Revision needed

Again, the reviewer commented. I think the discussion is getting quite mature, but there are still some insufficiencies in the in-text explanation and rationale of the justification for priors. In a normal peer review, I would prefer not to go back and forth too many times because it consumes both parties' resources, but I think it is important that key points are agreed upon between the authors and the reviewer, and I would like to continue for another round. I hope you will find those points of agreement, taking into account the past comments of this reviewer.

Yuki Yamada, Recommender

Dear Prof Yamada,

Thanks again to the reviewer and the editor. We have addressed Dr. Dienes's comments carefully, please see detailed point-by-point response below and the changed traces in our manuscript. We updated the data from Jones et al. (2021) after re-checking the data so that the number of participants used in our analysis is consistent with that reported by Jones et al. (2021).

Sincerely,

Hu Chuan-Peng,
Lei Yue

Reviews
Reviewed by Zoltan Dienes, 18 Feb 2023 09:15

The authors have made a major step in addressing my point. (Incidentally I didn't find the supplementary materials, so I don't comment on them.) I think however the discussion of this point should be in the main text when they introduce the prior (model of H1) that they use and they need to elaborate more to justify the scientific relevance of this model of H1 given these results. In effect if their reference sample had 50% males, their comparison would need about 67% or more males (or 33% or less) to be detected as different -is this reasonable in this context? (The model of H1 presumes any difference in proportion is as probable as any other - and one consequence of that vague assumption is a large difference is needed to detect any difference.) I am not sure I find that reasonable. (That is why I always use scientifically informed priors.) Can the authors argue for a reasonable prior or argue that this prior is reasonable *given their particular scientific/empirical context*?

Response 1: Thanks a lot for this suggestion, this is very helpful. We now calibrate our Bayesian method based on the smallest effect size of interest, which is estimated based on data from the literature.

Firstly, we estimated the smallest effect size of interest based on previous studies. We did it in the following steps:

Step 1, we extracted data from Rad et al (2018)'s supplementary material and searched for articles that reported sex ratio. We found 21 papers published in *Psychological Science* in 2014, which include 35 studies. One study surveyed households instead of individuals, thus the sex ratio is not available. The final articles are 20 and the studies are 33.

Step 2, we extracted the sex ratio of participants in these 33 studies and calculated the absolute deviations from a balanced ratio (0.5).

Step 3, we calculated the mean deviation and its 95% confidence intervals using bootstrap (bootES package in R). The results revealed that the mean deviation from 0.5 is 0.085, 95% CI [0.061 0.115].

Step 4, we used a value slightly smaller than the lower limit of the confidence intervals, i.e., a difference of 0.06 from $p = 0.5$, as the smallest effect size of interest.

Secondly, we used two approaches for simulation with the smallest effect size of interest.

The first strategy is fixed the prior as non-informative and varying the N. Our simulation revealed that when $N >= 1200$, we have 90% chance to detect the deviation of 0.06 with BF >= 6. Also, we have 90% chance to support the H_0 if there is no effect with $N >= 1200$.

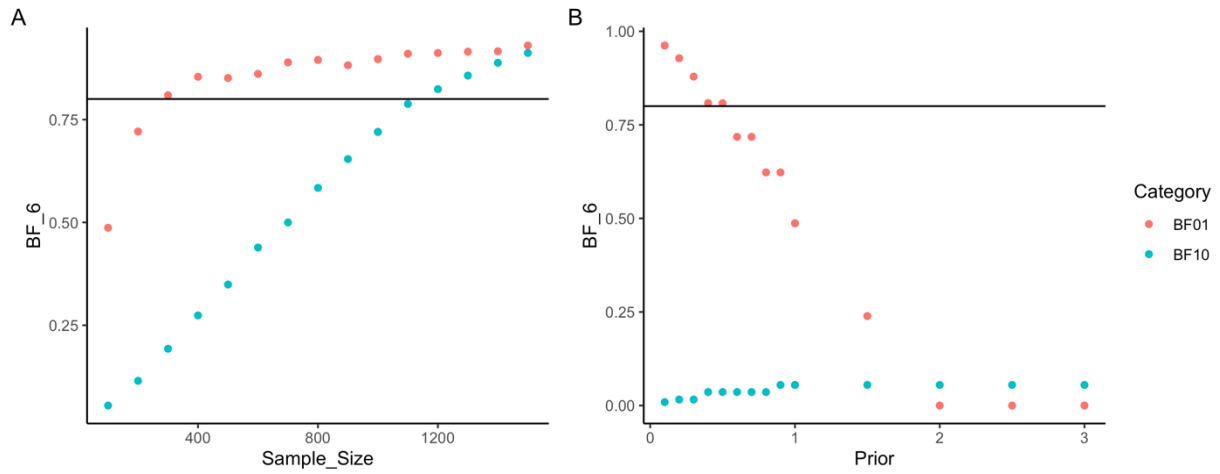

Fig R1. Simulation results. (A), Simulation with varying *N* and fixed prior, the x-axis is varying *N*, and the y-axis is the proportion of BF values that are greater than 6. (B) Simulation with varying prior and fixed *N*, the x-axis is the varying prior, and the y-axis is the proportion of BF values that are greater than 6.

