UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Chiba, G., Ozaki, Y., Fujii, S., & Savage, P. E. (2023). Sight vs. Sound Judgments of Music Performance Depend on Relative Performer Quality: Cross-cultural Evidence From Classical Piano and Tsugaru Shamisen Competitions. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.73641

Cognitive Psychology

Sight vs. Sound Judgments of Music Performance Depend on Relative Performer Quality: Cross-cultural Evidence From Classical Piano and Tsugaru Shamisen Competitions

Gakuto Chiba[1], Yuto Ozaki[1a], Shinya Fujii[2], Patrick E. Savage[2b]

[1] Graduate School of Media and Governance, Keio University, Fujisawa, Japan, [2] Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan

Keywords: music, performance, multimodal, auditory, visual, cross-cultural

https://doi.org/10.1525/collabra.73641

Collabra: Psychology

Vol. 9, Issue 1, 2023

Which information dominates in evaluating performance in music? Both experts and laypeople consistently report believing that sound should be the most important domain when judging music competitions, but experimental studies of Western participants rating video-only vs. audio-only versions of 6-second excerpts of Western classical performances have shown that in at least some cases visual information can play a stronger role. However, whether this phenomenon applies generally to music competitions or is restricted to specific repertoires or contexts is disputed. In this Registered Report, we focus on testing the generalizability of sight vs. sound effects by replicating previous studies of classical piano competitions with Japanese participants, while also expanding the same paradigm using new examples from competitions of a traditional Japanese folk musical instrument: the Tsugaru shamisen. For both classical piano and Tsugaru shamisen, we ask participants to choose the winner between the 1st- and 2nd- placing performers in 5 competitions and the 1st-place and low-ranking performers in 5 competitions (i.e., 40 performers total from 10 piano and 10 shamisen competitions). We tested the following three predictions twice each (once for piano and once for shamisen): 1) an interaction was predicted between domain (video-only vs. audio-only) and variance in quality (choosing between 1st and 2nd place vs. choosing between 1st and low-placing performers); 2) visuals were predicted to trump sound when variation in quality is low (1st vs. 2nd place); and 3) sound was predicted to trump visuals when variation in quality is high (1st vs. low-placing). Our experiments (*n* = 155 participants) confirmed our first predicted interaction between audio/visual domain and relative performer quality for both piano and shamisen conditions, suggesting that this interaction is cross-culturally general. In contrast, the second prediction was only supported for the piano stimuli and the third prediction was only supported for the shamisen condition, suggesting culturally dependent factors in the specific balance between sight and sound in the judgment of musical performance. Our results resolve discrepancies and debates from previous sight-vs-sound studies by replicating and extending them to include non-Western participants and musical traditions. Our findings may also have practical applications to evaluation criteria for performers, judges, and organizers of competitions, concerts, and auditions.

1. Introduction

Music is often defined primarily in auditory terms (e.g., “humanly organized sound”; Blacking, 1973). Indeed, sound is consistently reported to be the most important information for evaluating musical performance (Murnighan & Conlon, 1991; Sloboda et al., 2008). Yet there is also a rich literature across fields and methodological traditions showcasing the recognition that music is a multimodal phenomenon (Bergeron & Lopes, 2009; Leman, 2008; Savage et al., 2021; Vines et al., 2006). For example, visuals play an important role in evaluating musical performance, with elaborate costumes, make-up, and dancing characteristic

[a] Gakuto Chiba and Yuto Ozaki are equal-contribution first authors.

[b] Please direct correspondence to psavage@sfc.keio.ac.jp

of both traditional and contemporary music performance (Nettl, 2015). The popular international song competition is called "Eurovision", not "Eurosound" (cf. Haan et al., 2005).

Not only do visuals have the power to affect how it is that we hear the most basic aspects of musical sound (Thompson & Russo, 2007), visuals can also have societal consequences for hiring practices and issues of equity. In a seminal paper that has spurred policy changes, economists found that after the implementation of blind auditions by orchestral organizations, significantly more female musicians were hired (Goldin & Rouse, 2000). These findings underline how much the presence of visuals altered evaluations made of musicians and their performances.

Experimental evidence demonstrating cross-domain effects of visual information on auditory perception in music has accumulated over the past few decades and continue to spur interest across fields (Goebl & Palmer, 2009; Platz & Kopiez, 2012, 2013; Schutz & Lipscomb, 2007; Tsay, 2013, 2014; Wapnick et al., 1998). Although the findings regarding cross-modal influences from work in music are consistent with those of evaluations made across a range of domains beyond music (Campanella & Belin, 2007; Collignon et al., 2008; de Gelder et al., 1999; McGurk & MacDonald, 1976), there is debate about the relative effects of the roles of visuals vs. sound in music competitions and how general such effects may be. For example, two studies of Western classical music competitions came to contrasting conclusions regarding the roles of sight vs. sound: Tsay (2013) argued that "people actually depend primarily on visual information when making judgments about music performance", while Mehr et al. (2018) concluded from direct and conceptual replications of Tsay's study that "the sight-over-sound effect holds only under limited conditions". Yet reanalysis of Mehr et al.'s data suggests alternative possible interpretations (see below), and the generalizability of sight vs. sound effects beyond specific Western classical traditions and Western participants remains untested despite being arguably a question of even greater importance (Jacoby et al., 2020).

1.1. Re-analysis of Mehr et al. (2018)'s "failure to replicate" Tsay (2013)

Tsay (2013) found that, when choosing between 6-second excerpts of the 1st, 2nd, and 3rd-place performers in classical piano competitions, participants were able to choose correctly 46% of the time when watching silent videos without audio, compared to only 28% accuracy when listening to audio only without video (Tsay, 2013 Experiment 3).

Mehr et al. (2018) conducted a direct replication using mostly the same stimuli as Tsay (2013) Experiment 3 (9 of the 10 original sets of 1st-3rd placed performers), which they successfully replicated albeit with slightly weaker results (39% accuracy with video-only vs. 30% with sound-only; data plotted in Fig. 1a). Mehr et al. also conducted two conceptual replications using different stimuli, which they argued represented a "failure to replicate" Tsay's findings. However, Mehr et al. did not actually plot their data and relied only on selected statistical comparisons to argue that their results failed to replicate Tsay's. Specifically, they interpret the fact that video-only accuracy was not significantly above chance (50% in their modified design using only 1st and 2nd-place performances, rather than 33% in the original design using 1st-3rd place) as failure to replicate sight-over-sound effects. Yet when their data are visualized, it is clear that their Study 2 results (51% accuracy with video-only vs. 45% with audio-only) are qualitatively very similar to their Study 1 results (39% vs. 30%, respectively; Fig. 1b). Throughout their analyses, Mehr et al.'s only reported inferential statistics are one-sample t-tests comparing accuracy in each condition to chance, and do not report the statistics more theoretically relevant for sight-over-sound effects - namely the two-sample t-tests reported previously by Tsay (2013). When Mehr et al.'s data are reanalyzed using two-sample t-tests, both Study 1 and Study 2 replicate Tsay's finding of greater accuracy with video-only vs. audio-only (Study 1: t = -4.5, Cohen's d = 0.57, df = 243, p = 9.9 x 10^{-6}; Study 2: t = -3.0, Cohen's d = 0.42, df = 185, p (two-tailed) = 0.003). Thus, Mehr et al.'s claim that Study 2 failed to replicate Tsay's findings is inaccurate.

On the other hand, Mehr et al.'s claim that Study 3 failed to conceptually replicate Tsay is better supported by their data. Specifically, when differences in performance quality were made clearer by comparing the winning performer with lower-ranked performers rather than 2nd place performers, higher accuracy was found with audio-only (68%) than video-only (45%; Fig. 1c; t = 6.1, Cohen's d = 1.2, df = 98, p = 2.6 x 10^{-8}). Mehr et al.'s claim that "sight does not necessarily trump sound in the judgment of music performance" is thus clearly supported. However, this may be partially consistent with Experiment S3 in Tsay (2013), which found practically no difference in accuracy between video-only and audio-only performances when using stimuli from youth (pre-college) competitions where differences in quality are greater than found in professional competitions (Experiment S3-1: video-only 70% vs. audio-only 69%; Experiment S3-2: video-only 56% vs. audio-only 53%).

1.2. Study aims and hypotheses

To examine the generalizability of sight vs. sound effects in music performance, we will replicate previous studies using stimuli from Western classical music with Japanese participants and then repeat the same paradigm using stimuli from competitions on the Tsugaru *shamisen*, a traditional Japanese folk musical instrument that GC (first author) has experience performing as a national champion (https://www.gakuto-chiba.com/profile1).

