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.

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

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.

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?

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.

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.

Signed

Shinichi Nakagawa