慶應義塾
Keio University

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

February 13th, 2023

RE: https://doi.org/10.31234/osf.io/xky4j, *Sight vs. sound judgments of music performance depend on relative performer quality: Cross-cultural evidence from classical piano and Tsugaru shamisen competitions*

Dear Dr. Yamada,

We appreciate you managing this review process and providing the final feedback on our submission. We are grateful for the chance to use additional constructive feedback from two Reviewers to refine our manuscript. In addition to changes accommodating the given comments, we have also corrected minor formatting errors (e.g. italicizing subsection titles). Note that some of these minor errors were in Sections 1-2 that are "locked" from substantive changes, but none of these make any substantive change in the content. For the one substantive change to Section 2 suggested by reviewer 2, we have chosen to respond not by changing Section 2 but rather by adding an additional unregistered analysis to the Stage 2 portion of the manuscript, as you suggested. All changes are tracked for transparency.

We feel that the review process finally made our manuscript deliverable to readers. We hope you will find the revised manuscript acceptable for formal Stage 2 Recommendation.

Sincerely,

Patrick E. Savage
(on behalf of the authors)

Editor summary (Yuki Yamada)

Minor Revision

Thank you for submitting a Stage 2 manuscript with very intriguing results and discussion. I think this paper needs only minor revisions.

As you can see, we received peer review comments from two experts.

One gave detailed advice on how to graphically present and describe the results, and how to treat claims in the discussion. These would benefit the manuscript from serious consideration.

Thank you, we believe we have incorporated all suggestions (see below for details).

The second reviewer was also quite satisfied with the manuscript, but commented on the multiple comparisons. This comment calls for a change in Section 2.4.3, which is locked in Stage 1 and cannot be directly revised. Therefore, this point can be mentioned in the discussion if necessary or added to the results section as an unregistered analysis. Alternatively, you may want to simply disagree with the reviewer. Whichever approach you choose, please let us know why in your reply.

We chose to address this by taking up your suggestion of adding additional unregistered analyses, which confirmed the robustness of our results to this issue (see below).

Please see the individual peer review comments for details. We look forward to your corrections and re-submission.

Yuki Yamada, Recommender

Reviewer #1 (Kyoshiro Sasaki):

Summary
The present study performed conceptual replications examining which information (visual or auditory) dominates in evaluating performance in music; importantly, the present study used classical piano competitions as like the previous studies and expanded the same paradigm for the Tsugaru shamisen (i.e., a traditional Japanese folk musical instrument) competitions. As a result, the present study found significant interactions between domain (visual vs. auditory) and variance in quality (1st and 2nd place vs. 1st and low-placing performers) in both instruments. However, the trends were slightly different between the instruments. In the low variance, visuals dominated the judgment of the piano performance, while this fashion was not found in that of the Tsugaru shamisen one. Moreover, in the high variance, auditories dominated the judgment of the Tsugaru shamisen performance, while this fashion was not found in that of the piano one. They discussed their findings along with the original and other replication studies (including the recently published paper of Wilbiks and Yi (2022)).

#General comments
I would apologize for the delay. The authors completed a good work and I'm happy to review this full paper. This 2nd-stage manuscript followed the protocol and the authors discuss their results appropriately. Thus, there is little problem. The following comments may help to brush up the present study.

We sincerely appreciate you taking the time to refine our manuscript at both Stage 1 and Stage 2.

#Specific comments
- 2nd paragraph in the "2.5 Power analysis" section: ".... re-analysis of Mehr et al.'s data using using the parametric t-tests ...." -> ".... re-analysis of Mehr et al.'s data using the parametric t-tests ...." (FYI, there are other errors of this kind)

We thank the reviewer for catching our careless mistakes and apologize for the inconvenience. We went through the manuscript and have corrected these minor errors.

-It is unclear what "H4, H5 and H6" mean, although I guess that H1-3 are for the piano condtion while H4-6 are for the Tsugaru-shamisen one. Please clarify these.

We added the sentences in the Hypothesis section to explicitly state H1-3 are for the piano condition and H4-6 are for the Tsugaru-shamisen one.

"H1-H3 are the hypotheses for the outcome of the experiments using piano stimuli. Similarly, we also formalize the exactly same hypotheses for the case of Tsugaru-shamisen, which are labeled as H4-H6."

-It is better to keep the significant figures of the statistics consistent within the manuscript.

We thank the reviewer for pointing out this matter. We have updated the following figures to make them consistent. However, we would like to mention that we hold 2 decimal places for ANOVA-type statistics (corresponding to F statistic of parametric ANOVA tests) since this seems more commonly used than 2 significant digits and can retain the necessary precision for p-values .

