Dear Prof Yamada,

Thank a lot for your patience. We are extremely sorry for this delayed response, which was partly caused by the recent spread of COVID in China. Now we have addressed three reviewers' comments. Please see our point-by-point response below and the changed traces in our manuscript.

We hope you will find the revised manuscript acceptable for in-principle acceptance.

Sincerely,

Hu Chuan-Peng,
Lei Yue

Reviewed by Kai Hiraishi

Thank you for revising the manuscript. I like the authors' responses and the revisions. I have some minor comments and suggestions. Please see below.

Bayesian analysis

Thank you for explaining the planned Bayesian multinomial test. I think the explanation below is very informative. It helps me a lot to understand what the authors intend to test.

"The null hypothesis (H_0) is that the sample counts are generated by a specified set of population proportions. The alternative hypothesis (H_1) is that the sample counts are not generated by those population proportions."

This means that in JASP, the alternative hypothesis for Bayesian multinomial test is an unconstrained alternative hypothesis. That is, H_0 is specified as a multinomial distribution with a specific set of parameters, but the alternative hypothesis H_a is simply not H_0, i.e., all other possible parameters.

Then, I have few questions.

First, I am concerned about the relationship between the above explanation and the description on page 9 stated below.

The percentage data from one source is treated as the observed and the other is treated as expected. The null hypothesis (H_0) is that the observed percentage data are sampled from a multinomial distribution with parameters as defined by the expected percentage, the alternative hypothesis (H_a) is that the observed proportion data are sample from a multinomial distribution with equal probability for each cell.

Here, it seems that the definition of Ha is different from that in the first explanation (i.e., H_a is not H_0). Could you give more explanation?

Response 1: Thanks for pointing out this inconsistency, now these two parts are as follows:

"*Usually, the H_0 is tested against the encompassing hypotheses, or alternative hypothesis (H_a), that all category proportions are free to vary.*"

Second, I am concerned what will be the H_0 for the current study, especially for the first question.

The authors wrote that the first question is,

Firstly, whether the characteristics of Chinese samples reported in large-scale international collaborations are similar to those reported in Chinese psychological journals. (page 4)

Given the sentence, I have supposed that the sample characteristics of the studies reported in Chinese psychological journals will be the H_0, and that of the international collaborative studies (hereafter big-team studies) will be the "sample count." The Bayesian multinomial test will examine if the big-team study participants are sampled from the same (or similar) population as the participants reported in the Chinese psychological journals. However, the planned analysis is the other way around (on page 11 of the manuscript).

As I am unfamiliar with Bayesian analysis, I may be making a fundamental mistake. Nevertheless, it seems more natural to treat the big-team study data as the "sample count" if the authors are mainly concerned with the characteristics of the big-team samples. I suspect that many potential readers unfamiliar with Bayesian analysis would have a similar question. I would like the authors to explain in more detail why they set the big-team sample characteristics as H_0.

Response 2: Thank you for carefully reading our phrase, now we have revised the second paragraph on page 11 to make these two parts consistent with each other. Also, see below for the revised paragraph:

"*The data from international collaborations will be as observed and the data from Chinese psychological journals will be used as expected. More specifically, for the sex distribution, we will test whether the sex ratio of subjects from international collaborations is sampled from a distribution with parameters equal to the proportions of data extracted from Chinese psychology journals. The null hypothesis (H_0) is that the parameters of observed data are equal to the Chinese samples from Chinese psychology journals. The H_a is that the parameters of the observed data are free to vary instead of a fixed point."*

Target population coding

I agree with the authors that representativeness has not been taken seriously in psychology. Given the situation, I like how the authors plan to code the target population, to code the explicitness of the target population description, and extract the exact sentences/words that describe the target population.

Response 3: Thanks for this suggestion. First, we revised the codebook so that the revised version will code every article's target population. If the article did not explicitly state its target population, we will infer the target population based on certain words used in the discussion and

introduction part, such as “humans”, “people”, etc. In the end, the target population of every study will belong to one of the four categories: stated specific population; inferred specific population; inferred general population; stated general population.

As you suggested, we will compare sample characteristics between Chinese samples in big-team science and samples in Chinese journals based on their target population. More specifically, samples from studies that targeted at generalizing to all humans or Chinese population will be compared with each other, while samples from studies that targeted at generalizing to a specific population will be compared with each other. If the target population is human beings or Chinese population in general, inferred or stated, we will also compare the characteristics of these studies to the census data/CFPS.

We have revised our manuscript as below (see page 11, the first paragraph):

Given that studies from Chinese psychological journals may have different target populations as compared to international collaborative projects, we compare samples from studies that share the same target population. More specifically, only if articles from Chinese psychological journals and international collaborative projects targeted the general population (inferred or stated Chinese population or humans), we will compare their sample characteristics. In the same vein, samples from other shared target populations by both Chinese psychological journal articles and international collaborative projects, e.g., adolescents, will also be compared.

