慶應義塾
Keio University

5322 Endo, Fujisawa
Kanagawa 252-0882, Japan
(+81) 80-6551-4063

December 14th, 2021

RE: https://doi.org/10.31234/osf.io/xky4j, *Sight vs. sound in the judgment of music performance: Cross-cultural evidence from classical piano and Tsugaru shamisen competitions*

Dear Dr. Yamada,

We appreciate your invitation to revise and resubmit our manuscript based on the constructive comments of the two Reviewers. We are grateful for the chance to use their constructive feedback to completely redesign the experiment to allow us to include the classical piano stimuli used in previous experiments in a unified experimental design. This is precisely the kind of fundamental redesign that would have been impossible to implement efficiently in a traditional non-Registered Report review process, so we are delighted to have chosen the Registered Report format. We have modified the subtitle from “Evidence from Tsugaru shamisen competitions in Japan” to “Cross-cultural evidence from classical piano and Tsugaru shamisen competitions” to reflect this change. We have appended a version with track changes to this response letter for your convenience.

In addition to this major redesign, we have also addressed the reviewers’ other points regarding use of more appropriate statistics and clarifying terminology for a broad audience (see detailed point-by-point response below). To allow us to fully address the statistical issues, we recruited an additional author (Yuto Ozaki) to help conduct the more complex statistical analyses.

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,

Patrick E. Savage
(on behalf of the authors)

Editor summary (Yuki Yamada)

Major Revision

I appreciate your submission to PCI RR. As you can see, we were able to receive peer reviews from two relevant researchers: one is a cognitive psychologist who is very familiar with registered reports. The other is a widely experienced expert in Japanese historical music. Before introducing the individual peer review results, I would like to inform you that this manuscript requires a major revision before it can be recommended. The reasons for this are as follows.

The former reviewer seems to acknowledge the potential significance of this work, but also points out several major issues. These may be summarized in the appropriateness of the experimental design and the justifiability of the sample size design. In particular, if this study is truly considering that historical factors (I am not sure if that is the suitable terminology) are related to the audience's performance evaluation, it should be specifically stated as a hypothesis, as pointed out by the reviewer, and the experimental design should be capable of examining it. That is, it is worth considering using a method that can detect the effects of knowledge about the historical background of the Tsugaru shamisen and about traditional performers (not the recent popular ones), and adding other popular music as an additional control condition.

We have completely redesigned the experiment and modified the rationale as requested by Reviewer 1. Note that this resulted in comparing Tsugaru shamisen music not with popular music but the Western classical music stimuli used in previous similar studies. We have modified the subtitle from "Evidence from Tsugaru shamisen competitions in Japan" to "Cross-cultural evidence from classical piano and Tsugaru shamisen competitions" to reflect this change.

The latter reviewer appreciates the article very much, but says that some technical expressions should be annotated. In fact, readers who read registered reports and are familiar with hypothesis-testing studies will have no difficulty in understanding the meaning of statistics and methodological abbreviations. However, this study has a very unique focus of research subject (i.e., Tsugaru shamisen), and the readership may be much broader than the authors envisioned. Therefore, it would be beneficial for the social impact of this study to supplement the descriptions with points even if the authors might feel they are redundant in writing usual manuscripts.However, in my opinion, adding detailed explanations of statistics in the text may reduce the readability for experts, so I thought it would be a good idea to use footnotes.

Thus, I am looking forward to receiving this manuscript again, which has been greatly improved by the review comments of both reviewers.

We have implemented this helpful suggestion in a footnote within the abstract as follows:.

> ***This Stage 1 Registered Report is a proposed protocol designed to be used for collecting full data after the initial protocol has been reviewed and approved. It includes a power analysis to determine what is a reasonable number of participants to recruit to***

> ***appropriately balance logistical feasibility against the risks of false negative and false positive results. This involves terminology that may be unfamiliar to some readers without a background in statistics (e.g., "Cohen's d"; "80% power"). For accessible introductions to Registered Reports and power analysis, see Chambers (2019) and Braebart (2019), respectively.***

Reviewer #1 (Kyoshiro Sasaki):

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.

