Editor summary (Yuki Yamada)

This report merits a revision

Thank you very much for granting PCI-RR the opportunity to peer-review your paper.
At the same time, I sincerely apologize for the long delay in responding to you.

This manuscript was reviewed by two very experienced researchers who are interested in the WEIRD issue. Frankly speaking, the purpose of this study is favorably viewed by the reviewers as well as by myself, and it is desirable that this study be properly carried out.

This will require careful elaboration by a major revision, especially with respect to sample representativeness, coding, and sampling methods, as the reviewers have pointed out.

One reviewer also raised concerns about the placement of hypotheses and the setting of prior distributions in the Bayes factor analysis. It would be good to have this point clarified, but if necessary, please let us know that we can ask an expert in Bayesian statistics to check this point as an additional reviewer. In that case, we will ask them to focus the scope of their review on this point, so we do not expect it to take as long as it has so far.

I am looking forward to your revised manuscript.

Yuki Yamada

Dear Prof Yamada,

We appreciate your invitation to revise our manuscript. We are also grateful for constructive suggestions of two reviewers, which helps us to improve the quality of our protocol.

We have addressed reviewers' points carefully, please see detailed point-by-point response below and the changed traces in our manuscript. Also, we correct one error in Fig 3 B in the previous version. Also, we updated our census data to the 7th census data of China, which was made public recently.

As Reviewer 1 questioned the Bayes factor method and either we or the reviewer are confident about the method, we'd like to further recommend the following potential reviewers to review our method section to ensure the appropriateness of our method:

Dr. Zoltan Dienes, University of Sussex, email: zoltan.dienes@googlemail.com;
Dr. Herbert Hoijtink, Utrecht University, email: h.hoijtink@uu.nl;
Dr. Xin Gu, East China Normal University, email: guxin57@hotmail.com;
Dr. Eric-Jan Wagenmakers, University of Amsterdam, email: E.M.Wagenmakers@uva.nl;
Dr. Alexander Ly, University of Amsterdam, email: a.ly@uva.nl;

Dr. Richard D. Moery, Cardiff University, email: moreyr@cardiff.ac.uk;
Dr. Tahira Jamil, King Abdullah University of Science and Technology, email: tahira.jamil@kaust.edu.sa;
Dr. Jeffrey N. Rouder, University of California, email: jrouder@uci.edu

We feel that the review process prompted substantial improvements to 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

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 population 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).

Response 1: We appreciate the reviewer's detailed comments about "representativeness of sample" and how it is related to our coding schema.

We realized that defining "representativeness of a sample" or a "representative sample" itself is far more complicated than we had (naively) expected. There are at least six meanings of a "representative sample" (Kruskal & Mosteller, 1979a, 1979b, 1979c, 1980; Kukull & Ganguli, 2012): 1) "general, unjustified acclaim for the data"; 2) "absence (or presence) of selective

forces"; 3) "mirror or miniature of the population"; 4) "typical or ideal case … that represents it (the population) on average"; 5) "coverage of the population … (sample) containing at least one item from each stratum …"; 6) "a vague term to be made precise" by specification of a particular statistical sampling scheme, e.g., simple random sampling.

Among these, the core of the scientific meanings of representative sample is "miniature of the (target) population" and "absence of personal bias", which was firstly proposed by Norwegian statistician Anders Nicolai Kiær at the 1895 Berne meeting of the International Statistical Institute. Later, Neyman (1934) laid a theoretical foundation for random sampling, which put the randomization at the core of "rational" design. In the framework of Neyman, representative sample means samples that follows the random sampling design. However, as pointed by Zhao (2021), random sampling can not eliminate the chance bias and it is extremely difficult to achieve in reality. **In short, a representative sample should be "lack of bias", or a "miniature of the target population", or randomly sampled from the target population (even it is hard to achieve in reality)**.

As we are surveying a relative large sample of Chinese psychological studies (1000 Chinese papers & Chinese sample in a number of big-team science projects), we assume that the representativeness of these samples together should be a miniature of their targeted population.

We added a sentence to define "representativeness" in the current study in the introduction section (page 4, paragraph 2):

"*By the word "represent", we mean the sample in a study (or studies) should be a miniature of the targeted population without selection biases, or theoretically can be a miniature of the targeted population and without selection bias (Kruskal & Mosteller, 1979a, 1979b, 1979c, 1980; Kukull & Ganguli, 2012).*"

Secondly, we added the three items related to target population (or "generality statement") in our coding manual. The first item is the target population to which the study is intended to generalize. The second item is the level of explicitness on the statement about the target population (explicitly stated or inferred from the text). The third item is the exact sentences/words excerpted from the full text of the paper that are associated with the statement about the target population. Coder will instruct to search the statement of target population in conclusion. If no related information was found, they will search information in other parts of articles (firstly introduction, and then, results or other parts). These three items will code both the targeted population but also keep transparency of the coding process. However, we welcome to any further suggestion on coding target population (or generality).