The shamisen is a fretless chordophone (stringed instrument) similar to the Chinese sanxian, Arab oud, or European lute. Tsugaru shamisen is a folk shamisen genre traditionally played by blind folk musicians called "Bosama" in northeastern Japan (Daijo, 1995). In recent decades, Tsugaru shamisen has become popular among the general populace throughout Japan, even featuring in the popular 2016 animated movie "Kubo and the Two Strings". Importantly for our purposes, thousands of Tsugaru shamisen perform-

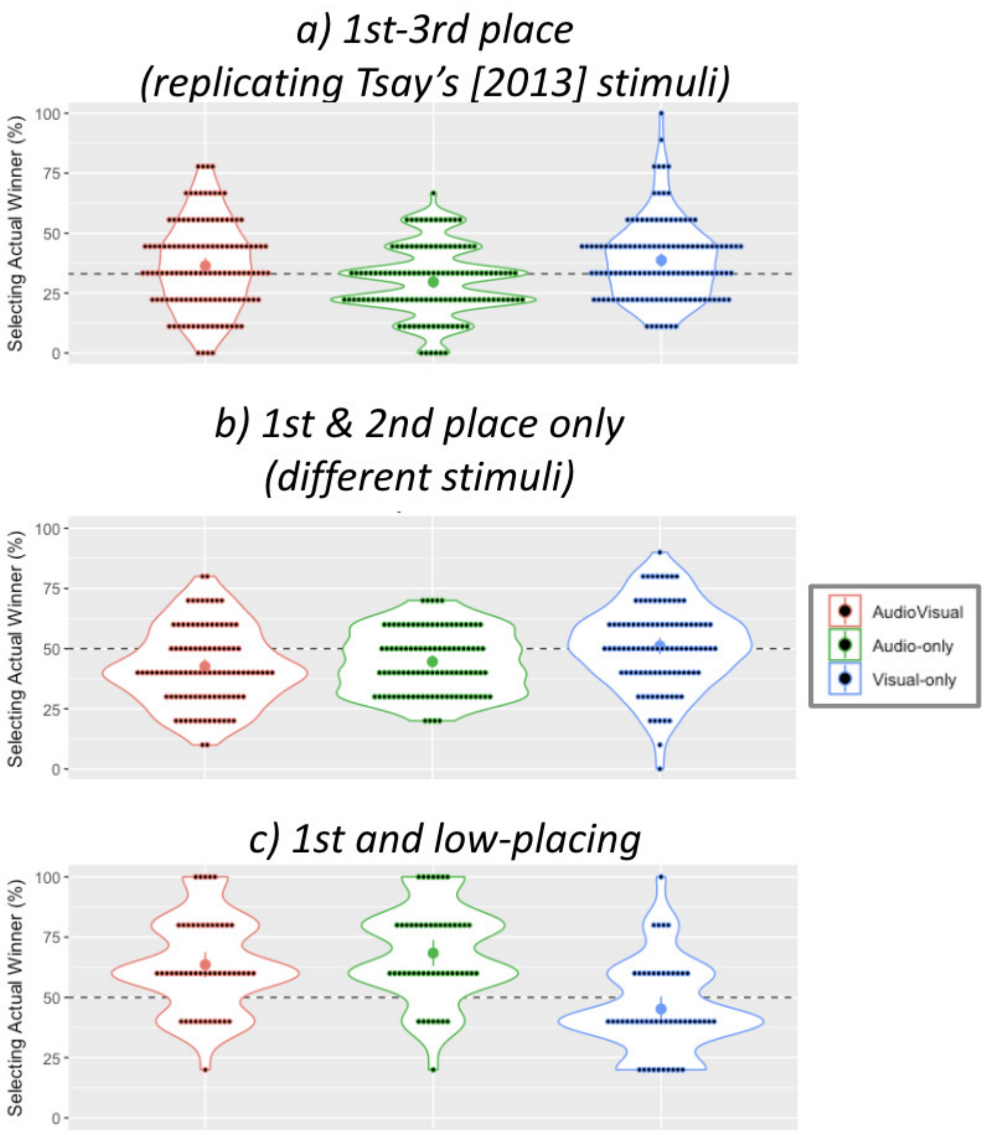

Figure 1. Violin plots visualizing Mehr et al.'s (2018) previous experimental results of sight vs. sound effects in judging piano performances (data were not visualized in the original publication). Panels a-c correspond to Studies 1-3 (see text for details).

Dots indicate individual participants (a: n=375 participants; b: n=300 participants; c: n=150 participants), large dots indicate means and bars indicate 95% confidence intervals. The colour legend indicates whether the 6-second excerpts participants played were audiovisual, audio-only, or visual-only. The y-axis indicates the percent of performers correctly choosing the winning performer. Dashed lines indicate chance levels (33% when choosing between 3 performers, 50% when choosing between only 2).

ers compete annually in dozens of regional, national, and even international competitions (Hughes, 2008). The large collection of recorded and ranked performances thus allows us to collect examples analogous to those from Western classical competition previously used in the experiments described above to allow direct comparison between Western classical competitions and competitions in a traditional non-Western folk genre.

1.3. Hypotheses

Based on previous findings from Western classical competitions described above (Mehr et al., 2018; Tsay, 2013), we made the following three predictions for piano and Tsugaru shamisen competitions (i.e., 3 predictions x 2 instrument types = 6 predictions total):

H1: We predict that there is an interaction effect between the modality factor (audio-only vs. video-only) and the quality variance factor (low vs. high variance) such that sight vs. sound effects depend on the performance quality gap of competitors. (Null hypothesis: sight vs. sound effects do not depend on the performance quality gap of competitors).

H2: We predict that visuals will dominate the judgment of piano performance among upper ranks (1st vs. 2nd place), due to low variance trials with relatively little differences in performance quality. (Null hypothesis: there is no difference between visual and audio judgments when variance in performer quality is low).

H3: We predict that sound will dominate the judgment of piano performance between upper and lower ranks (1st place vs. low-placing), where there are high variance trials with relatively greater differences in performance quality. (Null hypothesis: there is no difference between visual and audio judgments when variance in performer quality is high).

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. In the event that our predictions are not statistically significant, we will evaluate support for the null hypothesis through the use of relative effect sizes and confidence intervals, which are conceptually similar to parametric equivalence testing but can be applied to non-parametric data (see Methods).

2. Methods

We built upon standard designs of testing predictions of behaviors (Ambady & Rosenthal, 1993; Ballew & Todorov, 2007; Rule & Ambady, 2008; Todorov et al., 2005; Tsay, 2013, 2014, 2021) in a within-subjects experiment to maximize statistical power and interpretability. Our experimental design was based on the literature on thin slices of behaviors (Ambady et al., 2000, 2006; Ambady & Rosenthal, 1993), especially the studies of visuals vs. sound in music competition evaluation described above (Mehr et al., 2018; Tsay, 2013).

2.1. Stimulus choice

2.1.1. Confirmatory sample

To enable us to replicate and generalize previous studies we designed a paradigm that allowed us to compare our results as directly as possible with Tsay (2013) and Mehr et al. (2018) by having the same participants rate both piano and shamisen performance stimuli in the same experiment. However, each of the three paradigms reported in Mehr et al. used slightly different designs: Study 1 used 9 out of 10 sets of excerpts of three performers (1st-3rd place) previously used by Tsay (2013); Study 2 used 10 sets of only two performers; and Study 3 used 5 sets of 2 performers (see https://osf.io/6nx4d for details). As Mehr et al. explain, this meant that they could not conclusively determine whether differences in their results were due to differences in experimental design or differences in the independent variables of interest (i.e., audio vs. visual domain or high vs. low variance).

To avoid these confounds, we chose to unify our experimental design based on the paradigm with the smallest number of stimuli, namely the 5 pairs of performers used in Mehr et al.'s (2018) Study 3 (high-variance condition). We thus collected analogous 6-second excerpts of performances from 10 pairs of Tsugaru shamisen performers: 5 "high-variance" pairs (1st place and low-placing performers, as in Mehr et al. Study 3) and 5 "low-variance" pairs (1st and 2nd place performers, as in Mehr et al., 2018 Study 2). These performers were selected from different competitions so the 1st-place performers would not overlap between the high-variance and low-variance conditions. For all Tsugaru shamisen performers, GC (1st author) selected an excerpt from the same portion of the opening of the piece "Tsugaru Jongara Bushi", because it is the most famous piece among Tsugaru shamisen players, and it is a compulsory component of all competitions, which allows us to collect a large number of comparable samples.

To choose 5 "low-variance" pairs from the 9 1st/2nd place performers previously used by Mehr et al. and Tsay, we removed four pairs that seemed least appropriate to compare. These included:

- two sets of violin performances (all other performances were of piano and all our performances were also of a single instrument, Tsugaru shamisen)
- one set including a 4-second clip rather than a 6-second clip after audience applause was edited out
- one set including a 1st-place performer that overlapped with one of the sets used in Study 3.