- **4.1 Confirmatory analysis: 2 figures**
- **Table S2: 4 figures**
- **Table S4: 11 figures**

-It is difficult to interpret the graphs of the relative effects (Fig4b and 4c). The relative effect values in the text do not seem to match the values in the graph; for example, they reported "relative effect = .69" in text for the visual domain in the low variance for the piano condition, while the relative effects seem to be under .625. Why? Perhaps, is the "relative effect" different between in-text and in- graph? Anyway, it might be better to added some explanations.

As you noticed, the relative effects depicted in figures (b) and (c) and the quantities in table 3 are different, and we found an appropriate explanation for that needs to be included. We have added the following clarification in the figure caption.

> ***"Note that the relative effects shown in (b) and (c) indicate the superiority of each percent accuracy within the 4 conditions (visual × low-variance vs. audio × low-variance vs. visual × high-variance vs. audio × high-variance) for the sake of measuring interaction effects among these conditions, but the relative effects tested in H2, H3, H5, and H6 (cf. table 3) are the superiority of the percent accuracy between the 2 conditions of interest (visual × low-variance vs. audio × low-variance, or visual × high-variance vs. audio × high-variance), so the relative effects on (b) and (c) and table 3 are different."***

-The visualization of the accuracy of each pair (i.e., Fig 6) is highly helpful. If my understanding is correct, the left-most pair (i.e., M.F. vs D.L.) has the opposite tendency to the results of the low variance for the piano condition in the confirmatory analysis. Like this, one of the Tsugaru-shamisen pairs (i.e., Y.W. vs N.K.) has a different trend from the results of the confirmatory analysis. It might be beneficial to discuss what could trigger these differences; of course, this might stem from just noise.

We appreciate your taking the time to look into the details of this result. We have investigated these two cases, which we agree make for very instructive examples, and in fact led us to add a brief additional exploratory analyses of the possible role of sex biases (and update Fig. 4 to include performer sex), as follows:

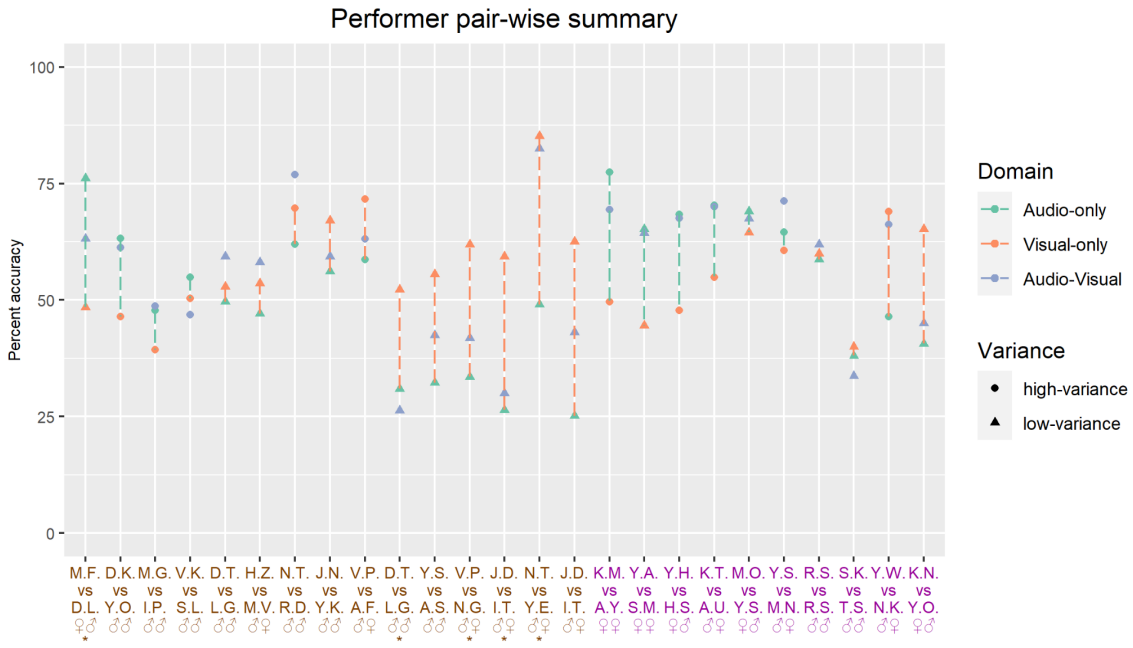

Figure 4. The average percent accuracy of each pair of clips. The x-axis labels show the initials of performers (1st place appears on the top/left) and whether they are piano players (brown color) or Tsugaru-shamisen players (purple color). ***The ♂ (male) and ♀ (female) symbols indicate the sex of the performers****, and the asterisks indicate the stimuli only appearing in the exploratory analysis. Dashed lines indicate the difference in the average percent accuracy between the audio-only condition and the visual-only condition. The color of the dashed lines is green if the percent accuracy of the audio-only condition is higher than the visual-only condition, and the orange color is used for the opposite case.*