1) 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.

We thank the reviewer for this important point. On reflection, we agree the design did not properly match the rationale. The Reviewer's suggestion of adding a group of participants who know the historical background of the Tsugaru shamisen might theoretically allow us to better match our original rationale. However, it would be infeasible to recruit enough participants who are Tsugaru shamisen players but don't already know the results of these competitions.

Furthermore, as the Reviewer suggested later, a main part of our rationale was testing the generality of sight vs. sound effects, and doing this would require re-testing the original stimuli used in previous experiments. We have therefore rewritten and reframed the manuscript to focus on the generalizability of sight vs. sound effects, not on the specific details of Tsugaru shamisen's historical background. We have also added clarification that part of our rationale is that the lead author is himself a Tsugaru shamisen national champion. The main newly framed paragraph (Lines 207-212) addressing this now reads as follows:

> *1.1 Study aims and hypotheses:*
> ***To examine the generalizability of sight vs. sound effects in music performance, we will replicate previous studies using stimuli from Western classical music with Japanese***

> ***participants and then repeat the same paradigm using stimuli from competitions on the Tsugaru shamisen, a traditional Japanese folk musical instrument that GC (first author) has experience performing as a national champion (https://www.gakuto-chiba.com/profile1).***

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.

We have removed the first paragraph and modified the succeeding paragraphs to better match the present rationale (including moving the previous Appendix reanalysis to the introduction as requested below).

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.

As described in section 1.1 ("Study aims and hypotheses") pasted above, this is no longer one of the aims of this Registered Report.

2) 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.

We thank the reviewer for pointing out this issue. Although we updated our experimental design as described in the current version of the manuscript, the dependent variable is still metric discrete taking the values of 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. As questioned, we considered the normal distribution would not necessarily be appropriate to model this data. Therefore we decided to adopt more general rank-based nonparametric statistical methods. Our new tests basically use the relative effect (also known by several different names such as the common language effect size statistic, a probability-based measure of effect size, stochastic superiority, probability index, etc.) as a test statistic that provides more interpretable information on whether there is a superiority in the scores of choosing correct competition winners by each condition.

The revised description now reads as follows:

> *2.3 Dependent variables*
>
> ***The dependent variable will be the percentage of participants correctly choosing the 1st-placed performer in a two-choice forced-choice paradigm. As described above, participants will be asked to choose the actual 1st-place winner five times in each domain × variance combination. Therefore, the dependent variable will be metric discrete data taking values of 0.0 (no correct choices), 0.2, 0.4, 0.6, 0.8 and 1.0 (all correct choices). This data will not necessarily approximate the normal distribution, so we will adopt nonparametric testing approaches (while also reporting parametric t-tests to enable***

exploratory comparison with Tsay's and Mehr et al.'s original analyses). After being presented with all tasks, participants then provide demographic information including gender, age, and musical experience.

2.4 Statistical analysis

2.4.1 H1 (prediction of interaction effects between the domain and the variance)

We will use a rank-based procedure factorial design which is designed for the general nonparametric testing of treatment effects (Noguchi et al., 2012; Friedrich et al., 2017; Brunner et al., 2018). The null hypothesis is that the interaction effect of the two factors (i.e. the domain and variance) is zero. The ANOVA-type statistic will be used as a test statistic and we rely on the R-package nparLD for its calculation for repeated measurements (Noguchi et al., 2012). Regarding the use of nparLD, it is known that the ANOVA-type statistic does not lead to asymptotically correct statistical decisions (Friedrich et al., 2017). However, we consider it is still useful for the following two reasons. Firstly, Friedrich et al. (2017) proposed to use a wild bootstrap method to improve the asymptotic correctness of the ANOVA-type statistic but they also mentioned that both the classical way of calculation by nparLD and their wild bootstrap method brought similar conclusions even though the latter method is more accurate. Furthermore, Umlauft et al. (2019) remarked from their simulations that the classical ANOVA-type statistic can still be relied on for global testing (i.e. testing the existence of interaction effects rather than post-hoc analysis) and our test is 2 × 2 factorial design, so the theoretical issue of the ANOVA-type statistic is not practically relevant in this study.