Pilot experiments (see below) suggested that restricting the stimuli to only 5 of the 9 previously used by Tsay (2013, Study 3) and Mehr et al. (2018, Study 1) did not appear to change the main sight-over-sound result reported by both.

This gave us a full set of 40 performances from 20 competitions for our main confirmatory analyses: 5 low-variance piano, 5 high-variance piano, 5 low-variance shamisen, and 5 high-variance shamisen (Table 1).

2.1.2. Exploratory sample

Tsay (2013) and Mehr et al. (2018) used a between-subjects design where different participants independently rated audio-only, visual-only, or audio-visual stimuli, but the same participant did not evaluate different domains. However, to increase statistical power and comparability we designed ours to be within-subjects, so the same participant evaluates all examples across all domains. To eliminate the possibility of order effects by which participants' judgments of audio-only or video-only samples would be affected if they had previously seen the audiovisual condition, we chose to focus our confirmatory analysis only on the key conditions of interest - audio-only vs. visual-only - and present these stimuli first. For exploratory comparison, audiovisual examples were also included at the end of the experiment, but these are not included in our confirmatory hypothesis testing. (The order of stimuli within the audio-only/video-only block and the audiovisual block is randomly determined.)

Also, although we chose to use 1st and 2nd-place performers from Mehr et al.'s Study 1 in order to allow us to also compare with Tsay (2013) who originally reported these stimuli, we also added stimuli from Mehr et al.'s Study 2 in order to allow exploratory analysis of the effect of changing the precise stimuli used. To choose a matched set of 5 pairs from the original 10 prepared by Mehr et al., we again excluded violin performances and also excluded two sets that included partial overlap with the stimuli used in Experiment 1 (i.e., the 6-second excerpts only differed by including/excluding 1-2 seconds). Thus each participant evaluates a total of 50 6-second excerpts from 25 pairs (40 performances / 20 pairs confirmatory [Table 1], 10 / 5 exploratory), and each performance is evaluated in three dif-

Table 1. Overview of the experimental stimuli selected: 6-second excerpts from 40 performances from 10 Tsugaru shamisen competitions and 10 classical piano competitions (see https://osf.io/nqkv8/ for detailed metadata). Piano excerpts were previously used by Tsay (2013) and/or Mehr et al. (2018; cf. https://osf.io/6nx4d/).

ID	Instrument	Variance	Competition	Place	Video excerpt
1	Piano	Low	1997 Van Cliburn International	1st	https://osf.io/t6nvf/
2	Piano	Low	1997 Van Cliburn International	2nd	https://osf.io/py5d6/
3	Piano	Low	2002 International Franz Liszt	1st	https://osf.io/p8uy6/
4	Piano	Low	2002 International Franz Liszt	2nd	https://osf.io/f48kg/
5	Piano	Low	2005 International Franz Liszt	1st	https://osf.io/q859w/
6	Piano	Low	2005 International Franz Liszt	2nd	https://osf.io/psgct/
7	Piano	Low	2008 San Marino	1st	https://osf.io/ynxjk/
8	Piano	Low	2008 San Marino	2nd	https://osf.io/k2etj/
9	Piano	Low	2009 Van Cliburn International	1st	https://osf.io/mcb7w/
10	Piano	Low	2009 Van Cliburn International	2nd	https://osf.io/rxw7n/
11	Piano	High	2009 Van Cliburn International	1st	https://osf.io/yrb7j/
12	Piano	High	2009 Van Cliburn International	Semifinalist	https://osf.io/mbgtz/
13	Piano	High	2007 International Franz Liszt	1st	https://osf.io/v5j3a/
14	Piano	High	2007 International Franz Liszt	3rd	https://osf.io/dqbcv/
15	Piano	High	2010 San Marino	1st	https://osf.io/67c9f/
16	Piano	High	2010 San Marino	Earlier competitor	https://osf.io/j2zv4/
17	Piano	High	2013 Van Cliburn International	1st	https://osf.io/vb4jq/
18	Piano	High	2013 Van Cliburn International	Preliminary competitor	https://osf.io/6rnuy/
19	Piano	High	2011 International Franz Liszt	1st	https://osf.io/dg2wy/
20	Piano	High	2011 International Franz Liszt	Semifinalist	https://osf.io/g7v3e/
21	Shamisen	Low	2019 Michinoku (general women)	1st	https://osf.io/cywh2/
22	Shamisen	Low	2019 Michinoku (general women)	2nd	https://osf.io/ydwcn/
23	Shamisen	Low	2019 Michinoku (general men)	1st	https://osf.io/gk7qe/
24	Shamisen	Low	2019 Michinoku (general men)	2nd	https://osf.io/rxsdg/
25	Shamisen	Low	2019 Biwako (boys and girls)	1st	https://osf.io/jg4x9/
26	Shamisen	Low	2019 Biwako (boys and girls)	2nd	https://osf.io/8bhvy/

27	Shamisen	Low	2019 Biwako (senior)	1st	https://osf.io/gcpe6/
28	Shamisen	Low	2019 Biwako (senior)	2nd	https://osf.io/y3m6f/
29	Shamisen	Low	2019 Hirosaki (personal B)	1st	https://osf.io/5fjy6/
30	Shamisen	Low	2019 Hirosaki (personal B)	2nd	https://osf.io/ntd2h/
31	Shamisen	High	2019 Michinoku (junior high school and high school students)	1st	https://osf.io/5vbjt/
32	Shamisen	High	2019 Michinoku (junior high school and high school students)	8th	https://osf.io/nsjmy/
33	Shamisen	High	2019 Biwako (general women)	1st	https://osf.io/b3j72/
34	Shamisen	High	2019 Biwako (general women)	21~47th	https://osf.io/x5hs2/
35	Shamisen	High	2019 Biwako (beginner)	1st	https://osf.io/p5uca/
36	Shamisen	High	2019 Biwako (beginner)	21~50th	https://osf.io/48tb2/
37	Shamisen	High	2019 Hirosaki (youth C)	1st	https://osf.io/dzxys/
38	Shamisen	High	2019 Hirosaki (youth C)	9~57th	https://osf.io/p26j8/
39	Shamisen	High	2019 Hirosaki (senior C)	1st	https://osf.io/fn4cr/
40	Shamisen	High	2019 Hirosaki (senior C)	8~31th	https://osf.io/8m7a6/

ferent formats; audio-only (confirmatory), video-only (confirmatory), and audiovisual (exploratory, saved for after the randomized audio-only/video-only block). This gives 50 excerpts x 6 seconds x 3 domains = 15 minutes worth of stimuli. This took pilot participants approximately 45 minutes to listen/watch and evaluate. The full pilot experiment can be accessed at https://gakuto101207.github.io/.

2.2. Independent variable

We have two independent variables: 1) stimulus domain (Audio-only vs. Visual-only [plus Audio-Visual for exploratory analysis]) and 2) the ranking gap of two performers as a proxy of the variance in their performance quality (High-variance and Low-variance). As a factorial design analysis, our experiment belongs to the repeated measures two-factor crossed design (domain × variance) where each factor has two factor-levels. Incidentally, studying the interaction effects brought by musical instrument/genre (Western classical piano vs. Japanese folk Tsugaru shamisen) is not within the scope of our hypotheses so this is not counted as a factor, but we will add this into our factorial design model in the exploratory analysis.

2.3. Dependent variable

The dependent variable will be the percentage of participants correctly choosing the 1st-placed performer in a two-choice forced-choice paradigm. As described above, participants will be asked to choose the actual 1st-place winner five times in each domain × variance combination. Therefore, the dependent variable will be metric discrete data taking values of 0.0 (no correct choices), 0.2, 0.4, 0.6, 0.8 and 1.0 (all correct choices). This data will not necessarily approximate the normal distribution, so we will adopt nonparametric testing approaches (while also reporting parametric t-tests to enable exploratory comparison with Tsay's and Mehr et al.'s original analyses). After being presented with all tasks, participants then provide demographic information including gender, age, and musical experience.

2.4. Statistical analysis

2.4.1. H1 (prediction of interaction effects between the domain and the variance)

We will use a rank-based procedure factorial design which is designed for the general nonparametric testing of treatment effects (Brunner et al., 2018; Friedrich et al., 2017; Noguchi et al., 2012). The null hypothesis is that the interaction effect of the two factors (i.e. the domain and variance) is zero. The ANOVA-type statistic will be used as a test statistic and we rely on the R-package nparLD for its calculation for repeated measurements (Noguchi et al., 2012). Regarding the use of nparLD, it is known that the ANOVA-type statistic does not lead to asymptotically correct statistical decisions (Friedrich et al., 2017). However, we consider it is still useful for the following two reasons.

