慶應義塾
Keio University

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

March 28, 2024

RE: Cross-cultural relationships between music, emotion, and visual imagery: A comparative study of Iran, Canada, and Japan [Stage 1 Registered Report] (doi:10.31234/osf.io/26yg5)

Dear Dr. Schwarzkopf,

Thank you very much for the opportunity to submit another round of revisions on our manuscript. We are grateful for the extremely thorough and helpful feedback from the reviewers. We have revised our manuscript accordingly, with point-by-point responses below (our responses in blue below).

The manuscript has made great improvements and we look forward to receiving your feedback. We would appreciate it if you grant our manuscript in-principle acceptance if you deem it appropriate in this stage.

Sincerely,

Shafagh Hadavi and Patrick E. Savage (on behalf of the authors)

Round 2 Review

Decision by D. Samuel Schwarzkopf, 30 Sep 2023 22:09

Manuscript: https://psyarxiv.com/26yg5/ version Version 3

Please submit another revision

Dear authors

Your revised RR Stage 1 manuscript has now been reviewed by the same two previous reviewers. Their comments are generally positive but as you see one reviewer still has several substantial concerns that I would ask you to address in another round of review. Some aspects are probably judgement calls, in particular the question of the effect size of interest and the resulting necessary sample size. As suggested by the reviewer, I may make a final decision about granting in-principle acceptance based on that in the next round.

Both reviewers also raised concerns about the statistical approach and it took me a while to figure out why. In section 1.2 "Analysis Plan" you still refer to paired t-tests instead of the new analysis approach you are now using. (I stared at the relevant page for some time without noticing it, so I thank the reviewers and Chris Chambers for pointing it out!) I assume this is a mistake and can be fixed easily. But please go through the whole manuscript carefully to ensure all parts are up-to-date as this is the version that will define your preregistered methodology.

Best wishes

Sam

Apologies for this oversight - we have now removed the old text about paired t tests and have gone through the rest of the manuscript to ensure it properly reflects the revised analysis, paying particular attention to the sections pointed out by the reviewers. We are happy for you to make a final decision regarding the necessary sample size, as we still respectfully disagree with Reviewer 2 (Karakashevska) on this point, while appreciating their constructive spirit and many helpful suggestions.

–

Reviewed by Nadine Dijkstra, 28 Sep 2023 16:25

I thank the authors for addressing most of my comments. My only remaining issue is with the proposed analyses: three paired t-tests to determine whether the same effect is present in each country. To establish whether there are differences between the countries, they need to be compared within the same statistical test. One option is already pointed out by reviewer 3: an ANOVA with country as a between-subject factor.

Apologies - as pointed out above by Sam Schwarzkopf, this wording was unintentionally preserved from the previous version. We have now updated this with the improved testing suggested by reviewer 3 (now coauthor), Leongómez.

—

Reviewed by Elena Karakashevska, 19 Sep 2023 16:52

I appreciate the author's efforts for revising this Stage 1 manuscript and commend the improvements made with the wording issues and statistical analysis. With regards to the power analysis the authors have adopted the approach suggested by reviewer 1 which is well conducted. The Appendix contains excellent detail describing the process of the power analysis. However, there are still inconsistencies in the analysis plan that have not been addressed in the manuscript. The authors have not explained how the CLMM will be implemented; provided justifications for choosing effect sizes of interest; explained how they will perform equivalence testing if they don't find any significant effects. The analysis of the pilot data is also not explained. I believe the manuscript can still be improved prior to being accepted.

Thanks for this thorough re-evaluation - we have responded to each point in depth below.

Major issues

While effect sizes estimated from pilot data are by definition unreliable (Brysbaert, 2019; Albers & Lakens, 2018), we note that all of our pilot data groups demonstrated effect sizes greater than 1 (minimum: ~1.2 [effect of tempo on visual density for solo music]; maximum: ~3.2 [effect of tempo on visual density for group music]).

The above paragraph seems contradictory to me. I would please ask the authors to clarify why they chose d = 1.0 as their smallest effect size of interest considering they claim it is not based on the pilot data report in the Stage 1 manuscript. Is this the smallest effect size that is theoretically relevant in terms of differences in emotional arousal caused by tempo changes? Would an effect of 0.8 not be interesting?