2.4.2 H2-H3 (prediction of the dominant domain for each variance condition)

We will use a studentized permutation test for the nonparametric paired data (Konietschke & Pauly, 2012) which is designed for the nonparametric Behrens-Fisher problem and is not requiring symmetry in the distribution as like the Wilcoxon signed-rank test. Formally, this method tests the relative effect [1] [2] q = 0.5 as a null hypothesis which means there is no difference between the paired data. In this study, we predict q > 0.5 as a one-tailed alternative hypothesis (i.e. a particular domain condition yields a higher percent correct). In H1, the two samples to be compared are the low-variance × visual-only condition and the low-variance × audio-only condition paired by participants. Similarly, the high-variance × visual-only condition and the high-variance × audio-only condition paired by participants are the target two samples of H2. The R-package nparcomp (Konietschke et al., 2015) will be used for the implementation.

3) 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.

In 2.4 Power analysis, the authors should explain the details of the power analysis (e.g., tools and type of the statistical test).

We thank the reviewers for this important question. We have now added discussion of effect sizes from previous studies into our power analysis It turned out that the necessary sample size was still based on the effect size corresponding to Cohen's d of 0.4, but we consider now that this number has reasonable rationales. We have also added more details on the precise methodology as requested, as follows:

> ***2.5 Power analysis***
>
> **A priori power analysis requires estimating the effect size before collecting data, which is notoriously difficult (Brysbaert, 2019). In this paper, we rely in part on previously published data from several hundred participants from Tsay's (2013) original study and Mehr et al.'s (2018) direct and conceptual replications. Because replications tend to more accurately estimate effect sizes than first publications due to publication bias (Open Science Collaboration, 2015), we focus on Mehr et al.'s data over Tsay's. We will set acceptable false negative and false negative parameters based on commonly used power guidelines of 80% and a family-wise *alpha* level of 0.05 (i.e., .0083 for each of 6 hypothesis test; see above for rationale).**
>
> **As described in section 1.1, re-analysis of Mehr et al.'s data using using the parametric t-tests originally used by Tsay and by Mehr et al. suggests a range of effect sizes ranging from a minimum of Cohen's d = 0.42 (for Study 2) to 0.57 (for Study 1 directly replicating Tsay) to 1.2 (for Study 3). When these data are reanalyzed using the non-parametric methods planned for our confirmatory analysis, these correspond to relative effect sizes ranging from 0.62 (Study 2) to 0.64 (Study 1) to 0.80 (Study 3). Since all data in our within-subjects experiment are collected from the same participants, our necessary sample size will be determined only by the smallest effect size of interest. Given that the smallest effect size found previously (Cohen's d = 0.42) is slightly larger than the value of 0.4 often cited as an approximation of the "smallest effect size of interest" (SESOI; Lakens, 2017), we will use the more conservative SESOI of d = 0.4, corresponding to a minimum relative effect of 0.61, giving a required sample size of n=155 participants. Note that this estimate is based on a between-subjects design, so because within-subjects designs are considered to potentially have higher power than between-subjects designs (Lakens, 2013) this is likely a conservative overestimate of the true sample needed to achieve power of 80%.**
>
> **Regarding the interaction effect, we obtained a partial eta squared of 0.20 from the ANOVA-type statistics. By using this value as an input of G*Power (Faul et al., 2009), the required sample size was estimated as 53 participants in total. This estimation was based on the fixed-effects ANOVA setting as in the above presumptions. Since this estimate gives a substantially**

lower minimum sample size than described above, we will again use the more conservative estimate of n=155 participants described above.

