慶應義塾
Keio University

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

Aug 29, 2023

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 your invitation to revise and resubmit our manuscript based on the constructive comments of the three Reviewers. We appreciate the opportunity to improve our manuscript in regards to the clarity of the research question as well as the analysis plan. We are submitting our updated version with tracked changes in addition to a point by point response to all comments.

Once again, we are grateful to have received valuable feedback resulting in substantial improvements to our manuscript. We hope that the updated version meets the requirements for In Principle Acceptance.

Note that, after consulting with you and PCI-RR Managing Board member Chris Chambers, we confirmed that it is permissible to invite Reviewer #1 (Dr. Juan David Leongómez) to join us as a coauthor and collaborator to implement his excellent suggested statistical analyses. We have done so and he has graciously accepted our offer. We all understand that he will be recused from any further reviewing of this and future versions of the manuscript.

Sincerely,

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

Editor’s comments

Rationale:

I concur with the reviewers' concerns that the rationale and the whole premise of the study lacks clarity. For example, it should be explained early in the manuscript what is meant by "visual density", how this relates to visual imagery, and why you expect that to depend on music.

We thank you and the reviewers for pointing out this important issue. We have added the following explanation in the introduction and hope that it clarifies our aim.

> **We use metaphors from other domains to understand music as an abstract experience. Many music-related metaphors are related to other sensory experiences such as visuals. Additionally, one of the ways of studying music-related visual imagery is through music to visual associations (Athanasopoulos, 2022) where visual metaphors such as height, size, roughness, horizontal placement, and color are presented to match with sonic stimuli. These cross-modal associations are shown to be mediated by factors such as musical training (Kussner and Leech-Wilkinson, 2014), and language (Dolschied et al., 2022). Additionally, research about music-based visual imagery has examined how musical features such as pitch, volume, and (the representation of) time in music are associated with visual imagery, and revealed correlations between musical and visual features such as pitch and spatial height (e.g. Athanasopoulos and Moran, 2013; Dolschied et al., 2022; Eitan and Timmers, 2010; Kussner 2014; Rusconi et al., 2006; Tan and Kelly, 2004), and horizontal time representation (Athanasopoulos and Moran, 2013; Kussner, 2014; Walker, 1987).**
>
> **On the other hand, there are other visual features whose association to music is understudied. For example, there have only been a few studies that examined the cross-modal relationship between musical and visual textures such as the study about the association between musical dissonance and visual roughness conducted by Gianno et al. (2021). Musical texture can refer to melodic, rhythmic, and harmonic layering of musical material (Sarrazin, 2016) and is often conceptualized with notions that describe visual texture such as dense/sparse, thick/thin, rough/smooth, etc. Additionally, visual textures are either plane or volumetric and a plane simple visual texture is characterized by three dimensions of directionality, size, and density. (Caivano, 1990) Density, as Caivano (1990) states, “depends on the relation between the texturing elements and the background”. By examining the relationship between rhythmic density (i.e. tempo) and visual density, we aim to investigate this understudied aspects of cross-modal mappings that could add to our understanding of music-induced visual imagery and music cognition.**

We have also included the following paragraph in our study aims and hypothesis section to clarify some of the suspected underlying mechanisms.

In our study, variation in rhythmic density (number of note onsets per second, or "tempo"[1]) is captured through a selection of solo and group excerpts, which are also directly manipulated by speeding up/slowing down the recording. Meanwhile, visual density is presented through horizontal straight lines against a white blank background and ranges from low to high on a scale from 1-5 (Fig. 2). For example, a higher number of lines within the same background area is equal to a higher density and vice versa. Visual density has yet to be empirically examined in cross-modal studies, however, it can provide some insight into music cognition across individuals and cultures since we use metaphors from visual textures to conceptualize musical complexity in terms of texture.

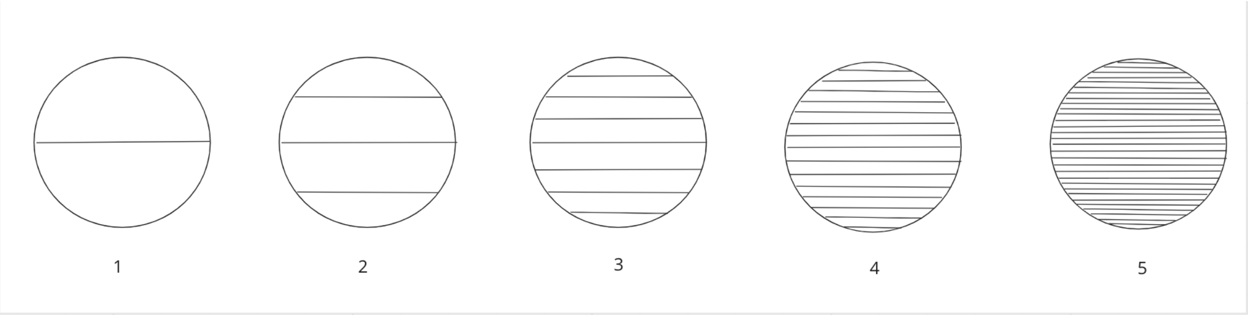

Fig 2. Adapted from the stimuli by Langlois et al. (2014).

