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 the plan for both of the hypotheses is the same.

*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.

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.

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

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.

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

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.

*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?

**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.

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

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

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.

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.

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.

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

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?

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.