Thirdly, when it comes to interpretation of "the results if we find the distributions differ between the samples of papers in Chinese journals and the samples of international collaborative projects", as the reviewer has mentioned, coding the target population (generality statement) will help us to address this issue. After coding the target population, we will only compare the samples between Chinese journals and big-team science projects if they are intended to generalize to the same

(sub-)population. We will encode the sampling plan and recruitment approach described in the papers, these information may help us understand why the sample are as they are now.

Hypotheses in Bayes Factor analysis

Let me first state that I am not familiar with Bayesian 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 Bayesian 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 (international 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 Chinese populations" and we will test it with the data (actual characteristics 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 $P = Pr(x1, x2, \ldots x7 \mid n = 100, p1, p2, \ldots p7)$ 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 setting 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 Bayesian 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 Bayesian analysis.

Response 2: Thanks for pointing this out. This comment includes two important points about the Bayes factor method we are using. The first concern is about the choice of the alternative hypothesis. The second concern is about the prior, is a flat prior made of ones really uninformative.

Before answering these two questions, we have to admit that we are not experts on Bayesian statistics either. However, we are trying to Bayes factor because it has a great advantage: it can both support and against the null hypothesis.

As for the first concern about the alternative hypothesis, we realized that the method we are choosing actually has an unconstrained alternative hypothesis, which means that we did not choose a specific distribution with specific set of parameters. Firstly, we used R code for calculating Bayes factor for the Bayesian multinomial test, which is the same as in JASP. Thus, the null hypothesis and alternative hypothesis should the same as in JASP. In the section about the Bayesian multinomial test (page 51) in the tutorial documents by Goss-Sampson, van Doorn, & Wagenmakers (http://static.jasp-stats.org/Manuals/Bayesian_Guide_v0_12_2_1.pdf), it reads as below:

"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. This is also reflected in the code: `log(BF) = (lbeta.xa-lbeta.a)` where lbeta.xa is the posterior and lbeta.a is the prior.

Secondly, in our previous communication with editor, Dr. Dienes, we stated that the alternative

was also specified as an equal probabilities for different bins. It seems that we were wrong regarding the meaning of the alternative hypothesis. We are sorry for this mistake.

Thirdly, the approach used in JASP is similar to the null hypothesis significance testing and is not recommended in general. However, in our case, we have very specific null hypotheses and providing evidence for or against these null hypotheses suits our research purpose. For example, for the first question, our null hypothesis is "*observed data are sampled from the population with parameter equals to that of Chinese samples from international collaborative projects*", refuting or confirming this null hypothesis will answer our question. The exact parameters of samples from Chinese journals or international collaborative projects can be examined further.

Fourthly, previous methodological papers suggested that using an unconstrained alternative hypothesis is a valid approach too (e.g., page 541 ~ 542 in Hoijtink et al.., 2019). Following the notation in Hoijtink et al. (2019), we have changed the notation of the alternative hypothesis as H_a, instead of H_1, and changed the meaning of H_a as "not H_0".

As for the second concern about prior, we are sure that a flat prior of ones are weak and non-informative prior in Bayesian multinomial test. The code for calculating the posterior in multinomial test is as below:

```
lbeta.xa <- sum(lgamma(alphas + counts)) - lgamma(sum(alphas + counts))
```