Exploring some of the exceptions in Figure 4 illustrates some of the factors influencing sight vs. sound dynamics. For example, the M.F. vs. D.L. example (far left of Fig. 4) shows an exception where participants performed substantially better when choosing between 1st- and 2nd-placed piano performers using audio-only, and were below chance with video-only. One possibility is that the combination of virtuosic playing with unusual facial expressions the 1st-placed M.F. made in this excerpt (https://osf.io/sw8ck) compared to D.L.'s more subdued performance and neutral expression (https://osf.io/3kue8) may have contributed to this exception. The opposite kind of exception (1st- vs. low-ranked shamisen performances with higher accuracy in the video-only condition) appears in the example of Y.W. vs. N.K. (2nd from right in Fig. 4). Here, it is possible that the striking acoustic characteristics of 21-50th-placed N.K.'s hard-bodied, tightly-stretched shamisen (https://osf.io/48tb2) may have led non-expert participants to choose this over the more subtle performance of the 1st-placed Y.W. (https://osf.io/p5uca). Importantly, both exceptions involve one male and one female performer, and both might be partially explained by a tendency for participants watching video-only to guess that the male won. In the piano case, this assumption is incorrect (possibly explaining the higher audio-only accuracy), while in the shamisen case, the assumption is correct (possibly explaining the higher video-only accuracy). To explore this possibility more systematically, we conducted an exploratory analysis of the video-only results for all 13 examples where participants had to choose between performers of different sexes (cf. Fig. 4). For all 9 cases where the male came 1st,

participant accuracy was greater than 50% (mean: 64%). For the 4 cases where the female came 1st, participant accuracy was greater than 50% in two cases and less than 50% in two cases (mean: 56%). This exploratory analysis is consistent with a weak bias toward choosing males, but cannot be treated as conclusive, since our study was not designed to rigorously test for such biases in a controlled manner and the trend was not consistent for all examples (participants did not always tend to choose male performers when audio was not available). Future controlled studies would be required to conclusively test for the existence of specific biases regarding sex or other factors (e.g., age, race).

-On p.23, they stated "we suspect that the different cultural backgrounds of participants may have played a stronger role". It should be better to suppress the tone of this. At this time, as the authors mentioned, there is little evidence supporting this claim, and it is quite possible that the failure of the replication is due to the slight differences in experimental design.

We agree with the reviewer's comment. We have replaced "we suspect" with "one possible speculation is" to soften it.

-The authors proposed that visual salience might play a role in the participants' prediction of competition winners, in particular, in the case of the piano. This is an interesting idea, and the attempt to capture the mechanism of this phenomenon from the framework of information processing is commendable. For that reason, it should be desirable to discuss this issue in more detail (for example, how visual salience is involved, and how this salience is used in making judgments, etc).

We appreciate this suggestion. We have expanded the discussion on this point by inserting the following sentences in the referred paragraph.

"Our study revealed a cross-culturally consistent pattern of the sight-vs-sound effect on selecting the winners of musical competitions. This finding suggests that when people choose musical talent, they tend to base their decisions on the audio information if the variance among the performance qualities is large enough. However, once the variance becomes small (as it tends to do during final stages of auditions and competitions), people increasingly rely on other information (e.g. visual) to evalute performance. Orquin et al. (2018) summarized six visual attention mechanisms that can bias decision making: visual salience, surface size, position, set size, random location and emotional stimuli. Amongst these mechanisms, we can hypothesize that visual salience has played a role in the participants' prediction of competition winners in the case of the piano. **Visual saliency is defined as the conspicuity of a visual element compared to the surrounding visual items, and it includes motion (Orquin et al., 2018). Attire and body movements have already been identified as features affecting the perceived quality of musical performance (Griffiths, 2008; Tsay, 2013). Our findings are potentially consistent with theories of decision-making behavior based on visual saliency, such as salient visual elements being processed as readily available information to make heuristic decisions (Tversky & Kahneman, 1973, but also see a hypothesis based on passion: Tsay 2013; Tolsá-Caballero & Tsay, 2021).** This hypothesis may also explain why the sight-over-sound effect was not observed in the case of Tsugaru-shamisen since the performers and the dynamics of the camera angle is relatively plain when compared to the piano clips.**"**

Reviewer #2 (Anonymous):

Overall, I found the study well planned and substantiated. The methodological aspects were well considered. Only the following point stands out:

- I recommend using the proposal by Noguchi et al in 'nonparametric multiple comparisons' (Behavior Research Methods, 2020) as it's in line with the nonparametric tests being used. Thus, update section 2.4.3. accordingly.

