Review for “The WEIRD problem in a “non-WEIRD” context: A meta-research on the representativeness of human subjects in Chinese psychological research”

Reviewer: Kai Hiraishi (Keio University, Tokyo, Japan, kaihiraishi@keio.jp)

Summary

This study aims to describe the characteristics of Chinese participants in psychological research. As a psychologist from a non-WEIRD country (Japan), I share the authors’ concerns on the representativeness of our samples and highly commend the plan's goals. That said, I would like to address several points that I believe need clarification. These are a) the definition of representativeness, 2) Bayes factor analysis plan, and 3) some concerns on the coding manual. I have some other concerns besides those mentioned here. But I would first like to have the authors’ response on the problems I have raised in this review before we proceed to elaborate on more detailed examinations of the research plan.

Definition of “representativeness”

I agree with the authors that we need to collect representative samples for psychological studies. Then, how should we define the “representativeness” of a sample? Two points should be noted.

Representativeness, Similarity, and Diversity

The first point is about the “similarity” and “diversity” of population and samples. Imagine that our population age distribution is so skewed that 80% of the population are in their 20s (of course, this is an extreme example). Then, we find that 80% of our psychological study samples are in their 20s. Does that mean that we have successfully collected representative samples? Not necessarily. Indeed, it depends on the research questions. If we are interested in the psychology of that particular population, we may be qualified to declare so. However, if we are to study human universal psychological phenomena, I do not think that such a claim is well justified. As there are plenty of reasons to suspect that the psychology of the 10s, 30s, 40s, and so on is different from that of the 20s, it may be better to collect more data from those individuals. Put differently, we may need to over-represent minority group members when we want to know universal human phenomena. Merely reflecting the population distribution may hinder the sample diversity and overlook important heterogeneity in human psychological phenomena.

Given the arguments above, I would request the authors clarify how they define the “representativeness” of the sample. More specifically, I would like to know the condition(s)

on which the authors conclude that the Chinese psychology samples well represent the population. If they think that the samples should have the same (or similar) distribution as the Chinese population, the samples composition should resemble the census. If the authors think that the sample should cover wider subgroups within the population, it may be required that the samples include equal numbers of data points (individuals) from each subgroup (e.g., age bin).

Representativeness/Diversity of what?

As mentioned above, the qualification of sample representativeness depends on the research question of a particular study. For instance, we will not conclude that a sample is biased even when it is composed of individuals from a very narrow age range (e.g., 0 to 12 months of age) if the study is about the development of infant vision. We need to carefully consider what the “bias” means when a meta-research finds it.

The apparent bias in sampling may not be a bias in participant sampling. Suppose that we have a population composed of equal numbers of individuals in their 20s, 30s, 40s, and 50s (i.e., 25% of pupulation in each age bin). A meta-research finds that 80% of the psychology samples are in their 20s. If most of the studies in the meta-research state that they deal with universal human phenomena, we can legitimately conclude that the sampling of participants in the field is biased and should be amended. However, if we find that most of the studies state that they are dealing with the psychology of adolescents, I do not think that their sampling of participants is biased. It is the research topic sampling that is biased in the field. That kind of bias may reflect the cultural, political, economic, and historical backgrounds of the field and the society. I am not sure if I can legitimately argue that the bias should be amended.

If we go one step further, we will find that even when the meta-research finds that the whole sample distribution matches that of a population, we are not necessarily entitled to conclude that the field successfully collects representative samples. Suppose that there are four major topics in the field, A, B, C, and D, all of which deal with universal human phenomena. Then we find that samples for topic A are mostly composed of individuals in their 20s. Likewise, samples of B are biased to the 30s, C to the 40s, and D to the 50s. Even though the samples for each research topic are highly biased, the field as a whole appears to have representative sample.

Given these considerations, I am particularly concerned about how we should interpret the results if we find the distributions differ between the samples of papers in Chinese journals and the samples of international collaborative projects. The difference may reflect the bias in participant sampling on whichever side, or may reflect the bias in topic sampling. In addition, even when they appear to have similar distributions as a whole, they may differ in the structures.

I suppose that coding data on generality conclusion of each target article may help to solve the problem. As such, I would request authors to elaborate on their plans on how to utilize the

article generality statement data in their analysis (please also see my comments on the coding manual).

Hypotheses in Bayes Factor analysis

Let me first state that I am not familiar with Baysian approach. What I write below may completely miss the point. I recommend the handling editor (Recommender) to find another one or two reviewers with expertise in Baysian statistics.

As I read the abstract and the introduction of the manuscript, I had an impression that the authors are mainly concerned with the characteristics of Chinese samples in international collaborative projects (hereafter, ICP). For instance, the authors wrote,

> *These (internatiolinal collaborative) projects, however, have not examined whether data collected from non-WEIRD regions are representative of the local population. Left this issue unaddressed, these large collaboration projects may create an illusion that the diversity problem can be solved by involving more researchers from non-WEIRD regions, ignoring the fact that data collected from non-WEIRD regions may suffer a problem of representativeness...*

If my understanding is correct, the hypothesis should be something like “the ICP samples are representative of Chinise populations” and we will test it with the data (actual charactersitics of the ICP samples). Specifically, in their Bayes Factor analysis on sample age distribution (Question 1), the authors may set H0 to the age distribution of subjects in Chinese psychology journals. That is, the null hypothesis is that the samples of ICP are as biased as those of psychology studies reported in Chinese journals. The H1 may be set as the age distribution in the census data. That is, the alternative hypothesis is that the ICP samples are representative of Chinese population age structure.[1] The Bayes factor (BF10) will indicate to what extent the data (actual age distribution of ICP data) supports the H1 relative to H0.