Firstly, Friedrich et al. (2017) proposed to use a wild bootstrap method to improve the asymptotic correctness of the ANOVA-type statistic but they also mentioned that both the classical way of calculation by nparLD and their wild bootstrap method brought similar conclusions even though the latter method is more accurate. Furthermore, Umlauft et al. (2019) remarked from their simulations that the classical ANOVA-type statistic can still be relied on for global testing (i.e. testing the existence of interaction effects rather than post-hoc analysis) and our test is 2 × 2 factorial design, so the theoretical issue of the ANOVA-type statistic is not practically relevant in this study.

2.4.2. H2-H3 (prediction of the dominant domain for each variance condition)

We will use a studentized permutation test for the nonparametric paired data (Konietschke & Pauly, 2012) which is designed for the nonparametric Behrens-Fisher problem and is not requiring symmetry in the distribution as like the Wilcoxon signed-rank test. Formally, this method tests the relative effect q = 0.5 as a null hypothesis which means there is no difference between the paired data. In this study, we predict q > 0.5 as a one-tailed alternative hypothesis (i.e. a particular domain condition yields a higher percent correct). In H1, the two samples to be compared are the low-variance × visual-only condition and the low-variance × audio-only condition paired by participants. Similarly, the high-variance × visual-only condition and the high-variance × audio-only condition paired by participants are the target two samples of H2. The R-package nparcomp (Konietschke et al., 2015) will be used for the implementation.

2.4.3. Significance level of Type-1 error

Because we are testing six predictions (3 each for piano and shamisen), we will use a Bonferroni correction to maintain an overall Type-1 Error alpha level of .05 (i.e., the critical significance *p*-value for each test will be set to .0083).

2.4.4. Evaluation of the support for the null hypothesis

If we fail to reject the null hypothesis for H2 or H3, we will conduct tests analogous to equivalence testing (Lakens, 2017; Schuirmann, 1987) based on the above nonparametric test methods. The original idea of the equivalence testing was developed for the t-test, and the test is performed by constructing the confidence interval around the test statistic (i.e. t-statistic) and then checking whether the pre-specified equivalence interval falls within the confidence interval. If yes, then the difference between the two groups is considered not exceeding the minimal meaningful difference expressed by the equivalence interval, and the two groups are deemed statistically equivalent.

Since the above nonparametric test methods involve the calculation of rank statistics which can provide an estimate of the relative effect, we will report the relative effect with its 90% confidence intervals as the effect size of each experiment, and we will assess the support for the null hypothesis by checking whether the confidence interval overlaps with the equivalence interval we consider meaningful. The reason for using 90% is to create a confidence interval same as the two one-sided tests procedure used in the equivalence testing (Lakens, 2017; Schuirmann, 1987). Specifically, we set the relative effect's equivalence interval of [0.39, 0.61] as the smallest effect size, corresponding to Cohen's d of ±0.4 (Ruscio, 2008), which is often considered a reasonable estimate of a "Smallest Effect Size Of Interest" (SESOI) for purposes of power analysis (Brysbaert, 2019; see additional justification of effect size in the "Power analysis" section below).

Regarding H1, we will create a confidence interval for the equivalence testing in a similar way to the methods proposed for fixed-effects ANOVA (Campbell & Lakens, 2021; Smithson, 2001). To be more precise, we will conduct the test according to the following steps if we fail to reject the null hypothesis for H1. Firstly, we calculate a finite denominator degrees of freedom of the ANOVA-type statistic (Brunner et al., 1997) which is set as infinity at the calculation of p-value (i.e. $F\left(df_1, \infty\right)$). Secondly, the non-centrality parameter of the underlying F-distribution is obtained and the 5% quantile value of F statistics is derived from the non-central F-distribution. Thirdly, the partial eta squared corresponding to the derived F statistics is calculated using the equation (4) of Smithson (2001) with the adjustment of positive bias (Mordkoff, 2019). We confirmed the use of Smithson (2001)'s equation can reproduce the 90% CI [0.31, 0.82] of partial eta squared presented in Lakens (2013)'s exemplary analysis of repeated measures ANOVA. Finally, by constructing a confidence interval of 5-100% of partial eta squared, we judge the non-inferiority of effect by whether a pre-specified threshold does not exist in this interval as similar to Campbell & Lakens (2021). We will use 0.01 for the threshold which is a borderline of the small effect of eta squared (Kirk, 1996). We acknowledge that eta squared and this 0.01 is basically used for between-subjects design so it is not compatible with our experimental design. Conceptually, it is recommended to set a meaningful "no effect" borderline from an ecological reason such as based on just noticeable differences (Lakens et al., 2018). Though there is no data we can rely on to set the threshold for the sight-vs-sound effect under within-subjects paradigms, we hope our study can be a basis for more precise analysis of performance judgment undertaken in future research.

2.5. Power analysis

A priori power analysis requires estimating the effect size before collecting data, which is notoriously difficult (Brysbaert, 2019). In this paper, we rely in part on previously published data from several hundred participants from Tsay's (2013) original study and Mehr et al.'s (2018) direct and conceptual replications. Because replications tend to more accurately estimate effect sizes than first publications due to publication bias (Open Science Collaboration, 2015), we focus on Mehr et al.'s data over Tsay's. We will set acceptable false negative parameters based on commonly used power guidelines of 80% and a family-wise *alpha* level of 0.05 (i.e., .0083 for each of 6 hypothesis test; see above for rationale).

As described in section 1.1, re-analysis of Mehr et al.'s data using the parametric t-tests originally used by Tsay and by Mehr et al. suggests a range of effect sizes ranging from a minimum of Cohen's d = 0.42 (for Study 2) to 0.57 (for Study 1 directly replicating Tsay) to 1.2 (for Study 3). When these data are reanalyzed using the non-parametric methods planned for our confirmatory analysis, these correspond to relative effect sizes ranging from 0.62 (Study 2) to 0.64 (Study 1) to 0.80 (Study 3). Since all data in our within-subjects experiment are collected from the same participants, our necessary sample size will be determined only by the smallest effect size of interest. Given that the smallest effect size found previously (Cohen's d = 0.42) is slightly larger than the value of 0.4 often cited as an approximation of the "smallest effect size of interest" (SESOI; Lakens, 2017), we will use the more conservative SESOI of d = 0.4, corresponding to a minimum relative effect of 0.61, giving a required sample size of n=155 participants. Note that this estimate is based on a between-subjects design, so because within-subjects designs are considered to potentially have higher power than between-subjects designs (Lakens, 2013) this is likely a conservative overestimate of the true sample needed to achieve power of 80%.

Regarding the interaction effect, we obtained a partial eta squared of 0.20 from the ANOVA-type statistics. By using this value as an input of G*Power (Faul et al., 2009), the required sample size was estimated as 53 participants in total. This estimation was based on the fixed-effects ANOVA setting as in the above presumptions. Since this estimate gives a substantially lower minimum sample size than described above, we will again use the more conservative estimate of n=155 participants described above.

2.6. Participants

Participants will be native Japanese speakers 18 and older who have no hearing or visual disabilities and who have read and consented to the online experiment. They will be recruited from Keio University and the surrounding communities through a combination of social media, printed flyers, and word-of-mouth advertisements. Participants will be reimbursed Keio University's standard rate (currently ¥1,050, approximately US$10). We ask them to respond to basic demographic items (e.g., Age, Gender, Native Tongue, general musical instrument experience, experience listening/performing Tsugaru shamisen, piano, or other music; and free response regarding factors they felt were relevant to evaluating piano and shamisen performances) after the experiment, and the online experiment will take approximately 45 minutes for completion.

2.7. Video editing method

All piano videos were taken directly from the supplementary materials published by Mehr et al. (2018). To edit the new Tsugaru shamisen videos, GC (1st author) used a video editing software called DaVinci Resolve. The Tsugaru shamisen tournament video included the tournament, category name, performer name, etc., so we masked these details. We also magnified the video to allow better viewing of the performers' movements, and adjusted the focus of footage such that performers would be in the center of the screen. Moreover, because sound volume between Tsugaru shamisen competition videos and Piano competition videos in our experiment was quite different, GC used a sound editing software called ffmpeg and matched max-volume to about -10dB. We also corrected for extraneous noises to maintain appropriate sound quality. Experimental stimuli excerpts and full original videos can be viewed at https://osf.io/p9fvs.

2.8. Pilot data

Pilot experiment data (*n* = 9 participants) were collected. Figure 2 shows pilot data for the percentage selected as the actual winner in each confirmatory condition (Audio-only and Visual-only). Most importantly, our results suggest that in most cases participants are able to correctly identify the actual winners at levels substantially greater than the 50% chance level using either audio-only or video-only stimuli (with the possible exception of low-variance shamisen condition). Even given this small amount of data, it suggests that the previous piano results by Tsay (2013) and Mehr et al. (2018) may be replicable with our new within-subjects design and unified criteria of 5 pairs per condition. Data of Tsugaru shamisen also suggest a possibly similar tendency to the piano data, though the effect appears weaker. Though we need to take into account the small amount of sample, these pilot data suggest that our experimental paradigm should be able to collect meaningful data to allow us to evaluate whether our hypotheses are supported.