Best,
F

We thank the reviewer for this suggestion. Since sections 1-2 are pre-registered, revising section 2.4.3. may violate the principle of Registered Reports, which requires researchers to fix a design and plan (hypothesis, data collection, analysis method, etc.) before collecting data and conducting analyses. However, we agree that a potential issue of using nonparametric tests for multiple comparisons, nontrasitivity paradox, would be better to control to ensure our result is not an artifact of methodological reason. Therefore, we have added a new exploratory analysis in the supplementary materials (S1.8) to re-analyze the data with the suggested multiple comparisons (copied below). We also added the following texts to the end of section 4.1 of the main text to signpost the readers to check the exploratory analysis.

> **"*Since our analysis is based on nonparametric statistics measuring the relative stochastic superiority of percent accuracy in each pair of conditions separately (H2-3, H5-6), there is a possibility that these relative effects can change when making superiority consistent among all pairs due to the nontransitivity paradox (Noguchi et al., 2020). However, our complimentary analysis (S1.8 for details) confirmed that our results are not affected by such a paradox and captures the relative effects consistently.*"**
>
> **…**
>
> ***S1.8 Nonparametric multiple comparisons controlling artifacts by nontransitive paradox.***
>
> ***Our original testing plan stipulated (1) testing interaction effects and performing two pairwise nonparametric tests for the experiments with piano stimuli and Tsugaru-shamisen stimuli, (2) grouping six hypotheses as a single family, and (3) employing Bonferroni's correction to control family-wise error rate. However, considering the hypotheses of interaction effects between the domain and variance (H1/H4), audio-only vs. visual-only under the low-variance condition (H2/H5), and audio-only vs. visual-only under the high-variance condition (H3/H6) as a family of multiple hypotheses testing, a more nuanced test can be conducted with control of potential nontransitive paradox (Noguchi et al., 2020). A nontransitive paradox is a paradox that does not preserve transitivity among multiple comparisons (e.g. A < B, B < C, but C < A, Blyth, 1972; Brown & Hettmansperger, 2002; Noguchi et al., 2020). This paradox can arise when multiple comparisons are executed with nonparametric statistics, especially when test statistics only measures pairwise relative superiority in each comparison. Noguchi et al.'s method of multiple comparisons controlling for the nontransitivity paradox can inform the consistent and***

overall superiority of samples among all pairs, and we re-analyzed our data with their multiple nonparametric comparisons with the R package nparcomp (Konietschke et al., 2015) as an exploratory analysis. In our case, this approach allows us to measure the superiority of percent accuracy among audio/visual × high/low-variance conditions, which more accurately demonstrates the effects of domain and variance. Moreover, we divide the original family into two families by instruments since we are not interested in performing multiple comparisons beyond instruments (e.g. piano audio-only under high-variance vs.Tsugaru-shamisen visual-only under high-variance). In other words, we define families in this exploratory analysis according to the unit of multiple comparisons and not by the set of multiple inferences as in the original text. Noguchi et al.'s method models p-values of multiple hypotheses simultaneously using multivariate distributions, so family-wise error control such as Bonferroni's correction is not necessary and the nominal alpha-level (α = .05 in our case) can be directly applied to interpret the result. This exploratory analysis confirmed the same results obtained in our main analysis which eliminates the possibility that our observation of the dependency of domain and variance via multiple comparisons is an artifact of a nontransitivity paradox.

Table S5-6 | Summary of nonparametric multiple comparisons (Noguchi et al., 2020). Please note this method uses test statistics and effect sizes different from the method used in the main analysis. The log odds ratio is used for effect sizes but this value is adjusted to approximate Cohen's d (Chinn, 2000; Noguchi et al., 2020; Sánchez-Meca et al., 2003). Therefore the rejection region of equivalence testing is also based on quantification by Cohen's d.

(Piano)

#	Test statistic	Obtained statistic	p-value (α=0.05)	Effect size	Obtained effect size	90% CI for equivalence testing (rejection region -0.4~0.4)
H2	Student's t-type test statistics	6.44	***4.7x10^{-10}**	Log odds ratio of relative effects	0.48	-
H3	Student's t-type test statistics	0.72	.42	Log odds ratio of relative effects	0.052	***-0.089 - 0.19**

(Tsugaru-shamisen)

#	Test statistic	Obtained statistic	p-value (α=0.05)	Effect size	Obtained effect size	90% CI for equivalence testing (rejection region -0.4~0.4)
H5	Student's t-type test statistics	0.24	.65	Log odds ratio of relative effects	0.017	***-0.12 - 0.16**
H6	Student's t-type test statistics	3.26	***1.2x10^{-3}**	Log odds ratio of relative effects	0.25	-