Apologies, this effect size is no longer Cohen's d, but now the fixed effect regression parameter (i.e., the beta coefficient). Given that in this case the main independent variable is Tempo, the effect size simply refers to the predicted change between Tempo conditions (β). However, given that the dependent variable is ordinal (a 5-point Likert scale), we use Cumulative Link Mixed Models (CLMM) which, unlike metric models, have been shown to provide consistent results for ordinal outcomes (Peng et al., 2023). CLMMs assume that the '*real*' dependent variable is a latent, continuous distribution that we cannot directly observe. Instead, what we observe is a categorisation of that latent variable (the 5 points in the Likerts scale). For these models, and in this case, this latent distribution is commonly modelled as a logit (log odds) probability of observing a given score, so all model predictions and coefficients are in the units of that estimated latent, continuous variable (log odds). Because of this, these models include threshold coefficients to help interpret coefficients in terms of the original 5 levels of the Likert scale (see e.g., Taylor et al., 2022).

Furthermore, it is important to consider the complexities of determining a standardised effect size for model terms in mixed models, and perhaps even more when dealing with ordinal (categorical) responses. In any case, the function *model_parameters* from the R package *parameters* (Lüdecke et al., 2020, https://doi.org/10.21105/joss.02445), allows the estimation of *z* values for fixed effects in Cumulative Link Mixed Models (https://search.r-project.org/CRAN/refmans/parameters/html/model_parameters.merMod.html), and the transformation between *z* and *d* effect sizes (or indeed between log odds and *d*) is common. This would mean that a log(OR) of 1 is equivalent to a Cohen's *d* of 0.55. However, we are not completely sure it makes sense to convert log odds to Cohen's *d* when (1) log odds are already an effect size and (2) the prediction of the model is a (log odds) probability of seeing a

score in the Likert categories, rather than the score itself. We have updated the manuscript to clarify this.

On the other hand, we chose an effect size of 1 as it is a more conservative estimate compared to that found in the pilot data. Assuming the effects in the pilot data may be biassed, we opt for a more conservative estimate as a precautionary measure. In this scenario, the effect size is not a Smallest Effect Size of Interest (SESOI), as determining a proper SESOI would be challenging. Instead, it simply ensures that the experiment would be adequately powered to detect effects of at least that size, which is smaller and more conservative than those observed in the pilot.

for our study to have 95% power to test both directional (one-tailed) hypotheses assuming an effect size of 1,

Can the authors explain why the effect of interest increased from 0.04 to 1? Is this because they are no longer reporting Cohen's d but a difference in SDs?

Apologies, the original 0.04 was a typo for (Cohen's d of) .4, but now our revised analysis no longer uses Cohen's d, hence the multiple changes.

we note that all of our pilot data groups demonstrated effect sizes greater than 1 (minimum: ~1.2 [effect of tempo on visual density for solo music]; maximum: ~3.2 [effect of tempo on visual density for group music]).

An explanation for the change in effect sizes and the analysis conducted here I believe is also necessary.

See above for this explanation. All changes and these reviews/responses will be publicly available, so we don't see the need to add further explanation to the main manuscript to avoid unneeded confusion.

We will perform paired *t* tests to test our hypothesis on our dependent variables: arousal, density, emotion arousal across tempo changes in all cultures.

In the analysis table the authors state they will perform a CLMM? The analysis section (1.2) does not reflect the changes stated in the response to reviewers' letter about the completely revised statistical analysis. I apologise if I'm looking at an outdate version, however I could not find anywhere in the tracked changes and psyarxiv versions, an explanation of how the CLMM will be implemented and overview of the new mixed model analysis apart from the mention in the table. If the authors will indeed perform t tests then their power analysis should be based on that, rather than CLMM.

Our apologies for this oversight! We have now corrected the main manuscript.

If any effect is not significant, we will use equivalence testing to test whether the effect is statistically equivalent to 0 (|effect size| < 1). If we find a significant effect in one or more cultures but a statistically equivalent result in one or more cultures, we will conclude the relationship is cross-culturally variable.