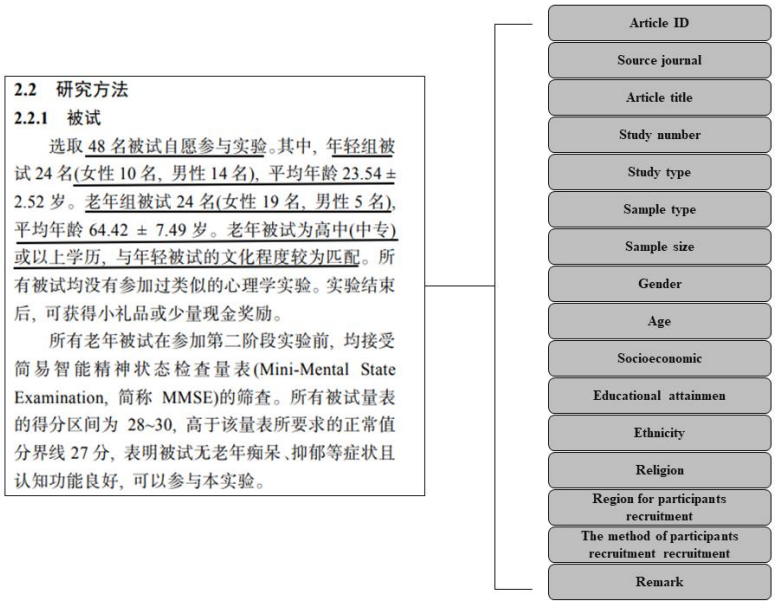
Where `alphas` are priors, which should be integers; `lgamma` is log gamma function to calculate the PMF of the multinomial distribution; `counts` is the observed data. From code we can see that if alphas are ones, then the posterior will be largely determined by the `counts`, the observed data. If we use larger priors, the observed data, `count`, will play a smaller role in the final posterior. Please see our online code for details on calculating the BF.

As we said, we are trying to use the best methods that are available to us, it'll be great if the editor can invite Bayesian statistician to further inspect our Bayes factor part. We have provided a list of experts on this.

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

Response 3: Thanks for pointing this out. We have reconsidered this situation and revised our coding manual. In our new coding manual, if multiple groups of samples were reported, we will code them separately, and all demographic information of each group will be recorded in detail.

Coding of the generality conclusion

This is related to the points I have mention 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 detailed 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)

Response 4: As we mentioned above, we will code the target population (generality statement) in the revised codebook. However, representativeness has not been taken seriously in the field (see, Thalmayer et al., 2021). This situation is similar to causality in psychology (Grosz et al., 2020): researchers may use vague statements about representativeness or generality. Our initial coding, based on a few papers from the same 5 journals but not the final sample of papers, suggested that it is difficult to code the target population. To make the coding task more doable, we added the three items related to target population (or generality) in our coding manual. The first item is the target population to which the study intended to generalize. The second item is the level of explicitness on the statement about the target population (explicitly stated or inferred from the text). The third item is the exact sentences/words excerpted from the full text of the paper that are associated with the statement about the target population. Coder will first search statement about these three items in conclusion. If no such information was found, coders will then search other parts of articles. These three items will both code the targeted population and keep transparency of the coding process.

Reviewed by Patrick Forscher

The authors propose to assess the representativeness of participants in Chinese psychology research. To this end, they propose to compare samples from five different sources:

1. Samples from five mainstream Chinese journals
2. Chinese samples from large-scale international collaborations
3. Non-Chinese samples from large-scale international collaborations
4. The National Bureau of Statistics of China
5. The Chinese Family Panel Study

They pursue their goal of assessing the representativeness of participants in research in China with five activities, which use the samples illustrated in 1-5:

1. Compare samples from Chinese journals (1) to Chinese samples in large-scale collaborations (2);
2. Compare samples from Chinese journals (1) to census data (4 and 5);
3. Compare Chinese samples from international collaborations (2) to non-Chinese samples in international collaborations (3)

I love the concept of this project. Psychology has long paid too little attention to sampling. Most of the time this problem gets described under the umbrella of the "WEIRD problem", but it can be construed more broadly as a "generalizability problem". The problem can even be construed more deeply as an issue with who defines the samples and topics that are interesting to study and how we draw conclusions about those samples and topics. I think this project could advance our understanding of these kinds of problems.

Response 1: Thanks a lot for your kind words.

I don't have many specific problems with the proposed study – instead, I have some suggestions for the authors to consider. Some of these suggestions may broaden the scope of the research. I think it would be fine if the authors declined some of these – so consider these suggestions as possibilities that the editor and authors can think about together as the protocol is revised.

My comments are divided into four sections:

1. Broad aims
2. Coded characteristics
3. Data sources
4. A note on the analysis plan

Broad aims

Although I love the topic of the project, I could imagine a skeptic wondering whether anyone would expect Chinese samples to perfectly represent the Chinese population. Researchers are supposed to choose the sampling methods that allow them to accomplish their research goals. Sometimes this involves random sampling to accomplish representativeness, but sometimes it doesn't – as is the case when researchers simply want to show that a psychological phenomenon exists in any population at all. This defense is actually the very one offered by Mook (1983) in response to early claims that psychology has a generalizability problem (https://www.vanderbilt.edu/psychological_sciences/graduate/programs/quantitative-methods/quantitative-content/mook_1983.pdf).