Empirical research from Antovic (2009a) and Eitan and Timmers (2010) has found diverse cross-modal mappings concerning musical pitch. To some extent, their results supported conceptual metaphor theory (CMT; understanding of a concept from a specific domain through the usage of metaphors in another domain) brought forward initially by Johnson and Lackoff (1980), but also suggested other potential mechanisms for these cross-modal mappings (Antovic, 2011). In this instance, do we make sense of rhythmic density in music through a visual density metaphor? If we associate excerpts with higher rhythmic density with images with higher visual density, we will have some evidence that suggests CMT might be the underlying mechanism. Musical features which are abstract in nature are conceptualized through more familiar notions and understood metaphorically.

On the other hand, embodied cognition of time (here highlighted as the passage of time in music) is another potential metaphorical mechanism through which we conceptualize tempo in music. It is not fully possible to use linguistic

[1] **For comparability with past research we use the term "tempo" to describe rhythmic density, but note that tempo is often defined in terms of beats per minute assuming a fixed pulse, which does not apply to some of the musical stimuli in our study, such as the ametric Japanese shakuhachi music (cf. Savage & Fujii, 2022). Our "rhythmic density" / "tempo" variable is also analogous to the "temporal rate" variable used by Ozaki et al. (2023) to compare rhythmic density in music and speech.**

> **reasoning to understand how musical features such as tempo interact with visual density e.g. whether and how tempo changes alter these music to visual density associations; however, can be attributed and supported by notions of Time-Metaphors and Space-Time Metonymies by Lackoff and Johnson.**
>
> **Lackoff and Johnson (1999) indicate that some metaphors for time are consistently found in different languages which explains how we relate motion to time i.e. we move in relation to other objects and people while time is passing.**
>
> **Through this idea of time-motion metaphor, our study might suggest that rhythmic density to visual density associations refer to the passage of time and the density of musical events representing the amount of motion through time (higher density of musical events attributed to more motion, and subsequently associate with higher visual density and in this case, more lines).**

Statistical approach:

Moreover, all reviewers raise issues with the statistical approach that parallel my own comments from pre-screening. The series of t-tests does not seem to accurately reflect the hypotheses. An effect across all cultures could be reflected in a significant main effect in an omnibus test. But as you see, reviewer *Juan David Leongómez* has an alternative suggestion and even provided some example code to be used for a potentially better-suited approach. I suggest you develop these proposals further - the reviewer offered to assist with this but of course you may wish to also draw upon the expertise of others.

As described above, we were so impressed with and grateful for Dr. Leongómez's analyses we invited him to join us as a coauthor to help implement them, and have rewritten all relevant statistical sections accordingly.

Sample size:

One reviewer calculates a sample size for your expected effect size of n=41 rather than 14, which makes me wonder if the 14 you've used is a typo that permeated through the manuscript? All reviewers raised a concern with this small sample size. You seem to expect a relatively large effect size that is perhaps justified for the arousal and visual density ratings, but for any difference between cultures I would expect this to be a lot smaller. As mentioned by several reviewers, given the fact you are planning online data collection, it should be entirely feasible to collect a *much larger* sample than n=14?

"14" was not a typo (indeed, the simulated power analyses in Dr. Leongómez's review also independently identify a "sample size of 14 participants (6 paired responses per participant)"). The confusion here is that each participant provides multiple paired responses, so the number of data points is actually six times

more than the number of participants. Also note that n=14 was the number of participants *per society*, so the total number of participants is three times more. That said, our revised statistical methodology and power analysis results in a larger sample size of n=72 participants total (n=24 per society for each of the three societies).

Typographic error:

Finally, there is still an error I already flagged up during prescreening: Under hypothesis 1.2.2 you talk about '*arousal ratings*' instead of *visual density* judgements.

Thank you for catching this typo. We have fixed section 1.2.2 as below:

1.2.2 Effects of tempo on density. We will test this by comparing the overall means of the visual density ratings for all excerpts in two high and low tempo conditions across all musical cultures and participant cultures. (Fig. 1B)

Note on one review:

Judging by their comments, one reviewer may have commented on the originally submitted version of the manuscript rather than the revision after my pre-screening. So some of their points may no longer be relevant. But it is also possible that I missed issues in my (hasty) pre-screening. Please check the comments carefully and respond to each point by this reviewer to clarify.