2.9. Exploratory analyses

Currently, three exploratory analyses are planned. Firstly, we will also perform comparative analysis with the Audio-Visual condition data. Secondly, regarding the piano, we will also collect data using the stimuli of Mehr et al. (2018)'s Experiment 2, so we will check whether the same sight-over-sound effect is replaced using stimuli different from the ones used in the confirmatory analysis and in Tsay's (2013) original analysis. Lastly, we will explore whether there may be differences in the sight-vs-sound effects for each of the 25 individual competitions (20 confirmatory + 5 exploratory).

3. Changes to Stage 1 Registered Report Protocol (Introduction and Methods Sections 1-2)

With two minor exceptions, we have left the Introductory and Methods (Sections 1-2) unchanged from the version granted In Principle Acceptance (accessible at https://osf.io/ry2b6), to avoid any appearance of adjusting hypotheses or analyses after results. This means we continue to use future tense in Sections 1-2 although in the following sections we use the past tense since the experiments are now complete. We also chose not to update the

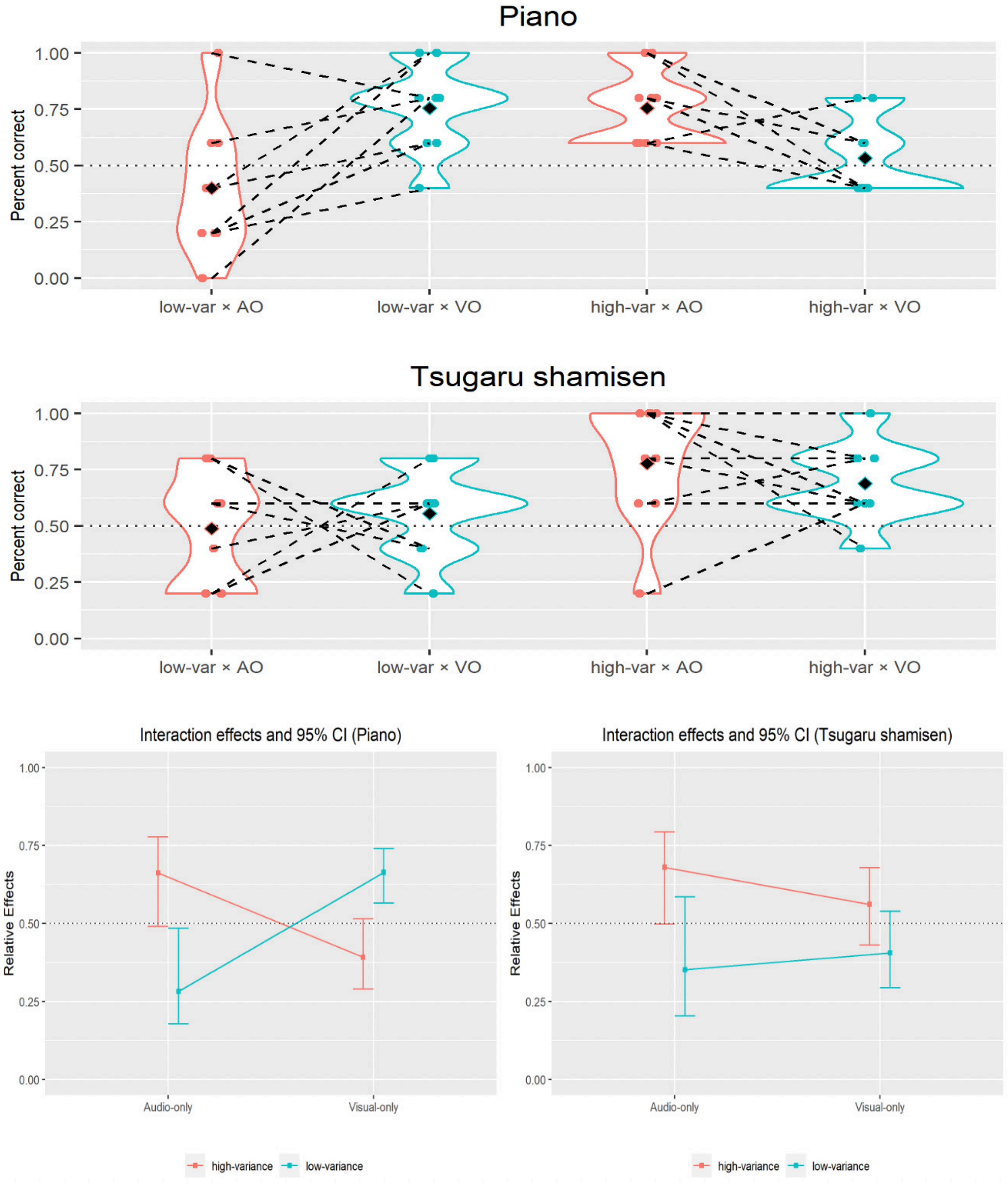

Figure 2. The top figure is the violin plots of the pilot data (n = 9). Black diamonds indicate mean values.

Dashed lines indicate paired data from the same participant. The bottom two figures show the relative effects of piano (left) and shamisen (right), and the bars are 95% confidence intervals based on the ANOVA-type statistics. Dashed lines (q = 0.5) indicate there is no effect. When the equivalent test is performed, confidence intervals will be calculated differently which is based on a studentized permutation test.

violin plot smoothing in Figs. 1 and 3 to the more intuitive visualization used in the following Figs. 4, S1 and S4, even though this smoothing makes the visualization clearer and has no effect on any statistical hypothesis testing. We also added new discussion of a relevant study by Wilbiks & Yi (2022) published after our Stage 1 protocol received In Principle Acceptance to the Discussion (Section 5), although normally such discussion would be appropriate in the Introduction.

One exception is that we have corrected “visuals” to read “sound” for H3 in section 1 (i.e., H3 now reads “*...**sound** will dominate the judgment of piano performance between upper and lower ranks (1st place vs. low-placing)...*” instead of “*...**visuals** will dominate the judgment of piano performance between upper and lower ranks (1st place vs. low-placing)...*”. Note that the original abstract and Table 1 correctly stated our intended prediction that sound would dominate in this high-variance condition (abstract: “*3) sound is predicted to trump visuals when variation in quality is high (1st vs. low-placing)*”; Table 1: “*H3: Sound dominates the judgment of piano performance between upper and lower ranks (1st place vs. low-placing), due to the high variance in trials.*”).

The other exception is that we corrected a typo in numbering (the “Exploratory sample” section that is now la-

Table 2. Registered Report design planner

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given outcome	Actual outcome
Does the dominance of the domain (audio or visual) depend on the variance in the performance qualities in performance?	H1: There is an interaction effect between the modality factor (audio-only vs. video-only) and the quality variance factor (low vs. high variance) such that sight vs. sound effects depend on the performance quality gap of competitors.	*n* = 155 (the rationale is given in 2.4)	Nonparametric repeated measurements using rank-based procedures and the ANOVA-type statistic (α = .0083).	There is/ is not an interaction between the domain and the variation in performance quality.	The hypothesis was supported in both the piano and Tsugaru-shamisen cases.
Which type of information, if any, has greater impact on the evaluation of piano performance in music?	H2: Visuals dominate the judgment of performance between upper ranks (1st vs. 2nd place), due to the low variance in trials.		A studentized permutation test for the nonparametric paired data of rate selecting actual winner in audio-only vs. video-only conditions (*α* = .0083). Equivalence testing if non-significant (α = .0083, 0.39 > *relative effect* ≤ 0.61)	Visuals or sound does/ does not dominate when judging between upper and lower ranks.	The hypothesis was only supported in the piano case. Regarding the Tsugaru-shamisen case, the equivalence test confirmed that there is no meaningful effect by the modality (i.e. audio or visual) for the accuracy of judging the actual competition winners.
(Same as above)	H3: Sound dominates the judgment of piano performance between upper and lower ranks (1st place vs. low-placing), due to the high variance in trials.		(Same as above)	(Same as above)	The hypothesis was only supported in the Tsugaru-shamisen case. Regarding the piano case, the equivalence test confirmed that there is no meaningful effect by the modality (i.e. audio or visual) for the accuracy of judging the actual competition winners.

