different people interpret items in the same way, and, therefore, these groups could still be equivalent in terms of measurement properties for a given construct. It is important to examine the issue directly." p. 10**

Page 11: RQ2, it could be made explicit here that this will be tested by looking at effects sizes as well as significance.

This has been updated to:

> **"RQ2. When measures are non-equivalent, does correcting for this change the statistical significance or effect sizes of the replications?" p. 12**

Page 12: "which is useful more broadly than just replication research" – good justification, if anything you are unselling the importance of these research questions for interpreting the results of any experiments!

We agree! And thank you for your endorsement of the study's importance. We have added the following to page 13 to emphasize the broader relevance of the study:

"Understanding whether different convenience samples are likely to display measurement non-equivalence will aid in the interpretation of all studies that use these samples and contribute to building a cumulative psychological science."

Page 12: "which requires a series of preliminary analyses to determine if the data satisfy assumptions". It would be useful to list all of the assumptions here.

We have edited the first sentence of this section to read:

"The primary proposed analysis is psychometric equivalence testing. We performed these tests using MG-CFA with maximum likelihood estimation, which requires that the data meet the assumptions of the estimation method (multivariate normal, sufficient response options to approximate continuous) and, additionally, that the baseline measurement model is adequately specified (French & Finch, 2011)." p. 13

Page 13: "Type I error rates for equivalence tests may be inflated when the baseline model is misspecified". For accessibility, could an example of model misspecification increasing Type 1 error rates be given here, even if hypothetical?

We added a sentence giving an example of misspecification inflating error rates:

"Type I error rates for equivalence tests may be inflated when the baseline model is misspecified (French & Finch, 2011), resulting in a higher probability of incorrectly concluding that a scale is non-equivalent across groups. For example, if a measure is modelled as unidimensional, but the items in fact load onto two factors, an equivalence test for this incorrectly specified unidimensional model would be more likely to find non-equivalence across groups, even though the true, 2-factor model is equivalent." p. 13/14

Page 17: The authors may be more aware of this than I am and this may be a non-issue, but some concerns have been raised about using scaled χ2-difference tests to compare nested models (see http://www.statmodel.com/chidiff.shtml)

Yes, this is a concern. One cannot simply subtract one χ2 value from another to construct the difference statistic if you are using scaled χ2 values. In this case, we are using the lavTestLRT function in R, with the approach from Satorra and Bentler (2001), which has been shown to be a valid approach to constructing scaled χ2-difference tests. We changed the relevant sentence in the following way to make this clear:

"To evaluate the tenability of each level of parameter restrictions, we compared each nested model to the next most restricted one using Satorra and Bentler's (2001) approach to calculating the scaled difference χ^2 statistic." p. 19

Results: The proposed data presentation looks appropriate, and thank-you to the authors for including these.

Thank you.

Finally, I don't think this is necessary to address in the current project, but I thought about this several times while reading the manuscript: While the focus is on detecting measurement inequivalence between samples, there is some ambiguity about what the authors consider the sample to be – is it the experimental units (individual participants) only, the experimental units plus the setting (participants + online/location), are the experimenters included too? Replications vary across experimental units, treatments, measurements and settings (Machery, E. [2020]. What is a replication?. Philosophy of Science, 87(4), 545-567), and so it might be worth considering the degree to which measurement invariance could be introduced from other sources, e.g. different experimenters, that would not normally be associated with the sample, but nevertheless would differ between them.
Best of luck with the project!

This is definitely an important consideration and something we have thought about as well. To better explore this issue, we have added the following to our discussion (p. 27):

> **"In order to examine equivalence across convenience samples in this project, we had to make decisions about how to deal with other plausible sources of non-equivalence. We opted to collapse across many groups, such as experimental conditions, participant gender, and participant race, all of which can contribute to non-equivalence. We also completely eliminated translated instruments, which are known to be a source of non-equivalence, by only using English versions of measures. If we considered every possible subgrouping, and clustered respondents into only those exactly like them, the groups would be too multitudinous and fine-grained to proceed with any examination. If we had sufficient data to do so, we could consider more groups simultaneously using the alignment method (Asparouhov & Muthén, 2014) for equivalence testing. However, given the available subgroup sample sizes in many cases, the issue necessitates some simplifying decisions regarding which features are likely to be relevant for a given measure. As a result, a limitation of this work is that we cannot be sure that the decisions we made are the right ones."**