Thank you very much for your note. Kindly find our point by point response to each reviewer's comment below.

Best regards
Sam Schwarzkopf

Review #1

Reviewed by Juan David Leongómez, 11 May 2023 02:27
This is a Stage 1 Registered Report that aims to investigate the relationship between emotional arousal and visual density induced by six musical excerpts differing in tempo and texture (solo vs group) in participants from Iran, Canada, and Japan. The study design is relatively simple and straightforward, with clear independent and dependent variables, and I commend that the authors commit to best practices in cross-cultural studies, as well as including both participant samples and music excerpts from non-WEIRD countries.

However, the authors acknowledge that the study violates some of the assumptions of the statistical analysis, such as using 5-point Likert scales instead of normally distributed continuous data, and the 6 paired responses from each participant not being independent of one another. For this, I would like to offer some recommendations and took the liberty of doing a simulation-based power analysis in R for different models, that hopefully will assist the authors in this regard. Once this limitation is addressed, I think the authors should move forward and start data collection.

Statistical power and test

I appreciate the power analysis made by the authors, and their direct statements regarding limitations. Also, the decision of testing each hypothesis three times (once per country) and only confirming predictions if all three tests is sensible. However, there are several important issues here:

First, it is not good practice to conduct separate analyses and infer differences between populations from them (see a summary in the section 'Interpreting comparisons between two effects without directly comparing them' in Makin & Orban De Xivry, 2019). An omnibus test to test the significance of the main effect of tempo (irrespective of group) and any interaction (in case the effect differs by culture) is an alternative, but maybe not the best.

Second, treating a 5-point Likert scale as if it was normal is problematic. Not only it is a discrete variable, but it also involves a finite set of possible values. This should be modelled as an ordinal scale. For this, there are several options, including ordinal logistic regression.

However, there is also the problem that (as the authors mention), there are 6 paired responses from each participant. This could be addressed by using a generalised mixed model, with random effects for each participant. For an ordinal dependent variable, this could be a Cumulative Link Mixed Model. In R, this can be achieved, for example, using the clmm function from the ordinal package.

Finally, in their response to the Triage at pre-screen, the authors mention that they could not find an appropriate test that would also allow to perform the appropriate matching power analysis. This is absolutely true, but there is always the possibility of doing a simulation-based power analysis, which I attempted and will try to summarise:

In this case, I simulated a population of 60,000 values, in two conditions (A and B, which could be low and high tempo), from binomial distributions in 5 attempts (meaning each "person" could get a score from 0 to 4), but I modified the likelihoods, so that conditions A and B have different distributions, and added 1 to each value so that the possible results range from 1 to 5. Then, I randomly assigned a participant ID to 6 values for both the A and B conditions. Finally, to match the authors assumptions, I made sure that the (paired) difference between conditions was d = 0.4 (Cohen's d is NOT an appropriate effect size in this case, but it serves the point).

Then, I simulated 1,000 samples from that population, and tested the power reached with different sample sizes (basically, the number of simulations in which the p value was below α), using 4 different statistical tests/models:

- t-test (not considering the multiple responses from each participant): with a sample size of 82, a power of 0.946 was obtained
- Wilcoxon signed-rank tests (not considering the multiple responses from each participant): with a sample size of 83, a power of 0.951 was obtained.
- Linear Mixed Model (considering the multiple responses from each participant): with a sample size of 14 participants (6 paired responses per participant), the power was 0.961. These models were fitted with the call: lmer(Value ~ Condition + (1 | Participant))
- Cumulative Link Mixed Model (considering the multiple responses from each participant): while I was able to fit the model on single samples, I was not able to run 1,000 simulations as before (or even more that 3, for reasons I have not yet understood) and quite literaly ran out of time to submit my review. So, sadly, I haven't managed to run a proper simulation-based power analysis for Cumulative Link Mixed Models, but with more time (or, even better, the help of a statistician) this would definitely be possible. While I am not an expert, and I need time to make progress in this, I would be happy to assist the authors in this if needed.

This simulation, including all the code, is attached as an RMD file.

We are extremely grateful for this - this is going far beyond the normal call of duty for a reviewer! Indeed, Dr. Leongómez has kindly agreed to join as a coauthor to implement these suggestions.

After he completed running the simulation with the Cumulative Link Mixed Model, we confirmed that this estimated a required sample size of 540 paired responses from 72 participants total (24 per society for each of the three societies) to obtain a power of at least 0.95 to test our two hypotheses, while controlling for multiple comparisons. Since as noted in his review this model is most theoretically appropriate for our data given it is ordinal Likert scale data with multiple responses per participant, we chose to implement this model in our completely revised statistical analysis plan.