However, there are some powerful responses to Mook's argument:

- Although some of psychology research involves existence proofs, many research topics require going beyond existence proofs. This is especially the case with research that has applied aspirations – it doesn't really matter if you can get something to work in the lab if it doesn't work in the real world.

- If researchers focus too much on the experiences and concerns of a narrow sub-population, they will miss phenomena that are experienced outside of that sub-population (see https://osf.io/preprints/africarxiv/xd269/). I like to think of these missed phenomena as "unknown unknowns" – research psychologists can't even know that they are missing them because their measures and datasets don't include the necessary information to know this.

- Researchers choose research priorities based on their own experiences. If researchers are also drawn from a narrow subpopulation, they will choose research priorities that are important to that subpopulation, creating a distorted view of human psychology (see https://www.proquest.com/docview/527775905?pq-origsite=gscholar&fromopenview=true)

I think the authors should consider Mook's arguments and the responses to it. Doing so might inform the aims and design of this study, as well as the information that is coded from each data source (see my next point, below).

Response 2: Thanks for pointing this out! Now we have integrated Mook's argument and counter points in the manuscript in the first footnote in the introduction (see page 3, paragraph 1). Here is the added footnote:

While it is generally accepted that samples should be representative to the target population, Mook (1983) argued that generalization maybe misplace in some cases where showing some effects do exist, even in rare and artificial settings, is valuable. This argument is invalid because most psychological research aims higher than mere existence of certain effects (e.g., guide the policies, IJzerman et al., 2020). Also, focusing on a narrow sub-population, we may miss phenomena that are outside that sub-population and the consequence of these missed phenomena is unknown. Finally, the selection of samples reflects the fact that researchers themselves are from narrow sub-population, they may priorities the phenomena that are important to that sub-

population and thus distort the whole picture of psychology.

Coded characteristics

The characteristics that are coded should be selected to accomplish the project's broad aims. Because I think these broad aims might need a bit of adjustment, and because the specific adjustments should be decided by the authors, I won't be too prescriptive with my suggestions about what to code. However, I do have a few thoughts that will, hopefully, help the authors think through what sorts of characteristics to select.

If the authors want to show that research in China is too focused on a specific research aim, such as the existence proof, they might consider coding some characteristics that capture the match between aim and sampling method. This might include, for example, the type of sampling the authors implemented (convenience, online panel, probability, etc), the type of research (exploratory or confirmatory), and/or the setting (lab, field, online, etc). They might find some ideas of what to code in this article on Arabic social psychology by Saab and colleagues (2020; https://journals.sagepub.com/doi/abs/10.1177/1948550620925224).

If the authors want to capture the types of topics the authors select (and maybe, compare the topics in Chinese language journals to those in big team science initiatives), it might be worth coding something about the broad topic of study. Ideally, this would use a pre-existing coding system (such as the article keywords) to lower burden on the coders. This was a focus in a commentary I co-wrote on African psychology (https://osf.io/preprints/africarxiv/xd269/); we didn't do a systematic coding of topics, but instead tried to give a holistic sense of how African priorities might differ from Western priorities.

If the authors want to assess who's setting the research priorities, they might want to code where the lead authors of each article are from and/or what their background is. This is an approach taken in Thalmeyer and colleagues (2021; https://serval.unil.ch/resource/serval:BIB_38DE994E17E6.P001/REF) – a more recent update to Arnett (2008) that the authors might find useful to scan.

Response 3: Thanks for these valuable suggestions. Our primary goal is to examine the "match between aim and sampling method". The coding manual was revised to suit this purpose better. (See also our response to Reviewer 1's response 1 (see page 5, paragraph 6 of this document). Additionally, we have added new coded characteristics: sampling method (Unreported/Convenience sampling/Random sampling/Others).

For the other two interesting issues related to research topics, we plan to export the keywords of Chinese journals' articles and the big team projects through CNKI (China National Knowledge Infrastructure) and Web of Science, and then use bibliometric methods to compare the similarities and differences between the keywords of the big team projects and Chinese journals' articles. However, given that we will only include 1000 papers in the current research, the bibliometric

analysis will be preliminary. To fully explore the research topics and examine the common and distinct trends between psychological science in China and other regions requires studies that primarily focus on the research topics in Chinese psychology and psychological science in other regions (e.g., North America, Europe).

Another similar possibility is to code the abstracts of the papers the authors sample for whether the source of the sample is mentioned, which could tell the authors who researchers take as the implicit "default participant". This is an approach taken by Kahalon and colleagues (2021; https://journals.sagepub.com/doi/10.1177/19485506211024036).

