**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**)

***Round 2 Review***

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.

***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?

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?

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.

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.

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.

*‘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.

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.

***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.

‘variables which are tempo and rhythm complexity.

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

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 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?

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

tempi

randomly presented with 24 excerpts

why did the number of excerpts increase?

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

Will these participants be replaced?

**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.

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.

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’

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.