Minor suggestions

P.1, second paragraph of the introduction: I think "and" is missing in "Audiovisual associations are shown to be mediated by "psychological and socio-cultural" elements

(Taruffi and Kussner, 2022), musical training (Kussner and Leech-Wilkinson, 2014), and language (Dolschied et al., 2022).”

Thank you for catching this. It is now fixed in the manuscript as below:

> **These cross-modal associations are shown to be mediated by factors such as musical training (Kussner and Leech-Wilkinson, 2014) and language (Dolschied et al., 2022).**

P.1, second paragraph of the introduction: Probably the authors meant to say tempo instead of time.

By time, we meant the representation of time in a musical excerpt. Research has found that this passage of time has been visualized mostly from left to right in participants who were familiar with western musical notation (Athanasopoulos and Moran, 2016). We have added “the representation of” to clarify it for the reader.

> **Additionally, research about music-based visual imagery has examined how musical features such as pitch, volume, and the representation of time in music are associated with visual imagery, and revealed correlations between musical and visual features such as pitch and spatial height (e.g. Dolschied et al., 2022; Athanasopoulos and Moran, 2013; Kussner 2014; Tan and Kelly, 2004; Rusconi et al., 2006; Eitan and Timmers, 2010), and horizontal time representation (Athanasopoulos and Moran, 2013; Kussner, 2014; Walker, 1987).**

P.2, second paragraph: The authors provide information about dimensional models, but not about categorical models. I suggest adding a short description/example with relevant citations.

Thank you for pointing this out. We included a short description and an example in the manuscript.

> **Categorical emotions, on the other hand, are based on the basic emotion model (Darwin, 1872; Ekman, 1992; Izard, 2007) which consists of six or higher categories of emotion such as sadness, anger, happiness, fear, and disgust. Prior research has shown several categories of emotions which could mostly be detected by the listeners (Juslin and Laukka, 2004; Sloboda, 1992; Wells and Hakanen, 1991; Zentner et al., 2008).**

P.7, second paragraph: “Texture one consists of one horizontal line in a circle. (Fig 2) The subsequent textures…”. I think the period should be after the citation of Fig. 2

Thank you for catching this. It is now fixed in the manuscript.

Texture one consists of one horizontal line in a circle (Fig 2).

References

Makin, T. R., & Orban De Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. ELife, 8, e48175. https://doi.org/10.7554/eLife.48175

Review #2 (Nadine Dijkstra)

1A. The scientific validity of the research question(s)

This study aims to investigate the relationship between music, emotion and visual imagery across three different countries. This research question is interesting and bring together different fields. However, the scientific motivation behind the research questions is not always made entirely clear. For example, the sentence "By distinguishing between universality and culture-specificity of these associations, we aim to understand whether relationships between music, emotion, and visual imagery are cross-culturally universal or culturally specific." is circular.

We thank the reviewer for catching this. It is now revised in the manuscript as below. We have also added more explanation about the underlying ideas for the study.

> **The aim of this study is to understand whether relationships between music, emotion, and visual imagery are universally consistent or culturally specific.**

Another example is the question regarding 'visual density'; the authors state that investigating this "will add to our understanding of audio-visual associations…" but do not give any motivation for why this is the case. Why is this feature important to investigate and how exactly does it add to previous research on different visual features? Also, please explicitly define 'visual density' and 'solo and group excerpts' in the introduction.

Thank you very much for this suggestion. We have added definition of 'visual density' and more details about 'solo and group excerpts' in the manuscript, as follows:

Solo pieces involve a solo instrument and group pieces involve multiple instruments playing simultaneously (Japan: *matsuribayashi* (festival ensemble), Iran: Kamancheh and Dohol, Canada: fiddle, banjo, mandolin, bass, and guitar).

More broadly, why is the relationship between music, culture, emotion and visual imagery important to characterize outside of the context of previous studies on this exact topic? Why is it important in general? Some of this is discussed in the fourth paragraph of the introduction where some more concrete findings regarding this relationship are mentioned. Tackling this earlier on might help in making the general motivation clearer.

Thank you very much for this valuable insight. We have updated the first paragraph's last sentence and continued with more explanation to clarify the context a bit better as seen in our response to the editor's comments about the study rationale.