(H1-H3 are each tested twice: once replicating previous stimuli from piano competitions and once using novel stimuli from Tsugaru shamisen competitions)

beled 2.1.2 was erroneously labeled 2.1.1). Note that we left the following statement in Section 2, although we ultimately decided to pursue other exploratory analyses instead: "*Incidentally, studying the interaction effects brought by musical instrument/genre (Western classical piano vs. Japanese folk Tsugaru shamisen) is not within the scope of our hypotheses so this is not counted as a factor, but we will add this into our factorial design model in the exploratory analysis.*"

We have updated the abstract and title, and added an extra column to Table 2 to reflect our actual findings after collecting full Stage 2 data (original Stage 1 title: "*Sight vs. sound in the judgment of music performance: Cross-cultural evidence from classical piano and Tsugaru shamisen competitions [Stage 1 Registered Report]*").

We excluded data from the participants who met one of the following conditions:

a. Reported a native language that was not Japanese (despite being a native Japanese speaker listed as a requirement in the experiment recruitment announcement; n = 3 participants excluded)
b. Completed the online experiment multiple times (only the first complete experiment was used; n = 4 participants excluded), or
c. Completed the experiments faster than was physically possible if they were performed correctly (i.e., timestamp between completing Part 1 and Part 2 was less than the 6 minutes that would be required to fully listen/watch all excerpts, enter responses, and

fill out the questionnaire during Part 2; n = 6 participants excluded).

d. Missing data (because either participants did not pick winners or there was some problem with the online form; n = 1 participants excluded).

While our Stage 1 Registered Report protocol explicitly listed Japanese native language as an inclusion criteria, we only realized issues b-d after beginning data collection and so these were not explicitly listed as exclusion criteria in our Stage 1 protocol. Thus, for the sake of transparency, we report results without excluding the 10 participants from exclusion criteria b & c, although the results are the same as the confirmatory analysis (cf. S1.5; null hypotheses of H1, H2, H4, and H6 are rejected, and the equivalence testing for H3 and H5 rejects corresponding null hypotheses; note that the 1 participant with missing data [d] was excluded for both confirmatory and exploratory analyses because including them would have required additional decisions about how to impute the missing values).

Finally, we corrected a few typos and formatting errors, and added a clarification that H1-3 refer to our three hypotheses for the piano condition, while H4-6 refer to the same three hypotheses for the Tsugaru shamisen condition.

All statistical hypothesis testing used the same methods and code specified in our Stage 1 protocol that received In Principle Acceptance after peer review.

4. Results

4.1. Confirmatory analysis

After collecting the pre-determined full sample (n = 155 participants), our pre-specified analyses (Figs. 3-4; Table 3) confirmed our predicted interaction effects between audio/visual modality and variance in performer quality in both piano and Tsugaru-shamisen (H1 and H4; adjusted $\eta^2_{partial}$ = 0.038 and .012; p = 4.9x10^{-7} and .0039, respectively). We also confirmed our predictions that that visuals dominate in the evaluation of low variance in performer quality (1st vs. 2nd place) for the piano condition (H2; relative effect = .69; equivalent Cohen's d = 0.71; p < 1.0x10^{-15}) and that audio dominates in the evaluation of high variance in performance quality (1st vs. low rank) for the Tsugaru-shamisen condition (H6; relative effect = .40; equivalent Cohen's d = 0.39; p < 1.0x10^{-15}).

On the other hand, our analyses did not confirm our predictions for the high-variance condition with piano (H3; relative effect = .48; equivalent Cohen's d = 0.085; p = .22) or the the low-variance condition with Tsugaru shamisen (H5; relative effect = .51; equivalent Cohen's d = 0.021; p = .41). In fact, equivalence tests for these two conditions found that the confidence intervals of relative effects were entirely covered by the specified region of equivalence (i.e. relative effect size of 0.39-0.61). Therefore we concluded the difference in the prediction accuracy of competition winners between audio vs. visual modalities does not meaningfully differ for the high-variance piano or low-variance Tsugaru shamisen conditions (H3 and H5, respectively). 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.

4.2. Exploratory analysis

With two exceptions, we do not perform any statistical tests in this exploratory analysis section but report our observations from descriptive statistics and visualization of data. The only exceptions are for our t-tests designed to replicate the statistical methodology of Tsay (2013), Mehr et al (2018; Table S2, as previously described in Section 2.3 of the Stage 1 protocol), and Wilbiks & Yi (2022; Table S4; added during Stage 2). Although the parametric assumptions of the t-test are not met, these tests still give qualitatively identical results to our chosen non-parametric methods (statistically significant differences supporting predictions H2 and H6 but not H3 and H4, although obtained p-values and effect sizes vary slightly). Incidentally, please note that, due to a technical glitch, the participants who completed Part 2 (the experiments of the visual-audio condition and exploratory stimuli; n=160 participants) are not exactly matched with Part 1 consisting of only the confirmatory experiments (n=155). Some participants' answers were only recorded for Part 2 and they are also included in our exploratory results.

We first explored the percent accuracy under the Audio-Visual condition (red violin plots in Fig. 3). Given that differences between these results and the audio and visual were relatively minor, not statistically evaluated, and that the order of presentation may have affected them (audio-visual conditions were presented after participants had already evaluated the audio-only and visual-only versions), we refrain from extensive speculation about them. However, it is worth noting that the average percent accuracy in the audio-visual condition turns out to fall in-between the audio-only and visual-only conditions in the low-variance stimuli with piano, replicating the qualitative pattern shown by Tsay (2013: Fig. 3) and Mehr et al Study 1 (2018; cf. Fig. 1 in the current paper) for the same stimuli with visual-only highest accuracy, followed by audio-visual, followed by audio-only.

Next, we re-ran the analysis of the piano condition with the full set of piano stimuli used in Mehr et al. (2018)'s Experiment 2 (see 2.1.1 Exploratory samples). It generated the same result as in the confirmatory analysis (supporting predictions H1 and H2 but not H3; adjusted $\eta^2_{partial}$ = 0.070, relative effect = 0.77, and relative effect = 0.48; p = 1.6x10^{-11}, < 1.0x10^{-15}, and 2.18x10^{-1}, respectively; Fig. S2; Table S1).

To see whether sight-vs-sound effects may be driven by the choice of specific excerpts, we checked whether there are noticeable differences among the chosen excerpts (Fig. 4). We averaged percent accuracy by each trial, domain and variance conditions. This summary provides a view of data similar to the confirmatory analysis. For example,