Response 4: Thank you for your suggestion. We have revised our coding manual. Specifically, for the abstract part, we have added an item to record whether articles mentioned the participants' demographic information (mentioned vs not mentioned). Because of the prevalence of undergraduate samples in psychology research, we will code distinguish studies that relied only on college students' samples and studies used other samples. We will then compare the percentage of mentioned and not mentioned in two groups of studies, see below for the template of our table 2.

Table 2. Different study types and their sample mentions.

Study type	Samples mentioned	Sample not mentioned	Total
Only college students			
College students & other populations			
Only sample outside colleges			
Total			

Data sources

I two brief notes on the data sources the authors have chosen.

Chinese journals. I must admit to ignorance as to the landscape of Chinese-language psychology journals, so I can't really evaluate whether the five Chinese-language journals are a good representation of this landscape. For the benefit of readers like me, can the authors provide some description of how these journals were chosen – and maybe, of the landscape of Chinese journals generally?

Response 5: Thank you for your valuable feedback. We have revised our manuscript as below (see page 5, last paragraph).

These journals are chosen because the following reasons. First, these five journals are indexed by CSSCI (Chinese Social Sciences Citation Index), which is regarded as authoritative and

comprehensive database for bibliometric studies of China's social sciences (e.g., Gong & Cheng, 2022). Thus, all these five journals are selected as of high-quality among all Chinese psychological journals. Second, these five journals cover most fields of psychology. Among them, Acta Psychological Sinica, Journal of Psychological Science, and Psychological and Behavioral Studies are comprehensive journals, studies from all sub-fields of psychology are included; Psychological Development and Education is the only journal for developmental and educational psychology in China; Chinese Journal of Clinical Psychology focuses studies in clinical psychology and mental health.

Big team science initiatives. The landscape of ManyLabs-style initiatives (or, as I like to call them, "big team science" initiatives; see https://psyarxiv.com/2mdxh/) has grown a lot since the first ManyLabs studies. If you need a list of possible data sources for this style of study, you might want to consult this spreadsheet (https://docs.google.com/spreadsheets/d/1BUURnm0CvwubyYJSp_Yfj0ntWLCHHOl_TDXsQgkSN1Q/edit#gid=0), which is compiled and maintained by Dwayne Lieck and Daniel Lakens
A note on the analysis plan

Response 6: We appreciate that you provided these data sources. We have integrated these into the list we curated ourselves, please see the list we have now: https://docs.google.com/spreadsheets/d/18lF1KohchjKR5nM2utvlnd9A-zc3iKyf4EXLJ_jMPgk/edit#gid=1741341307

The proposed analysis is very detailed and uses Bayesian methods that I don't feel qualified to review in detail. However, I felt generally that the specific analyses may be too focused on evaluating whether Chinese samples are "exactly representative" of the Chinese population. I would advise more thought on the broad goals of the research and the characteristics that need to be coded to achieve those broad goals, then revising the analysis plan.

Response 7: Thanks for pointing this out. In this revised version, we also coded the targeted population of each study, thus, the comparison will be between samples and the targeted population. We added this detail in our "data analyses" section (see paragraph 5, page 10):

"Given that studies from Chinese psychological journals may have different target populations as compared to international collaborative projects, we will only select sample data from those studies share the same target population and conduct the statistical inference. For example, one cluster of studies may target the same population: all Chinese adults, we will then compare their sample with the Chinese adults' demographical data from Census data from the nearest year. If another cluster of studies targeted adolescents, for example, we will then compare the sample to that of adolescents in Census data. Meanwhile, if possible, we will further explore potential reasons why studies aimed at the same population used different samples."

Conclusion

I love the topic of this proposal and want to see the finished product. I don't have strong views on the direction authors ought to take the protocol – though I do think they might benefit from reflecting a bit on the project's broad aims. This would give them the opportunity to sharpen the specific goals and research activities so that their project is as impactful as possible.

I sign all my reviews,
Patrick S. Forscher

Research Lead, Busara Center for Behavioral Economics
patrick.forscher@busaracenter.org

(PS: I noticed a few minor English usage mistakes. These didn't factor into my evaluation at all, which is why I am writing this note at the bottom. However, if the authors want someone to do some quick copy edits whenever they're looking to submit this to a journal, I'd be willing to help them with this)

Response 8: Thank you for pointing this out. We have examined the language of the text carefully. If we need help later, we will contact you.