Addition to the introduction:

> **However, few studies (e.g., Palmer, 2013) have investigated the relationship between musical features, visual imagery, and emotions to understand universality and diversity in cross-modal associations, even though visual imagery is thought to be one of the underlying mechanisms for how musical emotion is induced (Juslin and Västfjäll, 2008).**

1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

The motivation for the first hypothesis is clearly based on previous studies about the relationship between tempo and emotional arousal. The rationale for the second hypothesis is less clear to me, not least because the concept of 'visual density' has not been defined explicitly in the introduction. Furthermore, both hypotheses state that these relationships are universal but, as far as I understood, this is an empirical question and there is no clear evidence given to a priori expect universality or specificity between cultures.

Thank you very much for your comment. The hypotheses are based on some of the previous research and our intuitions about music to visual density associations; however, these are hypotheses in need of empirical testing, and might not be supported after collecting and analyzing Stage2 full data.

Regarding the concept of visual density, we have added the following description to the manuscript:

> **Musical texture can refer to melodic, rhythmic, and harmonic layering of musical material (Sarrazin, 2016) and is often conceptualized with notions that describe visual texture such as dense/sparse, thick/thin, rough/smooth, etc. On the other hand, visual textures are either plane or volumetric and a plane simple visual texture is characterized by three dimensions of directionality, size, and density. (Caivano, 1990) Density, as Caivano (1990) states, "depends on the relation between the texturing elements and the background". By examining the association between rhythmic texture and density of the musical and visual stimuli, we aim to investigate these understudied aspects of cross-modal mappings that could add to our understanding of music-induced visual imagery and music cognition.**
>
> **In our study, density is presented through horizontal straight lines against a white blank background and ranges from low to high on a scale from 1-5. For example, a higher number of lines against the background is equal to a higher density and vice versa. Visual density has yet to be empirically examined in cross-modal studies, however, it can provide some insight into music cognition across individuals and cultures since we use metaphors from visual textures to conceptualize musical complexity.**

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

The proposed analyses are t-tests but this is a complicated design with more than 2 groups/measurements to compare. The authors added a foot-note for other potential analyses but have not provided any motivation for why these other analyses might be more suitable and which of them they will actually use. Some of the specific interpretations of the analysis outcomes are also incorrect. For example, "If our predicted effect is significant in each of the three cultural groups, we will conclude that tempo-arousal relationships are cross-culturally general". However, it is possible that the relationship is significant in each group but that there are still significant differences between the groups (e.g. the relationship is stronger in one culture than another), indicating cultural specificity. The next sentence regarding interpretation of a null-finding is also incorrect: the relationship might be significant in some but not other groups but if there is no significant difference between the groups you cannot claim cultural specificity. More generally, absence of evidence for an effect (i.e. a null-finding, or 'statistically equivalent') obtained with frequentist statistics does not provide evidence for the absence of an effect. I am also surprised by the small sample sizes proposed (N = 14 per group), especially since the pilot data already consisted of

N = 9 per group and given that it is an online study. Taken together, I unfortunately have to conclude that the proposed analysis pipeline is not very sound.

We thank the reviewer for pointing out these issues. Our revised n=72-participant power analysis building on the one proposed by Rev #1 addresses the issues (see above).

PS Though it is no longer relevant, note that "statistically equivalent" is not the same as a traditional "null finding" (i.e., p>0.05). The equivalence testing we originally planned is a legitimate method for testing for absence of a meaningful effect using frequentist statistics
(cf. Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4), 355–362. https://doi.org/10.1177/1948550617697177).
Also note that our pilot data from 9 participants represents 3 participants per group, not 9 per group).

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses
The methodological details are mostly clear but for exact replication the musical excerpts and experimental code need to be shared. Furthermore, there is some unclarity and flexibility in the proposed statistical analysis (also see previous question).

Thank you very much for your suggestion. We have created an OSF repository (separate to our previous GitHub data/code repository), uploaded our musical stimuli, and added the link to the "Data/code availability statement" as follows:

> **Data/code availability:**
> **Our data and code are available at https://github.com/comp-music-lab/VisualEars**
> **Our musical stimuli are available at**
> **https://osf.io/pkvw2/?view_only=0894653041ba4375afdcaa3a1989fe71**

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

The tempo of the excerpts is manipulated to be 20% slower or faster than the original. Is there a control that can ensure that these manipulations do not influence other aspects of the music, independent of tempo per se, such as the clarity of the lyrics, which might influence the results? Furthermore, the solo pieces were slower than the group pieces, so any difference between these two is likely due to tempo rather than solo vs group. Relatedly, the pieces from the different countries also all have different tempos, which means that any difference here might also be due to tempo rather than country. In the absence of these controls, it is hard to draw clear conclusions about the effects of these variables.