The second strategy is fixed the *N* to 100 (i.e., stick to percentage data) but vary the prior. This strategy allows us to keep using percentage data as in our previous versions. The results revealed that when the prior decreased the probability of parameters of the binomial distribution round $p = 0.5$ by decreasing the prior from [1, 1] to [0.5, 0.5] and even smaller, the chance of getting evidence for H_0 increased but the chance of getting evidence for the H_1 decreased. These results suggest that we can not get enough power to support H_1 by varying prior in this way.

Thus, we decided to use the non-informative prior but revise our analytical plan by directly using the number of participants from data instead of percentage data ($N = 100$), because the latter will not provide enough statistical power. Using the number of participants from the data means that the final *N* will be large enough (>=1200). For large-team studies, participating teams were usually asked to collect more than 100 participants per team, which means with twelve big-team studies, the Chinese participants will be more than 1200. So far, we have collected 21 big-team studies which included Chinese participants and the total *N* of Chinese participants will be larger than 1200. For studies published in Chinese journals, we will extract data from 1000 empirical studies, which means if each paper only contains one study, and each study only has 20 participants, we will have 2000 participants.

We have added our justification of our revised analytical strategy in the main text, please see page 11, paragraph 4, and page 12, paragraph 1.

Again, we really appreciate Dr. Dienes's comment, which helped us to find a more appropriate setting for our Bayesian factor analyses.

For the age result, the average reader will not be able to interpret what this means in terms of what age differences could be detected. That is why I suggested modeling age as a normal and

showing what diference in mean ages could be detected. Find a way to present the sensitivity of the method to age differences that is intuitively graspable.

Response 2: We agree that it is difficult to interpret the "effect" in the multinomial test for average readers. Fortunately, the figures, e.g., figure 3B and figure 4D, are intuitive for readers to infer the difference between the distribution across different age bins.

We tried to follow Dr. Dienes's suggestion, i.e. modeling age as a normal distribution and testing the difference in mean age, e.g., using *t*-test. Unfortunately, we find that this approach does not fit the purpose for comparing two distributions. To our knowledge, independent samples *t*-tests fail to detect the difference between two datasets that have different age distributions but with similar mean ages. In this situation, (1) the assumptions of performing *t*-test are not met because the two datasets do not have similar *SD*; (2) the false negative rate is higher for the *t*-test.

We simulated one such case in our R Notebook, following the steps below:

Step 1: we generate multinormal distributed data with two different parameter vectors: [0.05, 0.45, 0.35, 0.1, 0.05], [0.45, 0.05, 0.1, 0.35, 0.05]. Each parameter represents the probability of age bins as in our manuscript: 0 ~ 17 (children and adolescents), 18 ~ 25 (early adulthood), 26 ~ 40 (middle adulthood), 41 ~ 59 (later adulthood), and >= 60 (elders). We used sample size N = 1200 so that it is consistent with our previous simulations.

Step 2: we generated age data by the sample from each age bin (using `runif`). This step generated age of 1200 participants for both multinomial parameters.

Step 3: we then compared the age data from the above two sources with independent sample *t*-test, and record the *p*-value, if the *p*-value is smaller than 0.05, we regard it as statistically significant.

We repeated the above step 1000 times and found that the proportion of significant results is about 0.25, even though the age distribution is different.

We also calculated the BF using the Bayesian Multinomial test, using data from Step 1, treating multinormal data from one parameter vector as observed and the other as expected, we also recorded the number of BF values that were greater than 6. The results revealed that the proportion of BF values that are greater than 6 is almost 100%.

Based on this simulation and our reasoning above, we infer that the Bayesian multinomial test fits our research purpose better.

Again, we thank Dr. Dienes for pointing out a potential alternative. We hope that figure 3B and figure 4D will help readers to understand the difference in age distributions, and thus address Dr. Dienes's concern.

Finally, all the scripts for the simulation, plotting, and preliminary analysis can be found at (https://osf.io/avb7t/?view_only=a7e4610491374093851fc2b7da57e85c) or GitHub (https://github.com/hcp4715/chin-subj).