As I read the manuscript, though, the authors seem to set H0 to the age distribution of ICP samples while H0 to the multinomial distribution with equal probability for each age bin. This seems to be different from what I have proposed above.

> *The data from Chinese psychological journals will be as observed and the data from international collaborations will be used as the expected. More specifically, for the sex distribution, we will test whether sex ratio of subjects from Chinese psychology journals is sampled from the population with a sex ratio equals to that of the samples from international collaborative projects. The null hypothesis (H0) is that observed data are sampled from the population with parameter equals to that of Chinese samples from international collaborative projects). The H1 is that the observed data are sampled from a multinomial distribution with equal probability for each.*

[1] If authors put more emphasis on sample diversity, the alternative hypothesis (H1) may be something like that “the ICP sample have the same numbers of participants in each age bin,” as proposed by the authors in their analysis plan.

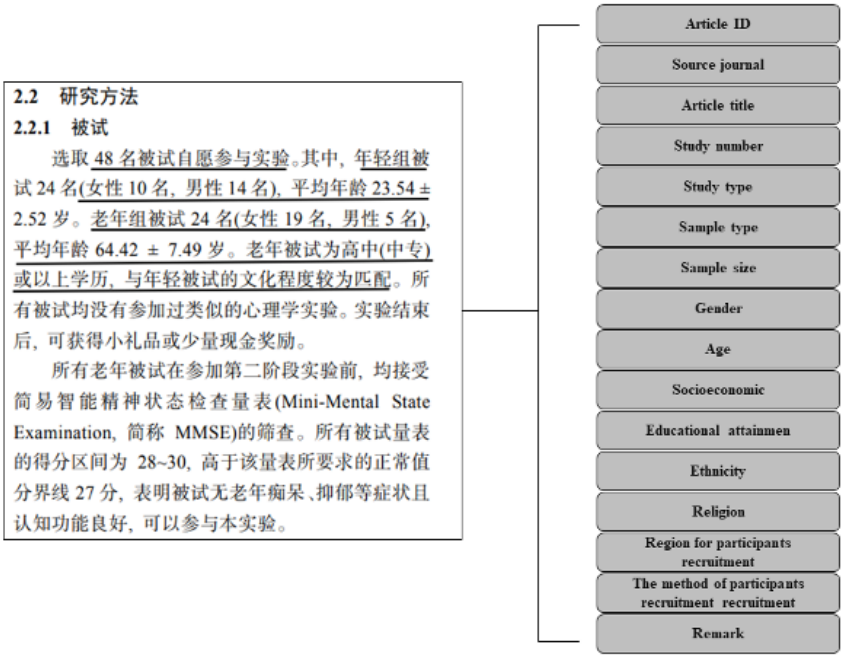
In addition, I am not sure if the prior employed by the authors is uninformative one. For the analysis of age distribution, authors employed a multinomial distribution with $P = Pr$ (*x1, x2, ... x7 | n = 100, p1, p2, ... p7*) and set *p1* to *p7* to be equal: [1/7, 1/7, ... 1/7]. I am afraid that setting the probability of each outcome (x1 to x7) to be equal is a relatively strong assumption. Put differently, isn't it the same as settig θ to be 0.5 in a binomial distribution? As I mentioned earlier, I do not necessarily think that H0 should have an uninformative prior (e.g., $\theta \sim$ Uniform [0, 1] in a binomial distribution) for the BF analysis. But, if the authors are to estimate the parameters' posterior probability distribution, it may be better to think of other less informative prior.

I repeat that I am not familiar with Baysian approach and am afraid that I may completely misunderstand the research questions and the analysis plan. Therefore, I would like to leave the problem to the handling editor and other reviewers who have expertise in Baysian analysis.

Concerns on the Coding manual

Coding of subgroup information

This is rather a minor comment. The coding manual instruct the coder to do unnecessary and problematic merging of information from two subgroups. The example article compared elderly participants with younger participants. The paper clearly described that they have collected 24 participants for each age group and also reported the age characteristics of each group. However, the coding manual requires to report only the total sample size (that is, 48). I am concerned that this procedure may distort the description of sample characteristics.



Sample size: 48

Gender:

reported

female: 19; male: 29

Age:

reported

23.54±2.52; 64.42±7.49

Coding of the generality conclusion

This is related to the points I have mentione above regarding the definition of representativeness. I think it is very important for the current study to collect data on generality/specificity statement of each target article. However, the manual does not provide detaild instruction on this dimension and simply refers to Rad et al. (2018). Even though Rad et al. (2018) provided relatively detailed description of their coding criterion, I do not think it is specific enough. I request authors to elaborate on this part before they start collecting data. For instance, the coding strategy employed by Rad et al. (2018) only requires to write down whether the article made any statements on constraints of generality of the results. But I think it is important to code the range of generalizability that was declared in each article (e.g., generalizable to children from 8 to 10 years-old in Eastern Asia).