If the authors are looking to show evidence for no effect, they need to explain how this will be implemented. Will they conduct a TOST analysis since they say 'whether the effect is statistically equivalent to 0' ?

Also, why is the critical effect size here < 0.41? Is this a typo? (the psyarxiv version has 1 here so I believe it's a typo in the tracked changes file)

For equivalence testing, they should justify the minimally interesting effect size in its raw units.

Yes, our apologies, 0.41 was a typo. Regarding the TOST procedure, we have decided on reflection not to implement it. This is because it would not be feasible with our current authorship team, even after having the added expertise of Reviewer #3 . While there are proposed options to do equivalence testing for fixed effects in linear mixed models (e.g., https://pedermisager.org/blog/mixed_model_equivalence/), and there are non-parametric TOST alternatives, there does not seem to be a feasible alternative for mixed models with categorical/ordinal outcome variables. Hence, we have removed any mention of equivalence testing or interpretation of null results from the manuscript.

'We disagree. In our opinion, best practice is to estimate the appropriate sample size that is needed to test the relevant prediction(s), and collect as much data as needed but not much more. Power analysis using two-tailed hypothesis tests for directional hypotheses does not make sense and leads to reduced power. Collecting more data than needed is a waste of limited time and resources.'

In response to this comment, I would in turn disagree with the authors here and state that best practice is to collect as much data as possible to be able to approximate the true population effect. My concern is that the effects the effects the authors re studying are highly variable and context-dependent, and all this variation will make it hard to find any consistent patterns. Collecting larger data sets can somewhat mitigate this.

Collecting data online for a relatively short duration experiment (30-40 minutes) should pose no 'waste of time and resources' to the authors considering the benefits of estimating true effect sizes to the literature. I would urge the authors to reconsider their approach to data collection, specifically in the design they have chosen, and increase their sample size > 24 participants per group. I understand the authors have done a great job estimating a sample size using simulated data, however I believe the study design allows for an increase in sample size which will be beneficial to the study. I will let the Recommender make the decision on this issue.

Estimating power using a two-tailed test doesn't reduce power, it forces an increase in sample size, bringing us one step closer to estimating the true population effect. In novel research questions (such as the present one), this is quite beneficial. This is similar as using power = 0.9, alpha = 0.02 rather than the conventional lower thresholds of power = 0.8, alpha = 0.05. Albeit this isn't a massive problem here considering the directional hypotheses.

We agree that the question of how much data to collect is an interesting philosophical one without a clear answer (see e.g., https://bsky.app/profile/patrickesavage.bsky.social/post/3kn2jtwqlr225 for some recent discussion and references). However, there are always trade-offs and opportunity costs (money, time, etc.) and we respectfully prefer to preserve our limited resources to the amount stated in this manuscript.

Since the power analysis is not based on the pilot study and the current analysis plan differs significantly from the pilot, I think it is not necessary to mention pilot results in the analysis section in this instance as it might make things confusing for the reader. If the authors wish to report the pilot results I recommend fitting them in the introduction or hypothesis section and including the appropriate analysis you conducted as well as the effect sizes.

We have implemented this helpful suggestion by removing pilot results from the analysis section.

Minor points

On the other hand, cross-cultural research in musical emotions has discovered both consistency and diversity in emotion appraisal in music

I think following this sentence authors should give a brief explanation of the consistency and diversity in this line of research before moving on to explain that only a few studies have looked at relationships between musical features, visual imagery and emotions.

We thank the reviewer for raising this point, however, we would like to point out that consistency and diversity in this line of research involves various dimensions of the relationship between music, visual stimuli, and emotion. In the following paragraphs in the introduction, we have dedicated one paragraph to each dimension and the way we are building upon previous research to give the readers a clearer idea of what has been done and what needs more research.

‘variables which are tempo and rhythm complexity.

Is there any reason authors decided not to include pitch and musical texture in their associations?

We thank the reviewer for raising this important comparison. As we have mentioned in the manuscript (copied below), due to the number of variables in our study, we have decided to keep our confirmatory hypothesis on tempo, emotional arousal, and visual density. We will conduct an exploratory analysis on other factors such as pitch, musical texture, etc.

