Review Reports (doi.org/10.31234/osf.io/xky4j)

#Summary

It is important to examine how each domain (visual/auditory) influences evaluating performance in music. The previous study has demonstrated that visual information can play a stronger role than auditory information in some cases. The present study will examine these issues by the use of different types of stimuli and participants. Based on the results, the authors will discuss the generalizability of phenomena and the roles of the unique tradition of instrument performances.

# General comments

The authors seem to address the interesting issues. Examination of the generalizability of phenomena is also important by the use of different types of stimuli and participants. However, there are several problems, and the authors should revise these before collecting the data.

#Major points

- The reason for addressing the Tsugaru shamisen (lines 92-96) seems not to fit the protocol. If my understanding is correct, one of the reasons why the authors will address the Tsugaru shamisen is due to its unique history: It was traditionally played by blind folk musicians. Specifically, the authors seem to examine whether traditionally excluding the role of visual information might have unique influences on evaluating performance. However, is the historical background of the Tsugaru shamisen really familiar, particularly among assumed participants? From the viewpoint of addressing the unique historical influences of the Tsugaru shamisen on evaluating performance, it would be appropriate that the authors set two groups of participants (those who know the historical background of the Tsugaru shamisen vs those who do not know the historical background of the Tsugaru shamisen) and compare the results between the groups. If the authors do not focus on the unique historical influences of the Tsugaru shamisen on evaluating performance, they should revise the descriptions at lines 92-96.
- As above, if one of the authors' aims is to examine the effect of the characteristics of the Tsugaru shamisen performance tradition, they should discuss how the characteristics influence observers' evaluation at Introduction. That is, they should explain the details of the assumed mechanism.

- Is parametric tests (i.e., *t*-tests) appropriate for the data of the present study? The possible score of each participant will be only 0, 33%, 66%, or 100%. Of course, I understand that the authors might use the same statistical procedures. However, this is a conceptual replication, not direct replication. It might be better to think the ways of the data analysis deeply.
- Considering 2.3. Statistical analysis and 2.4 Power analysis, the authors set Cohen's *d* = 0.4 for calculation required sample sizes and performing equivalence testing. This seems to be because they assumed Cohen's *d* = 0.4 as SESOI. Why did they choose this? The authors seem to think that estimating the effect size before collecting the data is notoriously difficult but I guess that they can estimate them based on the previous study and their pilot data.
- It is unclear whether the study design is appropriate for testing their hypotheses. The authors built two hypotheses, where one of the keys is the degree of the variances in the trials. However, the independent variable is only the stimulus domain (Audio-only, Visual-only, and AudioVisual). How will the authors examine the hypotheses at the present design? Moreover, there are two dependent variables: The percentage of participants correctly choosing 1) the 1st-placed performer, and 2) the lowest-placed performer. It would be better to build the hypotheses for both dependent variables.
- I think that they cannot conclude that the characteristics of the Tsugaru shamisen performance tradition mediate in the results even If the authors obtain the null result. This is because the cultural background of the participants as well as the instrument will be different between the present and previous studies. It is possible that the predicted phenomena will not occur in Japanese speakers (i.e., the participants in the present study). Thus, in addition to the present experiment, it would be better to perform an experiment, where the authors use the same stimuli as those of the previous studies (i.e., solo piano competitions) and collect the data from Japanese speakers, and to confirm whether the predicted phenomena will be observed in Japanese speakers. This examination should be beneficial for testing the generalizability of sight vs. sound effects; I think that one of the present study's motivations is to investigate the generalizability of sight vs. sound effects, considering the descriptions at lines 68-71.

#Minor points
- The first paragraph of the Introduction seems to be far away from the theme of the present study. It would be better to revise them to match the theme of the present study.

- The authors should explain the details of the main previous studies (Tsay, 2013; Mehr et al., 2018) in the Introduction, not the Appendix. These are the bases of their hypotheses.
- In 2.4 Power analysis, the authors should explain the details of the power analysis (e.g., tools and type of the statistical test).