We thank the reviewer for this comment. We have modified the manuscript as follows to clarify our choices:

> **The tempo of all 6 excerpts is separately manipulated in a way that it is lowered and raised by 20% for each excerpt. By comparing two excerpts manipulated equal amounts from the original rather than comparing the original with a modified version, both versions are equally different from the original and thus any differences due to the piece being heard faster/slower than intended should cancel each other out (if we compared the original with a modified version, the variable tempo would become confounded with the presence/absence of tempo manipulation). We selected these specific tempo after finding in our preliminary pilot analyses that these provided the optimal balance between maximizing acoustic differences while minimizing audible recording artifacts created by the manipulation process.**
>
> **Note that, because 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).**

Review #3 (Elena Karakashevska)

(Some of the comments in the review below were addressed in round 1 and are marked as resolved in this letter.)

This is a potentially interesting paper. Authors have noticed the lack of research in cross-cultural links between visual imagery, music and emotion. The study however needs some major improvements. The power analysis is something that needs work on, I think with 14 participants in each group the study is underpowered massively. Especially taking in consideration it is conducted online.

The authors aim to add to the literature by doing a cross cultural study to test for differences in emotional arousal and perception of density of visual imagery, across cultures by manipulating tempo in solo and group performance pieces. The authors aim to conduct this study online and compare within subjects effects only.

Major issues

The analysis plans

The analysis plan involves multiple t-tests. I don’t think considering the data and the hypotheses that this is the most suited statistical test. Testing 3 times for an effect in the same pool of data is not wise.

The authors can encompass everything they are hypothesising for in a single analysis. Perhaps two mixed ANOVAs would be best. The cultural group can be a between subjects factor and tempo a within subjects factor in each ANOVA looking at emotional arousal and visual imagery respectively. I understand you are dealing with ordinal data but this shouldn’t be a major problem as ANOVAs are commonly used for this.

The hypotheses

Our hypotheses (Table 1) are listed as:
1) Increasing tempo consistently increases emotional arousal across cultures. Arousal ratings and tempo changes are positively correlated in Iran, Japan, and Canada.
2) Increasing tempo consistently increases density of visual imagery across cultures. Visual density associations and tempo changes are positively correlated in Iran, Japan, and Canada.

These are well conveyed; however, there are no plans to test the first hypothesis? If the first hypothesis states that the authors predict a correlation, they need to state how they will test for a correlation. In the table the first hypothesis is worded differently, yet the plan for both of the hypotheses is the same.

We thank the reviewer for this valuable feedback. We have now unified the wording here and in the table, removing redundant terminology and removing confusing reference to correlations, as follows:

> **1) Increasing tempo consistently increases emotional arousal across cultures.**
>
> **2) Increasing tempo consistently increases density of visual imagery across cultures.**

The methods

Can the authors please state the length of the experiment, since this will dictate whether two attention checks are sufficient. Stating the number of trials and the average duration of trial + response time would give a good estimate of length.

Please add details about the visual stimuli used in the experiment, such as the dimensions of the patterns, how they are generated, etc. **(Resolved)**

We thank the reviewer for pointing this out. We added the following details about the visual stimuli in the pre-screening process (including moving an image of the stimuli from supplementary material to Fig. 2 in the main manuscript). We have also added the following clarification to the manuscript:

> **"The experiment takes about 30-40 minutes."**
>
> **They also select an image of a visual texture (adapted from Langlois et al., 2014) which they think most represents the excerpt on a scale of 1-5. 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. Texture one consists of one horizontal line in a circle. (Fig 2). The subsequent textures incorporate a steady increase in the number of horizontal lines leading to a steady increase in visual density levels. Y=2X+1 represents how the density increases, where Y is the number of lines in the next texture and X is the number of lines in the previous texture. These visual textures are generated through digital drawing.**

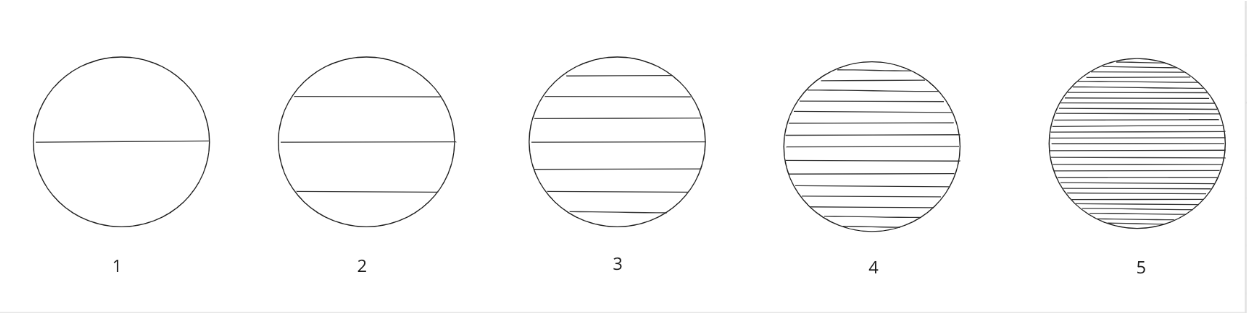

Fig 2. Adapted from the stimuli by Langlois et al. (2014).