Benjamin Farrar

R2
Review for "Convenience Samples and Measurement Equivalence in Replication Research"

I first write that I am not a psychologist, but I am an ecologist/evolutionary biologist who is interested in meta-research. Therefore, my comments sometimes may sound slightly strange to psychologists. Anyway, I will do my best

This study will use datasets from the Many Labs replication projects to test whether convenient samples (uni students and crowdsourced ones) can be considered equivalent. This seems like a great idea. I really enjoyed reading this. My lab has done many registered reports, published protocols and detailed pre-registrations. But this registered report is

much more detailed and thought thoroughly than we have ever done (I really like these empty tables and bullet points for Discussion). Having said this, I have several comments, all of which are just minor. I write them down not in order of importance.

1: Page 5 "RMSEA < .05, CFI > .95, SRMR < .08"
What are these abbreviations? They are mentioned for the first time. More explanation will be needed.

We regret this wasn't clear, we have altered the top of page 5 to read:

> **"of the six scales examined, none met all three fit index cut-offs selected (root mean square error of approximation [RMSEA] < .05, comparative fit index [CFI] > .95, standardized root mean squared residual [SRMR] < .08)."**

2: Page 5 "a multiple group confirmatory factor analytic (MG-CFA) approach "We need a bit more explanation of what it is and what it does

We have added the following to page 5 to explain the approach more clearly:

> **"Confirmatory factor analysis (CFA) is a statistical modelling approach which aims to represent "the causal relations between one or more unobserved, or latent, variables and a set of observed variables" (Flora, 2017), and MG-CFA is the extension of this approach to model multi-group data, allowing for the detection and modelling of differences due to group membership."**

3: "If this approach is to be effective, it's important that aggregated samples demonstrate ME" But it seems OK to aggregate as long as you model these two different samples.

This is a good clarifying point, and we agree. We have altered the quoted sentence to read as follows:

> **"If this approach is to be effective, it's important that aggregated samples demonstrate ME, or that researchers employ a statistical model that accounts for non-equivalence across samples." p. 9**

4: I think the authors are mainly talking about what is known as "selection bias" (in medicine) and "corridor bias" (in casual inference). Do the authors want to mention these words so this work will be more relevant to a broader audience?

We would be happy to add these terms, as we would like our paper to be relevant to a broad audience and understand that differences in disciplinary language may impede this. Unfortunately, we are not clear which specific part of the paper this is referring to or how to address it. We leave this to the recommender to provide guidance as to if and how we can address this comment.

5: OK, this is my biggest point. So the authors have huge sample-sized datasets. Naturally, inferential statistics will be significantly different, leading to non-equivalence. Any thoughts on this? Also, this relates to my Point 3. To me, it seems to be OK to have these two samples mixed regardless of equivalence or not. The default is to deal with these as different populations, including these as a predictor (fixed or random effects).

My point is that my default position is that everything is different, so if you have enough sample sizes, everything will be non-equivalent. So we should be modelling and treating these different samples differently from the start without ME tests. Maybe I missed something on the way.

We agree that these are important issues to think through, and have a few points of clarification to offer that we have also incorporated into the paper to address this thoroughly:

1. **The tests that we are employing use maximum likelihood estimation, which is a large sample technique different from least squares which is often used for regression or analysis of variance. Flagging negligible effects may be a concern for some of our tests, but not across the board, as a few of our samples are at or just below recommended size for adequate power. In response to R1, we have clarified the sample size constraints for the proposed analyses on page 18/19.**
2. **While we are not using these for decision making, we will also be examining dMACS (Nye & Drasgow, 2011), an effect size index for quantifying the importance of degrees of measurement non-equivalence. This will allow us to report and reflect on whether the degree of non-equivalence detected is of any practical importance. We have incorporated your concern on page 19 when we discuss this effect size index:**

   > **"Due to the fact that we may find statistically significant, but not practically significant non-equivalence, we also report dMACS effect sizes (Nye & Drasgow, 2011), though these were not used for decision making (code 2.3). Based on simulation studies by Nye et al. (2019), when less than 50% of the items are non-equivalent, we consider dMACS > .40 to be practically significant; and when 50% or more are non-equivalent, we consider dMACS > .20 to be practically significant."**