I agree with the authors that they can compare big-team and Chinese journal studies with the same target sample. I propose another set of analyses, comparing the big-team studies and the Chinese journal studies that lack an explicit description of their target population. In addition, the authors may compare those studies (big-team studies and Chinese journal studies combined) with the census and the family panel study.

It has been pointed out that, by not explicitly stating the target population, psychologists sometimes implicitly assume that their findings generalize to humans in general (Cheon et al., 2020; Kahalon et al., 2021). Thus I suppose that it is legitimate to assume that those studies, the studies that do not explicitly state the target population, should have collected representative samples that are representative of the general population. Of course, I do not expect it to be the case.

Response 4: For the analyses you proposed above, we will compare those studies (big-team studies and Chinese journal studies combined) with census data and other available data (see page 12, paragraph 3):

The second question of this study is whether all Chinese samples data available, regardless of the sources of the data (see Figure 1), come from a very narrow slice of the Chinese population. Given sample representativeness indeed depend on the target population, we will further distinguish two types of analyses. For studies that targeted the general population, inferred or stated, we will compare their sample characteristics to the whole census data from the National Bureau of

Statistics. For studies that targeted a specific population, we will compare the sample characteristics to that specific population selected from census data. If the information of that specific population is not available in census data, we will search for other reliable data sources as the reference data.

Updating the coding manual

I like the way the authors plan to revise and update their coding manual as below.

In the pre-coding stage, we first developed the initial version code manual based on the previous study (Arnett, 2008; Nielsen et al., 2017; Pollet & Saxton, 2019; Rad et al., 2018). Then, at least two coders will code ten random articles independently, they will compare the results, resolve the differences and revise the manual. After that, they will code another ten articles and compare the results and revise the coding manual again. This procedure will iterate until the disagreement between two coders is negligible. (page 7)

I have some minor questions regarding the coding manual1 as below. I am confident that these will be addressed with the revision process.

1. The definition of "convenience sampling" in the sampling method section. Does it include crowdsourcing? Compared to standard random sampling, crowdsourcing is easier and more convenient. On the other hand, some crowdsourcing services such as prolific.co provide a "representative sample" that matches a nation's population characteristics. While not random, they are expected to be more representative than the traditional convenience sampling via undergraduate psychology classes.

Response 5: We have two items related to the source of participants: sampling method and recruitment method. The first one corresponds to sampling methods in methodological textbooks, including convenience sampling, random sampling, and others methods (see https://en.wikipedia.org/wiki/Sampling_(statistics) for an explanation). For crowdsourcing, it is also possible to combine different sampling methods by setting restrictions, e.g., sex, age, region, and education. Thus, the sampling methods will be coded independently of the sampling methods.

For the "Participant Recruitment Method", we will code how participants are recruited. In this item, crowdsourcing will be coded by the platform they use, e.g. "MTurk", "prolific" or "crowdsourcing" if no platform name is available.

2. The classification of educational attainment. The classification consists of 1) lower than college and 2) college or higher. I wonder how coders should classify undergraduate students with the manual.

Response 6: We are aware of the variability of descriptions of the subjects' educational attainment (see another ongoing project of our team: https://osf.io/hwtxq).

Now, our strategy is to extract words that describe the participants' educational attainment in each article (if available). After that, we will try to have a secondary coding so that educational attainment data from different studies will use the same coding system.

3. The classification of sample type. The authors provide only four classes; university students, students but not university students, infants and toddlers, and preschool children. Are they sufficient to cover the Chinese psychology participants?

Response 7: We have added one more option for this item: Adults who are not students. Please see our updated coding manual. Please note that the current form of the manual is a preliminary one and the classification maybe change in the subsequent coding. To ensure rigor, we will also extract sentences or words from papers that are associated with sample type.

Reviewed by Patrick Forscher

I think the authors have done a thorough and admirable job addressing my comments. I only have two remaining (potential) and concerns.

First, I still wonder whether we should expect research samples to exactly represent the general population from which they were drawn. To take an extreme example, let's imagine that a research community goes through a period of doing lots of research on anxiety. One would expect that research field to include lots of highly anxious participants – more so than one would expect in the general population. But this lack of representativeness is intended and, I think, justified because it is necessary to achieve the researchers' goals. If this research community stays fixated on anxiety for an extended period of time, one could critique that community for being too focused on one topic at the expense of other valuable topics that are relevant to non-anxious people, but I think periods of focus on one topic can be justified.

To be fair to the authors, they have included some codes of the researchers' intended generalization -- but I think the findings will need to be interpreted carefully with the relationship between samples, populations, and research goals in mind. So, I don't see a strong need for revision -- this is just something to keep in mind for the discussion section (with, perhaps, a few tweaks to the framing of the paper's goals).

Response 1: We agree that the representativeness of samples depends on the research question. We will interpret carefully the relationship between research aims, populations, and research samples and take it as the limitation of our study to discuss.