“Since the solo and group excerpts differ not only in terms of solo/group but also in terms of tempo and other variables, it is not possible to compare solo vs. group recordings to draw strong conclusions about the effect of tempo or other variables. Instead, we will make such comparisons in a purely exploratory manner, and restrict our confirmatory hypothesis testing only to our controlled experiments manipulating tempo (i.e., the paired responses represent direct comparison of faster and slower versions of the same musical excerpt by the same participant).”

The figure on page 3 is missing a number and figure description.

Authors refer to fig 2 in the hypothesis section and present fig 2 in the materials section. Perhaps it would read better if fig 2 is moved to the introduction.

Table 1 is missing a title.

We thank the reviewer for catching these mistakes. We have revised the figures and added titles in the manuscript. We moved the figure on page 8 (now figure 1) to page 3.

We plan to recruit participants from Japan, Iran, and Canada (n=24 per group; see Power Analysis below), who are raised in these countries and whose first language is Japanese, Farsi, or English, respectively, with any level of musical training.

Are Canadian French speakers going to be excluded?

Yes, French Canadians will be excluded, since we are looking to recruit participants from Canada whose first language is English (we have added this clarification to the manuscript).

these specific tempo after finding in our preliminary pilot analyses that these tempi

We thank the reviewer for catching the typo. It is now corrected in the manuscript.

randomly presented with 24 excerpts

why did the number of excerpts increase?

Our stimuli consists of six excerpts manipulated in tempo and pitch. Each stimulus is presented in a high/low tempo and high/low pitch therefore resulting in 24 excerpts. As discussed earlier we will only include tempo changes in our confirmatory analysis and keep the pitch changes for the exploratory analysis. We have added the following clarification to the manuscript:

> **"(6 recordings each manipulated 4 ways to be higher/lower in tempo/pitch)"**

experiment or those who do not complete each section will be excluded.

Will these participants be replaced?

We have added "(and replaced with new participants until we meet our target number of participants)" to the manuscript to clarify that since we are looking to collect data from a total of 72 participants across the three cultures, in the event of missing data for any participants, we will replace them until we have reached our target number.

2.1.2. Independent variable

Our independent variable is tempo, which is manipulated by raising the tempo to be either 20% higher or 20% lower than the original tempo.

2.1.3. Dependent variables

Our visual stimuli consists of a texture that varies in density on a scale of 1 to 5, 1 being the least dense with one line and 5 being the most dense. The number of lines increases density=X2+1 . Scale 1 has one line and scale 5 has 31 lines. Arousal ratings will be done using scales 1-5, 1 being very subdued and calm, and 5 being very excited or aroused.

I believe these two sections are not necessary and the information already exists in the Methods and Materials sections, so it is repetitive.

We thank the reviewer for pointing this out, however we think it might be easier for the readers who refer to the subtitles to have the information in both places.

for our study to have 95% power to test both directional (one-tailed) hypotheses assuming an effect size of

Please also state the alpha level here.

The alpha value was previously provided below ("...while maintaining an overall false-positive rate of 5% (p < .025 after applying Bonferroni correction")

We will test our hypotheses by a cumulative link mixed model (CLMM) on our dependent variables (here, arousal) for high vs. low tempo versions of the 6 excerpts.

Missing ‘performing’

Thank you very much for catching the missing word. We revised it in the manuscript.

Participants will be presented with Fig 2 which is a series of 5 visual textures ascending in visual density represented by five circles with increasing numbers of parallel horizontal lines. Each circle has a diameter of 2cm.

Please state the size of the visual stimuli in visual angle.

We thank the reviewer for raising this point. We would like to draw your attention to the following:

1- In designing the experiment on the online platform, we will make sure that the stimuli looks undistorted and homogeneous on different devices.

2- We acknowledge that the scale of the stimuli will be different based on which device the participant is using to do the experiment, however, the five circles (scales 1-5) will be the same size for each participant regardless of the device and the screen scales. Therefore, it should not matter which device they are using since the variable in the visual stimuli lies in the number of lines within the circles from scale 1 to 5.