Peer Community In
Registered
Reports

Data from students and crowdsourced online platforms do not often measure the same thing

A recommendation by ***Corina Logan*** *based on peer reviews by* ***Benjamin Farrar*** *and* ***Shinichi Nakagawa*** *of the STAGE 2 REPORT:*

Lindsay J. Alley, Jordan Axt, Jessica Kay Flake (2023) Convenience Samples and Measurement Equivalence in Replication Research. OSF, ver. 2, peer-reviewed and recommended by Peer Community in Registered Reports. https://osf.io/s5t3v

Submitted: 01 September 2023, Recommended: 13 November 2023

Cite this recommendation as:
Logan, C. (2023) Data from students and crowdsourced online platforms do not often measure the same thing. *Peer Community in Registered Reports*, 100551. 10.24072/pci.rr.100551

Published: 13 November 2023

Comparative research is how evidence is generated to support or refute broad hypotheses (e.g., Pagel 1999). However, the foundations of such research must be solid if one is to arrive at the correct conclusions. Determining the external validity (the generalizability across situations/individuals/populations) of the building blocks of comparative data sets allows one to place appropriate caveats around the robustness of their conclusions (Steckler & McLeroy 2008).

In the current study, Alley and colleagues (2023) tackled the external validity of comparative research that relies on subjects who are either university students or participating in experiments via an online platform. They determined whether data from these two types of subjects have measurement equivalence - whether the same trait is measured in the same way across groups.

Although they use data from studies involved in the Many Labs replication project to evaluate this question, their results are of crucial importance to other comparative researchers whose data are generated from these two sources (students and online crowdsourcing). The authors show that these two types of subjects do not often have measurement equivalence, which is a warning to others to evaluate their experimental design to improve validity. They provide useful recommendations for researchers on how to to implement equivalence testing in their studies, and they facilitate the process by providing well annotated code that is openly available for others to use.

After one round of review and revision, the recommender judged that the manuscript met the Stage 2 criteria and awarded a positive recommendation. **URL to the preregistered Stage 1 protocol:** https://osf.io/7gtvf

Level of bias control achieved: Level 2. *At least some data/evidence that was used to answer the research question had been accessed and partially observed by the authors prior to Stage 1 IPA, but the authors certify that they had not yet observed the key variables within the data that were used to answer the research question AND they took additional steps to maximise bias control and rigour.*

List of eligible PCI RR-friendly journals:

- Advances in Methods and Practices in Psychological Science
- Collabra: Psychology
- F1000Research
- Peer Community Journal
- PeerJ
- Royal Society Open Science
- Studia Psychologica
- Swiss Psychology Open

References:

1. Pagel, M. (1999). Inferring the historical patterns of biological evolution. Nature, 401, 877-884. https://doi.org/10.1038/44766

2. Steckler, A. & McLeroy, K. R. (2008). The importance of external validity. American Journal of Public Health 98, 9-10. https://doi.org/10.2105/AJPH.2007.126847

3. Alley L. J., Axt, J., & Flake J. K. (2023). Convenience Samples and Measurement Equivalence in Replication Research [Stage 2 Registered Report] Acceptance of Version 2 by Peer Community in Registered Reports. https://osf.io/s5t3v

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.17605/OSF.IO/HT48Z
Version of the preprint: 1

Authors' reply, 06 November 2023

Download author's reply
Download tracked changes file

Decision by Corina Logan, posted 27 September 2023, validated 27 September 2023

Minor revision

Dear authors,

Congratulations on completing your study and finishing your Stage 2 article! The unexpected bumps that came up along the way were normal, and your solutions to the problems upheld the scientific integrity of the registered report - nice work. The same two reviewers who evaluated the Stage 1 came back to evaluate Stage 2, and both found that your manuscript meets the PCI RR Stage 2 criteria. I am ready to issue IPA after you revise per my and Nakagawa's comments. Note that my comments are so minor that you do not need to address them if you feel they are not useful, but please do make sure to address Nakagawa's comment.