4) 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.

5) 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.

<Response for 4 & 5>
We thank the reviewer for these important points. After careful consideration, we completely redesigned our experiment to be able to test our predictions in a more straightforward way and to investigate the difference in audience judgments against Western and non-Western music performances. Firstly, we decided to follow previous studies in using only one dependent variable - the percent correct of choosing actual competition winners through 5 trials per condition. Secondly, in order to explicitly factorize the influence of the domain (audio-only, visual-only, audio-visual) and the variance condition (high-variance, low-variance), we changed to present 2 clips per trial and the types of clip pair will be prepared for each combination of those two factors plus instrument (e.g. piano × audio-only × low-variance, Tsugaru-shamisen × visual-only × high-variance). As a result, we will present 12 (= 3 × 2 × 2) condition variations of paired clips stimulus and 5 stimuli will be given in each condition. However, the audio-visual data will be only used for the exploratory analysis. There is also an additional set of stimuli that are also only used for the exploratory analysis. We consider this new experimental design allows us to investigate what factor (i.e. modality and performance quality gap) affects the choice of performers and to generalize the previous studies cross-culturally.

Copied below is a substantial new section explaining this new method, including the explicit inclusion of variance in performance quality as an independent variable and the splitting of

stimuli into confirmatory (audio-only vs. video-only) and exploratory (audiovisual samples + Mehr et al. Study 2 replication):

2.1 Stimulus choice

2.1.1 Confirmatory sample

To enable us to replicate and generalize previous studies we designed a paradigm that allowed us to compare our results as directly as possible with Tsay (2013) and Mehr et al. (2018) by having the same participants rate both piano and shamisen performance stimuli in the same experiment. However, each of the three paradigms reported in Mehr et al. used slightly different designs: Study 1 used 9 out of 10 sets of excerpts of three performers (1st-3rd place) previously used by Tsay (2013); Study 2 used 10 sets of only two performers; and Study 3 used 5 sets of 2 performers (see https://osf.io/6nx4d for details). As Mehr et al. explain, this meant that they could not conclusively determine whether differences in their results were due to differences in experimental design or differences in the independent variables of interest (i.e., audio vs. visual domain or high vs. low variance).

To avoid these confounds, we chose to unify our experimental design based on the paradigm with the smallest number of stimuli, namely the 5 pairs of performers used in Mehr et al.'s (2018) Study 3 (high-variance condition). We thus collected analogous 6-second excerpts of performances from 10 pairs of Tsugaru shamisen performers: 5 "high-variance" pairs (1st place and low-placing performers, as in Mehr et al. Study 3) and 5 "low-variance" pairs (1st and 2nd place performers, as in Mehr et al. 2018 Study 2). These performers were selected from different competitions so the 1st-place performers would not overlap between the high-variance and low-variance conditions. For all Tsugaru shamisen performers, GC (1st author) selected an excerpt from the same portion of the opening of the piece "Tsugaru Jongara Bushi", because it is the most famous piece among Tsugaru shamisen players, and it is a compulsory component of all competitions, which allows us to collect a large number of comparable samples.

To choose 5 "low-variance" pairs from the 9 1st/2nd place performers previously used by Mehr et al. and Tsay, we removed four pairs that seemed least appropriate to compare. These included:

-two sets of violin performances (all other performances were of piano and all our performances were also of a single instrument, Tsugaru shamisen)

-one set including a 4-second clip rather than a 6-second clip after audience applause was edited out

-one set including a 1st-place performer that overlapped with one of the sets used in Study 3.

Pilot experiments (see below) suggested that restricting the stimuli to only 5 of the 9 previously used by Tsay (2013, Study 3) and Mehr et al. (2018, Study 1) did not appear to change the main sight-over-sound result reported by both.