3. **Additionally, with the sensitivity analysis we will examine whether correcting for non-equivalence changes the effect sizes or statistical test results of the replication studies from which the data were collected. This will speak to the practical importance of any non-equivalence we detect. It is entirely possible that the tests we run will detect non-equivalence, but the sensitivity analysis will show that it does not meaningfully impact the downstream results of the replication studies. Additionally, when and how much any non-equivalence changes results could be related to the amount and degree displayed. We hope to explore and discuss this issue once our results are complete. We have added the following to our point form discussion draft to acknowledge this issue (p. 25):**

   > **"The results of our sensitivity analysis will speak to the degree to which any non-equivalence detected is of practical importance to researchers. We will discuss the results here, including whether and how non-equivalence contributed to the replication effects, and what features of the non-equivalence (loading vs intercept, effect size, number of items, direction of non-equivalence) impacted results."**

4. **Finally, we agree that modelling measurement differences across groups in all cases would be ideal. However, it is clear that this does not happen in psychology, even for groups that are known to exhibit measurement differences (Boer et al.,**

2018). Additionally, in most datasets there are many plausible ways to create groups that might display non-equivalence, including culture, language, gender, educational background. These amount to more subgroups than is generally feasible to model. For this reason, it is useful to have research that can highlight which groups are important to model, and which it might be reasonable to ignore. In response to reviewer 1, we have added a point addressing these issues to our draft discussion section on page 26.

Of course, I can understand the authors' main purpose for doing this, as these two groups of convenience samples are quite different. Then, we should not be surprised if studies do not replicate using different types of samples. But I guess then it is really interesting if some phenomenon is robust regardless of their differences. I look forward to how this study will turn out.

This is a complex issue. It is true that differences across samples may impact whether an effect is found in both and may also impact whether there are measurement differences across these subgroups for a particular construct.

However, even more fundamentally, if these samples are interpreting the questions they are asked in a very different way, this alters the meaning of any results that do not account for these differences. Without exploring the issue of measurement equivalence, we cannot be confident in our interpretation of replication results. We hope to discuss this further in our paper, though what we have to say will be strongly impacted by our results. We have added the following to our draft discussion to indicate our intention to dig into this further:

> **"If there is undetected non-equivalence in replication studies, this may impact the meaning of the results. Given our findings, we will discuss the degree to which this may be a concern for convenience samples." p. 25**

Signed

Shinichi Nakagawa

Other changes not requested by reviewers:

On the advice of a colleague, we have switched from using Bartlett's factor scores for the sensitivity analysis to regression factor scores. We have changed the references to this on page 21, and altered code 3.1 and 3.2.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Boer, D., Hanke, K., & He, J. (2018). On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734.

Finch, W. H., & French, B. F. (2018). A Simulation Investigation of the Performance of Invariance Assessment Using Equivalence Testing Procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 673–686.

Flora, D. B. (2017). *Statistical Methods for the Social and Behavioural Sciences: A Model-Based Approach*. SAGE.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019.

French, B. F., & Finch, H. (2016). Factorial invariance testing under different levels of partial loading invariance within a multiple group confirmatory factor analysis model. *Journal of Modern Applied Statistical Methods: JMASM*, *15*(1), 511–538.

French, B. F., & Finch, W. H. (2006). Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402.

French, B. F., & Finch, W. H. (2011). Model Misspecification and Invariance Testing Using Confirmatory Factor Analytic Procedures. *Journal of Experimental Education*, *79*(4), 404–428.

Koziol, N. A., & Bovaird, J. A. (2018). The Impact of Model Parameterization and Estimation Methods on Tests of Measurement Invariance With Ordered Polytomous Data. *Educational and Psychological Measurement*, *78*(2), 272–296.

Meade, A. W., & Bauer, D. J. (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. In *Psychometrika* (Vol. 57, Issue 2, pp. 289–311). https://doi.org/10.1007/bf02294510

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How Big Are My Effects? Examining the Magnitude of Effect Sizes in Studies of Measurement Equivalence. *Organizational Research Methods*, *22*(3), 678–709.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. *The Journal of Applied Psychology*, *96*(5), 966–980.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662.