To answer your question, there are no space constraints at PCI RR so you don't need to move anything to supplementary material.

Here are my comments on the manuscript...

1) Results: I found it extremely useful that you clarified the size of the effects in relation to what your tests were powered for (e.g., "Item 1 ("I find satisfaction in deliberating hard for long hours") was the only item above the cut-off for a medium effect, all others were small or negligible"). I noticed that some paragraphs discussed the a small effect being the cut-off, while others discussed a medium effect being the cut-off. It might be even clearer if you noted in each paragraph that the effect size cut-off related to the power/sensitivity/etc analyses you conducted at Stage 1 for each analysis, which is why it differed.

2) Discussion: "power in ME testing is impact by the strength of inter-item correlations" - change "impact" to "impacted"

3) Discussion: "For this reason, researchers should not assume that different crowdsourced samples will be equivalent to each other, or even student samples collected in different settings". Could you please clarify what "different settings" refers to? Different countries/languages/etc.?

4) Study design table: you could add a column to the right that shows your findings.

I'm looking forward to receiving your revision.

All my best,

Corina

Reviewed by Benjamin Farrar, 12 September 2023

2A. Whether the data are able to test the authors' proposed hypotheses (or answer the proposed research question) by passing the approved outcome-neutral criteria, such as absence of floor and ceiling effects or success of positive controls or other quality checks.

Yes. The authors clearly outline when equivalence testing was unsuitable due to failing to find an appropriate anchor item, and also clearly discuss the pontential limits to the generalisability of their results.

2B. Whether the introduction, rationale and stated hypotheses (where applicable) are the same as the approved Stage 1 submission. This can be readily assessed by referring to the tracked-changes manuscript supplied by the authors. (Note that this is the DIRECT link to the tracked changes version: https://osf.io/download/hgku8/?direct%26mode=render)

Yes, the small number of tracked changes relate to changes of text or transparent and justified changes to the methods.

2C. Whether the authors adhered precisely to the registered study procedures.

As above, yes. In the occasion the authors added a criterion for determining configural equivalence after the Stage 1 process, this was clearly stated in the manuscript and was accompanied by a footnote to explain the nature of, and reason for, the deviation from the planned methodology. It was reported that this decision was made by an author blind to the results, that is appropriate.

2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative.

As above, yes.

2E. Whether the authors' conclusions are justified given the evidence.

In my view, the author's have conducted a very thorough study that followed the Stage 1 approval. The conclusions that are justified by the evidence. The writing is of a very high standard throughout, and the discussion of the generalisability of their results is fully appropriate.

I re-ran the code for one comparison (EMA implicit vs MTurk) as a reproducibility check, and I was able to fully reproduce the equivalence test results for this comparison, although I did note the mean age for MTurk was 34.98400 (35.0) rather than the 34.0 reported. The code is clear, and excellently commented on throughout.

Yours sincerely,
Ben Farrar

Reviewed by Shinichi Nakagawa, 26 September 2023

I have reviewed stage 1 of this MS and very much enjoyed it and was looking forward to reading stage 2. I first acknowledge that I am a quantitative ecologist so I do not know the relevant field and literature. Yet, I would be able to check whether the statistical analyses conducted were sound. Also, this is my first time reviewing stage 2, but my understanding is that I check whether they followed the stage 1 plan and check for deviations. The authors conducted the study with very minor deviations. I liked that the Discussion section had limitation and recommendation sections, which are very clearly and honestly written. Overall, I think this is a great stage 2. I have one question, tho. By reading this work, I got the impression that authors are encouraging to be cautious about mixing samples. Yet, some papers in biology encourage the mixing of samples knowing non-equivalence (differences, e.g. sex and strains). I wondered what authors make of this, and there should be some related discussion. I note this mixing process is called "heterogenization", which is encouraged by an increasing number of grant agencies. There is an example paper:

Voelkl, Bernhard, et al. "Reproducibility of animal research in light of biological variation." Nature Reviews Neuroscience 21.7 (2020): 384-393.