Second, I do have some lingering concerns about the analysis plan. One part of this concern is linked to my comments about whether one expects exact representativeness in research samples, as this expectation will be encoded in the prior. I just wonder whether the Bayes factors are comparing the right models. Maybe they are, as long as the discussion section contextualizes the results appropriately (ie it makes clear that sampling decisions are or should be a function of research goals) -- so perhaps no action is needed on this point. My other concern about the analysis plan is that it may need to be critiqued by someone with more Bayesian expertise than either I or the other reviewer can provide. I think this is an issue for the editor to decide.

Response 2: Yes, now we have a methodological reviewer, Dr. Dienes, to comment on our method and we had make revisions based on Dr. Dienes's comments.

At any rate, I think this is an interesting and valuable project and I'm looking forward to seeing where the authors go with it.

Response 3: thanks for your kind words.

I sign all my reviews,

Patrick S. Forscher

Research Lead, Busara Center for Behavioral Economics

patrick.forscher@busaracenter.org

Reviewed by Zoltan Dienes

I will comment just on the choice of analysis.

p 10: "and the Ha is not H0" and also footnote 2 "for others, the Bayesian hypothesis testing can be done without specifying the alternative hypothesis"

In fact, a Bayes factor always requires a specification of H1 because one has to calculate the probability of the data given H1, and this can only be done if H1 is some particular distribution. Where it seems not to be done, e.g. in the Hoijtink reference, it is done implicitly; and in the current case of a default, there is an explicit distribution, it is just that it is chosen without reference to the specific scientific problem. In this case, the authors themselves claim there is a distribution for H1, so the statements cited made above should be deleted. However, I think the model of H1 used as a default by the authors is not exactly equal fixed probabilities in each cell, as might be read from their description. Rather it is the distribution of probabilities in each cell is the same. What the authors need to do is say what this distribution is, and briefly justify its relevance.

Response 1: Thanks again for reviewing the Bayes factor part of our manuscript. We have re-read materials related to Bayesian multinomial test, especially Sarafoglou et al (2020), to get a deeper understanding of this test. As you pointed out, H_1 is necessary for calculating BF and the H_1 is not a fixed probability but a distribution of the probabilities in each cell. In the current case, it is the Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, i.e., $Dir(\alpha_1, \alpha_2, \ldots, \alpha_k)$. This distribution specifies distributions of probabilities in each cell.

When testing hypothesis, the H_0 is a set of fixed probabilities for each cell. This fixed value can be viewed as a point in Dirichlet distribution. That is, H_0 is a point null hypothesis. H_1, on the other hand, is a distribution of probabilities for each cell except H_0. Thus, here H_1 is an encompassing alternative hypothesis and is often noted as H_e.

The calculation of the BF_{0e} is then simplified by using Savage-Dickey likelihood ratio, i.e., the likelihood of the point under prior distribution (which is a Dirichlet distribution that is defined by prior) and the likelihood of the point under posterior distribution (which is also a Dirichlet distribution, but updated by the data).

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E.-J., & Marsman, M. (2020, August 19). Evaluating Multinomial Order Restrictions with Bridge Sampling. PsyArXiv. https://doi.org/10.31234/osf.io/bux7p

Incidentally, to see that there is a distribution involved, try in JASP "Bayesian multinomial test" which I think is what the authors are using, if one specifies the same expected counts as the counts for the prior (model of H1), the Bayes factor is not 1. That is because the prior/model of H1 uses a distribution of probabilities in each cell.

Response 2: Yes, we use Dirichlet distribution as the hyper-parameter of the multinomial distribution, the prior here is distribution for the parameters of Dirichlet instead of the multinomial distribution. With a prior distribution $Dir(\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_k = 1)$, we meant that all combinations of probabilities for categories of a multinomial distribution are equally possible.

In terms of justifying their model, the authors can show what Bayes factors are obtained for different deviations from expected proportions. This will indicate what size deviations their analysis is sensitive to, given their Ns. They should do this in order to show the severity of their tests: Is it likely that the tests will find evidence against their hypotheses, given reasonable assumptions about what size deviations there might be?

Response 3: In our case, we use the expected proportion as the point null hypothesis, and the BF_{0a} represents the ratio that the likelihood of this point given the prior and the likelihood of this point given the posterior (which reflects the effect of the data).

For sensitivity analysis, we did not conduct simulation directly based on the "effect size", i.e., the distance between two points in the Dirichlet distribution. Instead, we now use two different priors for calculating the Bayes factor. In addition to the default "non-informative" prior, we also choose another stronger prior, which is the proportions are equal to the expected (e.g., the proportion from census data when testing our second hypothesis). Under this prior and our N (sum up to 100 because we are using percentage), it's much harder to gain evidence to support H_0.
When applying multinomial test to age bins, the number of bins also matters. In this version, we choose two different ways to create the ages bins, one is based on psychological science, and the other is more based on the census data's age bins. We reported Bayes factor from the former data in the main text and the latter in the supplementary.

The Design Table needs to be more specific. List each hypothesis that will be tested, giving the exact test, and stating under what conditions the hypothesis will be deemed supported or refuted (e.g. what BF threshold).

Response 4: We have revised the Design table by adding more specific thresholds for supporting or refuting each hypothesis.