Can the authors please confirm if the effect size is 0.04 or 0.4. I assume the first is a typo in the power analysis section.

Thank you for catching this typo. We had intended to write 0.4. However the revised power analysis required different effect size statistics, so neither .04 nor 0.4 appears in the revised manuscript.

The power analysis

The biggest issue I have is the power analysis. I'm unsure about this power analysis and why it's based on responses.

Our thoroughly revised statistical analysis should resolve these issues. However, the reviewer appears confused about some aspects of power analysis, so for completeness we will clarify below (even though most of these points are no longer relevant).

I'm also unsure how the authors concluded that 84/6 = 14, not sure where this comes from, so an explanation is welcome.

84 paired responses divided by 6 paired responses per participant does equal 14 participants. (The non-independence issue with multiple responses per participant is now addressed in Dr. Leongomez's suggested mixed model analysis approach). Cf. Brysbaert (2019) section "Increasing the power by having multiple observations per condition per participant".

Assuming the effect (dz = 0.4) found in the pilot study is correct, running a simple G*power analysis using liberal constraints (alpha = 0.05, power of 0.8, one-tailed) on this comes to needing 41 participants for a paired samples t-test.

Our SESOI of d=0.4 was not estimated from pilot data, but based on previous recommendations. Cf. Brysbaert (2019) for why "Pilot studies are next to worthless to estimate effect sizes".

The authors are conducting an online survey and settling to 14 participants per group, when they're comparing means is simply not sufficient. With the methods of data collection, there is resources to recruit a larger sample and have more power to detect any differences.

In addition, the authors have decided to use liberal constraints like a one-tailed hypothesis testing which obviously reduces your need for a large sample. I understand they have a one-tailed prediction based on their pilots but it is good practice to keep it two tailed for power analysis.

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.

According to Brysbaert (2019) basing effects on such a single pilot study is ‘almost worthless’, especially since the pilot itself had a small n.

Correct, that is why we did not estimate our effect size based on our pilot data (see above).

Note that for our specific paradigm, each participant produces 6 paired responses (one pair for each of the 6 musical stimuli), and each cultural group will be tested separately, so we will need at least 84 /6 = 14 participants x 3 = 42 participants total across all three cultural groups.
If the authors concluded that the least number plausible is 14 per group, they should reconsider and adjust their sample as they are recruiting participants online and have the possibility to recruit more.
Again, collecting more data than needed is a waste of limited time and resources.

Wording issues
By making a distinction between cross-cultural consistency and diversity in these correspondences and emotion appraisals, we aim to understand whether 1) we can find any innate/physiologically derived connection that could explain our cross-modal associations, and 2) tempo mediates emotion appraisals and visual imagery.
It is not clear to me how this study will contribute towards understanding innate cross-modal associations. Can the authors please expand on this?

Thank you for your valuable feedback. We also think “innate” does not correctly reflect the aims and scope of our research, and consequently have revised the manuscript as below in all instances where “innate” was mentioned.

> **By making a distinction between cross-cultural consistency and diversity in these correspondences and emotion appraisals, we aim to understand whether 1) we can find any connections between the variables that could explain our cross-modal associations, and 2) tempo mediates emotion appraisals and visual imagery.**
>
> **Abstract:**
> **However, few studies have simultaneously investigated cross-cultural links between music, visual imagery, and emotion in order to distinguish the role of cultural experiences in contrast to more widespread perceptual capabilities.**
>
> **Page 2 paragraph 1:**

> **One way to empirically tackle these ideas is to conduct cross-cultural studies and measure both emotional models in order to gain insight into the mechanisms of emotion detection across cultures.**

Cross-modal associations, more precisely audiovisual associations for the sake of our study, are found to be mediated by language, music training, and in the case of music to colour associations, mediated by emotion. However, the mechanisms for these associations are still unclear. Several metaphors that are typically used for visual texture description (dense/sparse, thin/thick, rough/smooth) are also used to describe musical complexity e.g. rhythmic density which is higher when there are higher number of subdivisions in a beat and vice versa. One hypothesis, if found cross-culturally consistent, is the conceptual metaphor theory which refers to the understanding of one abstract idea in one domain through conceptualizing in another domain. In this case, we could hypothesize that musical tempo is understood through metaphors such as density that are also applicable to the visual domain.

Minor points

In the abstract the authors state their hypotheses first, then what they did in the study.