This gave us a full set of 40 performances from 20 competitions for our main confirmatory analyses: 5 low-variance piano, 5 high-variance piano, 5 low-variance shamisen, and 5 high-variance shamisen (Table 1).

2.1.1 Exploratory sample

Tsay (2013) and Mehr et al. (2018) used a between-subjects design where different participants independently rated audio-only, visual-only, or audio-visual stimuli, but the same participant did not evaluate different domains. However, to increase statistical power and comparability we designed ours to be within-subjects, so the same participant evaluates all examples across all domains. To eliminate the possibility of order effects by which participants' judgments of audio-only or video-only samples would be affected if they had previously seen the audiovisual condition, we chose to focus our confirmatory analysis only on the key conditions of interest - audio-only vs. visual-only - and present these stimuli first. For exploratory comparison, audiovisual examples were also included at the end of the experiment, but these are not included in our confirmatory hypothesis testing. (The order of stimuli within the audio-only/video-only block and the audiovisual block is randomly determined.)

Also, although we chose to use 1st and 2nd-place performers from Mehr et al.'s Study 1 in order to allow us to also compare with Tsay (2013) who originally reported these stimuli, we also added stimuli from Mehr et al.'s Study 2 in order to allow exploratory analysis of the effect of changing the precise stimuli used. To choose a matched set of 5 pairs from the original 10 prepared by Mehr et al., we again excluded violin performances and also excluded two sets that included partial overlap with the stimuli used in Experiment 1 (i.e., the 6-second excerpts only differed by including/excluding 1-2 seconds). Thus each participant evaluates a total of 50 6-second excerpts from 25 pairs (40 performances / 20 pairs confirmatory [Table 1], 10 / 5 exploratory), and each performance is evaluated in three different formats; audio-only (confirmatory), video-only (confirmatory), and audiovisual (exploratory, saved for after the randomized audio-only/video-only block). This gives 50 excerpts x 6 seconds x 3 domains = 15 minutes worth of stimuli. This took pilot participants approximately 45 minutes to listen/watch and evaluate. The full pilot experiment can be accessed at https://gakuto101207.github.io/.

2.2 Independent variable

> ***We have two independent variables: 1) stimulus domain (Audio-only vs. Visual-only [plus Audio-Visual for exploratory analysis]) and 2) the ranking gap of two performers as a proxy of the variance in their performance quality (High-variance and Low-variance). As a factorial design analysis, our experiment belongs to the repeated measures two-factor crossed design (domain × variance) where each factor has two factor-levels. Incidentally, studying the interaction effects brought by musical instrument/genre (Western classical piano vs. Japanese folk Tsugaru shamisen) is not within the scope of our hypotheses so this is not counted as a factor, but we will add this into our factorial design model in the exploratory analysis. Participants will be randomly assigned 9 tasks, 3 from each of these types. In the Audio-only condition, only the sounds of "Tsugaru Jongara Bushi" by the three players are heard in succession, with no visual input. In the Visual-only condition, only the three players are displayed on the video screen in succession, with no auditory input. In the AudioVisual condition, three performance videos with sound are presented. In these three conditions, participants are asked to evaluate all performances.***

Minor Point
6) 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.

We moved the Appendix to the Introduction and explained the details of previous studies (Fig.1).

Reviewer #2 (David Hughes):

This is a fascinating and valuable article. I myself have performed Tsugaru shamisen in Japan before audiences, and I have also (despite trying to avoid it!) been a judge at a few folk song contests, some of which have included Tsugaru shamisen performers.

I'd never thought about whether sight (i.e. the appearance of the performers, their facial expressions, movement, clothing etc) influenced my judgment of performers, sometimes overcoming sonal differences between contestants (not only shamisen players but also singers). This article finally makes me think about it – about music as a "multimodal phenomenon", to quote the authors.