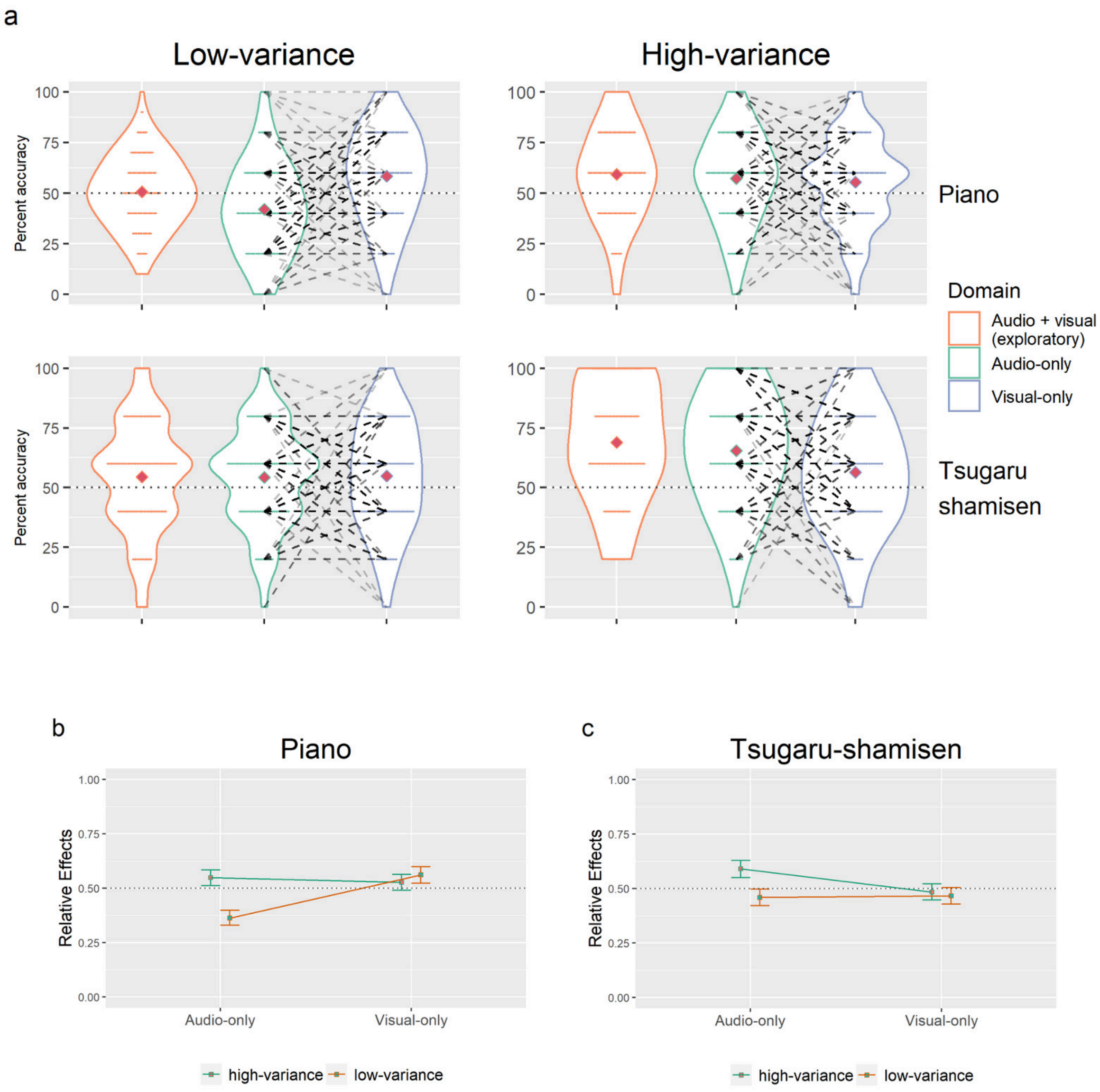

Figure 3. The top figure (a) shows violin plots of the full data (n = 155 participants for the audio-only and visual-only data) for the dependent variable of % correctly choosing the 1st-placed performer in a two-choice forced choice task. Red diamonds indicate mean values. Dashed lines indicate paired data from the same participant. The bottom two figures show the interaction effect of relative effects of piano (b) and Tsugaru-shamisen (c), and the bars are 95% confidence intervals based on the ANOVA-type statistics.

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. Dashed lines (q = 0.5) indicate there is no effect. The percent accuracy of the Audio-Visual condition (red violin plot; n=160 participants) is for the exploratory analysis and thus supplementary information.
NB: We have changed the smoothing interval for the violin plots from that used in the Stage 1 protocol (Fig. 2) for more intuitive visualization (this doesn't affect our statistical analyses).

we also observed the pattern that visual dominates in the low-variance condition under the piano stimuli and audio dominates in the high-variance condition under the Tsugaru-shamisen stimuli. Although we can see the visual-only condition tends to result in higher percent accuracy for the piano stimuli and vice versa for the Tsugaru-shamisen stimuli, further investigation is needed to test if this result is by chance or the realization of the difference between the piano and Tsugaru-shamisen.

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

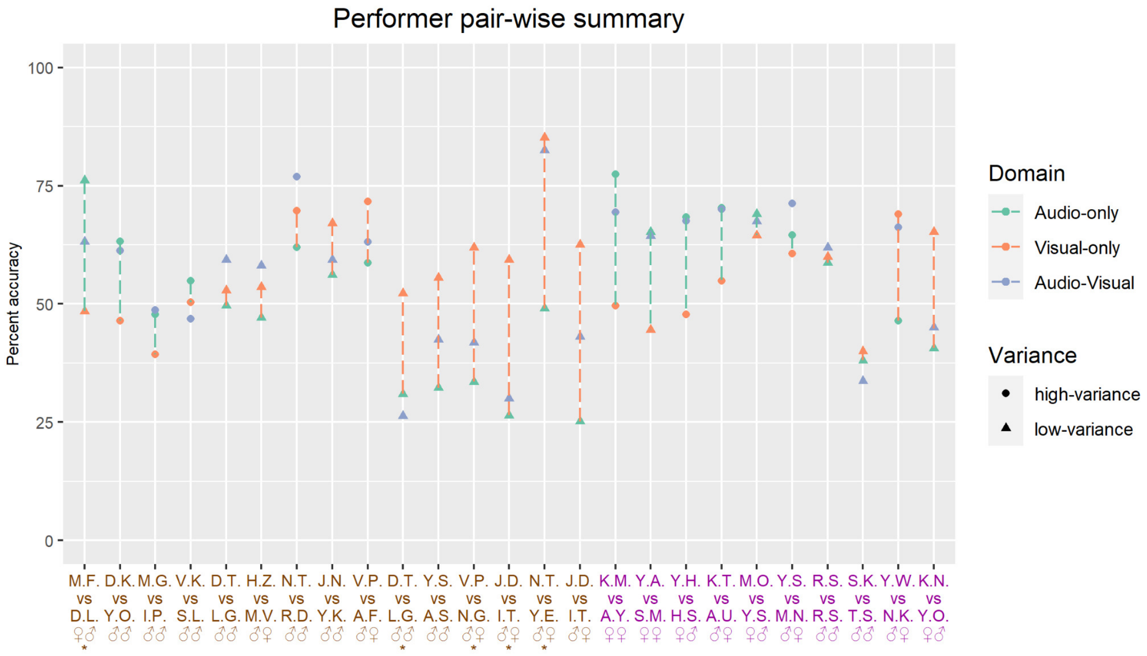

Figure 4. The average percent accuracy of each pair of clips.

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.

Table 3. Summary of the hypotheses tests and obtained effect sizes.

#	Test statistic	Obtained statistic	p-value (α=0.05/6)	90% CI for equivalence testing (rejection region 0.39-0.61)	Effect size translation	Obtained effect size
H1	ANOVA-type statistic	25.31	***$4.9x10^{-7}$**	-	Adjusted $\eta^2_{partial}$	0.038
H2	Relative effect	0.69	***< $1.0x10^{-15}$**	-	Cohen's D	0.71
H3	Relative effect	0.48	.22	***0.43-0.53**	Cohen's D	-0.085
H4	ANOVA-type statistic	8.33	***$3.9x10^{-3}$**	-	Adjusted $\eta^2_{partial}$	0.012
H5	Relative effect	0.51	.41	***0.45-0.56**	Cohen's D	0.021
H6	Relative effect	0.40	***< $1.0x10^{-15}$**	-	Cohen's D	-0.37

*Statistically significant (family-wise α < .05 after correcting for multiple comparisons). All values are given to two significant figures for consistency except for ANOVA-type statistics (approximation of F statistics). ANOVA-type statistics are given to two decimal places to retain the necessary precision for p-values.

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).

We additionally explored whether there is a noticeable difference in the sight-vs-sound effect among competitions (Fig. S3) and whether participants believed audio or visual information should be more important (S1.6). Regarding the former, the percent accuracy was summarized as the same in the case of the performer pair-wise summary, but the average was taken competition-wise. No clear trends emerged in sight-vs-sound effects among competitions, which suggests competition-level effects were unlikely to have driven any of the results of our analyses. The latter revealed that 91% of answers showed that audio is more important than visual though our confirmatory analysis found visual information is more reliable for predicting competition winners in the low-variance condition with piano, which is similar to the result of 83% found previously for novice US listeners of classical music competitions (Tsay, 2013: Fig. 1).

Finally, spurred by another claimed “failed replication” of Tsay’s results published after our Stage 1 protocol was published (Wilbiks & Yi, 2022), we checked whether the percent accuracy at each condition significantly differs from the chance level (50%; S1.7). In summary, except for the piano condition with the low-variance audio-visual stimuli, we found all conditions show a tendency of deviation from the chance level when using the same methodology as previous studies (i.e., $p < 0.05$ using one-sample t-tests; cf. S1.7). As already shown in Figure 3, only the case of the piano condition with the low-variance audio-only stimuli resulted in disagreement with the expert’s judgment (i.e. mean percent accuracy lower than 50%). However, we emphasize that this is an exploratory analysis not pre-registered in our Stage 1 Registered Report protocol that does not meet the same exacting standards of evidence of our main confirmatory analyses.

Our experiments presented pairs of performances that are different from the original experiment design by Tsay (2013) or Tsay (2014) using triads, so it is difficult to directly compare our results with theirs. But they also found the percent accuracy under the audio-visual condition was at a chance level (Experiment 3), which coincides with our result. Study 2 and Study 3 by Mehr et al. (2018) presented stimuli in pairs like ours, but they reported that the visual-only condition produced non-significant results. Recently, Wilbiks & Yi (2022) also reported the replication study of Tsay’s (2014) study. Although they only tested under the visual-only condition with triad stimuli presentation, what they observed was the same as the case of Mehr et al. (2018); percent accuracy was at chance level. All cited experiments differ in design, sample size and maybe participants demographics, and these mixed results suggest the assessment of whether participants can guess the true competition winners better than chance requires careful experiment design.

5. Discussion

Our study replicates and extends previous studies using a cross-cultural paradigm to confirm our prediction that sight-vs-sound effects in judgment of musical performance depend on the relative quality of the performers. Specifically, for both Western classical piano and Japanese Tsugaru shamisen competitions, the closer two performers are in quality, the more participants’ evaluation of their performance is affected by visual information. This supports Tsay’s (2013) claim that visual information can affect judgments of musical performance, but also supports Mehr et al.'s (2018) claim that the strength of such effects depends on the relative quality of the performers.