We hypothesize that there are cross-culturally consistent correlations between tempo changes and 1) visual density associations, and 2) arousal ratings. In this study, we investigate the relationship between emotional arousal and visual density induced by 6 musical excerpts differing in tempo and texture (solo vs group) in participants in Japan, Iran, and Canada. By distinguishing between universality and culture-specificity of these associations, we aim to understand whether relationships between music, emotion, and visual imagery are cross-culturally universal or culturally specific.

I think it would make more sense to the reader if this was reworded to state the current study then the hypotheses.

Thank you for pointing this out. We have swapped the order in the manuscript and revised the last sentence.

> **In this study, we investigate the relationship between emotional arousal and visual density induced by 6 musical excerpts differing in tempo and texture (solo vs group) in participants in Japan, Iran, and Canada. We hypothesize that there are cross-culturally consistent relationships between tempo changes and 1) visual density associations, and 2) arousal ratings. The aim of this study is to understand whether relationships between music, emotion, and visual imagery are universally consistent or culturally specific.**

Minor mistakes in ordering of references in-text in the introduction on page 1. Please double check these.

Thank you for catching this. It is now fixed in the manuscript.

Cross-modal audio-visual associations have been extensively studied and
revealed evidence for consistency between musical elements and visual features

There is no need to say cross-modal and audio-visual when they essentially mean the same thing in this sentence.

Please be consistent in the spelling of audio-visual. In places there is no hyphen.

Thank you for catching this. We have fixed this throughout the manuscript.

Audiovisual associations are shown to be mediated by "psychological and socio-cultural" elements (Taruffi and Kussner, 2022), musical training (Kussner and Leech-Wilkinson, 2014), language (Dolschied et al., 2022).

Unsure why "psychological and socio-cultural" are in quotation marks here. Can the authors please explain or perhaps mention what these elements consist of.

Thank you for the review for this excellent catch. Since we are not considering those elements in our study, we decided to revise the sentence to prevent confusion. The updated sentence is as below:

> **These cross-modal associations are shown to be mediated by factors such as musical training (Kussner and Leech-Wilkinson, 2014) and language (Dolschied et al., 2022). It is worth noting that implicit associations are very useful in cross-cultural studies as translations and naturally occurring connotations around language can be misleading and are preferably avoided. (Athanasopoulos, 2022)**

Importantly, most studies in the music cognition and perception literature incorporate mostly Western music and Western participants, and there is a need to test these findings in other cultures as well (Jacoby et al., 2020).

Sentence would read better if you remove the 'as well' from the end.

Thank you for your suggestion. It definitely reads better now.

> **Importantly, most studies in the music cognition and perception literature incorporate mostly Western music and Western participants, and there is a need to test these findings in other cultures (Jacoby et al., 2020).**

Our choice of countries is based on our access to the local communities and having native speakers as coauthors who can facilitate the process of data collection.
co-authors. Not sure having native speakers as co-authors is relevant enough to be in the manuscript. Having access to the data encompasses that.

Thank you for your comment. We added the following to the manuscript.

> **Since the surveys and scales we are using are not always translated to foreign languages, the co-authors will translate and help in data collection by reviewing the translations and making sure the study is clear and feasible in their native language.**

We delve deeper into this relationship through a comparative experiment in Japan, Iran, and Canada
Would be good to state the relationship prior to this claim, or within the sentence

Thank you for your suggestion. It is now revised in the manuscript.

> **We delve deeper into this relationship between musical tempo, visual density, and emotional arousal through a comparative experiment in Japan, Iran, and Canada to discover the similarities and differences across these three seemingly different populations; countries with distinct language, cultural practices, scripts, and musical cultures.**

We selected these specific tempo after finding in our preliminary pilot analyses that these provided the optimal balance between maximizing acoustic differences while minimizing audible recording artifacts created by the manipulation process.
This is a bit confusing for me to understand. Can you please explain in a clearer manner how you reached the conclusion that this selection of tempi is the optimal balance?

Thank you for your comment.

> **The musical stimuli were manipulated in Audacity and various tempo changes (10-50%) were examined in order to find the optimal change that was noticeable enough and natural-sounding. There was a consensus among the co-authors that tempi manipulations greater than 20% had excessive sound artifacts while manipulations less than 20% were not different enough from the original.**

Can authors please clarify if any of the authors took part in the pilot study from which the power analysis is based on? This can potentially have confounding effects.

Thank you for your attention to this matter. All except two of the coauthors participated in the pilot experiment, however, the pilot data will not be used in

the study after In Principle Acceptance (as required by PCI-RR) and the effect sizes calculated in the power analysis were not calculated based on the pilot data, so any confounds affect only the pilot data and will not affect our actual Stage 2 data/analyses.