I have toured the UK (as lecturer, co-performer etc) with two different folk music groups from Japan which included high-level Tsugaru shamisen players. Though these performances were concerts, not contests, I was definitely aware that bodily movement, facial expression and other visible elements impacted on the audience, and I even advised the performers of this. In fact, the performers were already aware of the importance of visuality. So indeed this article is important for bringing this to my conscious attention, and to realise that surely sight is indeed competing with sound for many judges and audiences.

We thank the reviewer for this encouraging assessment.

1) This article should definitely be published and publicized. But for readers like me, more detailed explanation of some terms and concepts will be needed.

This article is quite technical, sometimes using terms that elude me a bit. On p.1, line 27 of the abstract, “d = 0.4” confused me. I searched the internet for this phrase, and it seems to link with “Cohen’s d”, though even then I couldn’t completely grasp it. This article does explain it somewhat, as an “effect size”, but perhaps a bit more explanation is needed for non-specialist readers. Also in section 2.3, the terms “p-value”, “t-test” and “alpha level” need more explanation, at least for people like me! Also, “GC” in line 188 puzzled me, but I presume it refers to the co-author Gakuto Chiba.

Various small changes are needed to help non-specialist readers. For example, on p.4, line 119, they write that they selected “brief 6s excerpts” for one test. I’ve finally realized that 6s = 6 seconds! But this was only made clear to me in line 327. They should change this to “brief 6-second excerpts”.

We have corrected “d = 0.4” to “Cohen’s d” and added the explanation of effect size in new sections 2.4 and 2.5. In the same new sections, we also have added the more detailed explanation of the terms “p-value”, “t-test” and “alpha level” needed in our analysis. Moreover, we have corrected “GC” to “GC (1st author)”, “6s excerpts” to “6-second excerpts”. Finally, as suggested by the Editor, we have added the following footnote at the first appearance of “d” in the abstract to point non-specialists toward appropriate resources to help them understand the statistics and methodology of this Registered Report:

> ***1. This Stage 1 Registered Report is a proposed protocol designed to be used for collecting full data after the initial protocol has been reviewed and approved. It includes a power analysis to determine what is a reasonable number of participants to recruit to appropriately balance logistical feasibility against the risks of false negative and false positive results. This involves terminology that may be unfamiliar to some readers without a background in statistics (e.g., “Cohen’s d”; “80% power”). For accessible introductions to Registered Reports and power analysis, see Chambers (2019) and Braebart (2019), respectively.***

They have also cited many writings that I, as a “normal” ethnomusicologist, have never read or even heard of. This is useful, in that those technical writings are surely important to the more scientific readers at whom the article is clearly aimed. Much of the analysis and discussion in this article will indeed appeal to and interest people like me – as I noted above, it certainly made me think more clearly about the influence of visuals in music performance. But this article will also reach out to scholars focusing on sound perception and “cross-modal” analyses.

They also cite a range of writings about Western classical music, noting that scholars have often pursued “the role of visuals and sound” (p.3). Thus considering the sonic-visual differences in perception of other musical genres broadens such analysis in a valuable way.

One of their predictions (p.4, H1) is that “visuals will dominate the judgment” of the upper ranks of a Tsugaru shamisen competition, when the sonic performances are very close in quality. I’d never thought of that, but in fact I agree.

Still, the method for the test described there – having different people judge a performance in different ways, by only its audio, only its visual, or both together (audiovisual) – is excellent and indeed focusses on the theme of this article. And then comparing the participants’ judgments with the actual outcome of those performers in a contest helps us understand how audio and visual judgments can differ greatly.

Overall, despite some terminology eluding me, I truly look forward to the results of the full testing they will conduct, again focusing on the different perceptions by their test participants of audio, visual and audio-visual versions of performances. Thus I support their plans 100%. This article simply needs a bit more clarity in some places (mostly mentioned above) to make things clearer and easier for readers like me who are unlikely to read all the relevant publications in their bibliography.

CONCLUSION: Yes, this article deserves full support and publication.

We are delighted by this positive assessment.