However, our predictions for the precise form such effects would take in different contexts were only partially accurate. Specifically, we confirmed the predicted sight-over-sound effect for low variance (1st- vs. 2nd-placed) piano performers and the predicted sound-over-sight effect for high variance (1st- vs. low-ranked) Tsugaru shamisen performers, but did not find the corresponding predicted sight-over-sound effect for low-variance Tsugaru shamisen or sound-over-sight effect for high-variance piano. This suggests that, while the general phenomenon of cross-modal interactions is cross-culturally general, the specific ways in which they manifest vary depending on the cultural and performance context. For this reason, we use the more general term “sight vs. sound effects” rather than “sight-over-sound effect” in our title.

Our results contradict a recent paper entitled “Musical novices are unable to judge musical quality from brief video clips: A failed replication of Tsay (2014)” (Wilbiks & Yi, 2022). Although Wilbiks & Yi submitted their paper on September 28, 2022, more than 9 months after our published Stage 1 protocol received In Principle Acceptance at *Peer Community In Registered Reports* (Yamada, 2021), they did not cite or take into account our analysis of the problems with relying on one-sample t-tests to evaluate the replicability of Tsay’s claimed sight-over-sound effect, and so their analysis suffers from similar limitations as Mehr et al.'s (2018) already described in the introduction above. However, even judging Wilbiks & Yi’s results by their own chosen criteria, our results contradict theirs on multiple counts. First, unlike Wilbiks & Yi, we found that musical novice (Japanese) participants *were* consistently able to judge musical quality from brief 6s video clips at levels above chance, whether this was using Tsay’s (2013) original piano stimuli, Mehr et al.‘s (2018) alternative piano stimuli, or both the high-variance and low-variance versions of our new Tsugaru shamisen stimuli. We also found the same above-chance level results for the audio clips, with the exception of the audio-only low-variance piano condition, and even this exception was consistent with Tsay’s (2013) result that participants could not judge this specific condition above chance level. We also note that Wilbiks & Yi simply failed to reject the null hypothesis of no deviation from chance, rather than using an analysis such as equivalence testing that could have provided stronger evidence *for* the null hypothesis (Lakens, 2017). Finally, we note that, for reasons not made clear in the paper, Wilbiks & Yi chose to focus their replication study on a follow-up study by Tsay (2014) using slightly different stimuli and design rather than on the Tsay (2013) stimuli used here and by Mehr et al. (2018). Thus, while we cannot directly compare our results with their specific participant and stimulus sample, our general conclusions here using 6s stimuli contradict the broad claim from Wilbiks & Yi's (2022) abstract that "6s is not a sufficient amount of time for novices to judge the relative quality of musical performance, regardless of the modality in which they were presented". At the same time, our results do show that participants' judgments are only slightly above chance levels, which does support a weaker form of Wilbiks & Yi's claim, i.e., novices are not particularly good at judging musical quality from short 6s excerpts. However, a more comprehensive analysis should compare against appropriate controls (e.g., can expert judges achieve much higher accuracy when watching full performances with both audio and video?).

Curiously, while our sample of Japanese participants replicated the key sight-over-sound result reported by both Tsay (2013) and Mehr et al. (2018) for U.S. participants in the high-variance piano condition, and we replicated Mehr et al.'s key prediction of interaction effects between modality and variance in performer quality and even replicated their predicted sound-over-sight effect in Tsugaru shamisen performances, we failed to replicate Mehr et al.'s reported sound-over-sight effect for piano performances despite using identical experimental stimuli. While the slight differences in experimental design (e.g., our within-subjects paradigm vs. their between-subjects) might conceivably have caused this, one possible speculation is that the different cultural backgrounds of participants may have played a stronger role. However, we are not completely confident about this interpretation either because Western classical music has spread widely enough that our novice Japanese participants may well have had similar levels of exposure to it as Mehr et al.'s novice US participants (indeed, several of the competition winners were Japanese).

Importantly, the differences in performer quality were intentionally higher for the Tsugaru shamisen high-variance condition than for the piano high-variance condition, because we could not be confident from pilot experiments that choosing higher-ranked comparisons (e.g., 1st vs. 8th instead of 1st vs. 21~50th) would enable us to get results more accurate than chance for either audio-only or visual-only conditions. This selection was possible because Tsugaru-shamisen competitions disclosed the ranks of most competitors. In contrast, three out of the five lower-placed performers appearing in Mehr et al.'s high-variance piano condition can be actually regarded as top-level performers (i.e., 3rd place or semifinalists), so it is plausible that choosing the actual winners in this condition (e.g., 1st vs. 3rd place) was too challenging for our novice Japanese listeners (though this does not explain why novice US listeners were able to perform so well). Although this would not explain why we could not replicate Mehr et al.'s (2018) high-variance piano result, this difference in stimulus selection may have caused why a sound-over-sight effect was detected in the high-variance condition of Tsugaru-shamisen data.

Although we could not find a sight-over-sound effect in the low-variance condition of Tsugaru-shamisen data, if we subjectively compare the movie clips between piano and Tsugaru-shamisen, it is clear that Tsugaru-shamisen players tend not to dress up in fancy formal outfits like most of the classical piano performers but instead wear more casual attire (e.g. denim jeans). Furthermore, their camera angle was fixed in contrast with dynamic panning and zooming in/out employed in the piano movie clips. This difference may be attributable to the difference in the traditions of the musical performance of Western classical piano music and Tsugaru-shamisen music, with the former often considered a prestigious, elite art form and the latter considered a more down-to-earth folk style where dramatic visual effects may come off as overly pretentious. We also observed that the body movements of Tsugaru-shamisen players appear more moderate than those of the pianists, which could also influence the impression of their performance quality. Considering these cultural differences in the visual expression of the performance, we consider that the failure of our sight-over-sound prediction for the high-variance Tsugaru shamisen condition may be because the participants could not find salient visual cues from the movie clips of Tsugaru-shamisen that they could rely on to accurately judge their talents.

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: Tolsá-Caballero & Tsay, 2021; Tsay, 2013). 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.

Based on this new hypothesis, a research question further arising is what has made the difference in increasing the emphasis on visual saliency in performance among musical traditions. Regarding the Tsugaru-shamisen, it is acknowledged that the history began with one blind man (Nitabo, 仁太坊) and the early days of Tsugaru-shamisen were developed by blind men (Chiba & Savage, in press; Daijo, 1995). One of the most influential figures in the history of Tsugaru-shamisen, Chikuzan Takahashi (高橋竹山), was praised for making Tsugaru-shamisen an art (Matsuki, 2011), and he also lost his vision in the childhood. Therefore, one speculation is that prestige bias (Mesoudi, 2011) has resulted in sight-vs-sound effects evolving in opposite directions for classical piano and Tsugaru shamisen performing traditions, with an emphasis on visuals driven by factors such as associations with prestigious outfits in classical piano competitions, but an emphasis *away* from visual effects in Tsugaru-shamisen performance propagating from those pioneering blind performers.

Further experiments and research are needed to test such mechanisms, resolve lingering discrepancies between our current findings and previous findings, and clarify even more controversial debates, such as regarding the potential role of blind auditions in reducing or magnifying racial or gender biases (Sommers, 2019). However, by recruiting Japanese participants and adding stimuli from the unique genre of Tsugaru shamisen, we showed the generalizability of sight vs. sound effects beyond the framework of a specific Western classical tradition and Western participants. Our study adds an important new cross-modal (audio/visual) dimension to an emerging body of cross-cultural music cognition, providing further evidence for the complex interplay between cross-culturally universal and culturally-dependent aspects of music cognition in the important applied domain of evaluating musical performance. We hope our Registered Report approach has contributed robust findings regarding the replicability and cross-cultural generality of sight vs. sound judgments, which directly impact the livelihoods of musicians around the world.

Note

This is a full Stage 2 Registered Report that was peer-reviewed and recommended by *Peer Community In Registered Reports* and accepted for publication in the journal *Collabra: Psychology.* For full open peer reviews, author replies, and editorial recommendation, see https://doi.org/10.24072/pci.rr.100351.

Ethics

We have approval of the Keio University Shonan Fujisawa Campus Institutional Review Board to PES (approval #298). All participants provided informed consent.

Data/Code Accessibility Statement

Audiovisual stimuli are available at https://osf.io/p9fvs/

Data and analysis code are available at https://github.com/comp-music-lab/sight-vs-sound.git

The full experiment can be accessed at https://gakuto101207.github.io/

Authors Contributions

Conceptualization: Gakuto Chiba, Patrick E. Savage, Shinya Fujii

Investigation: Gakuto Chiba [prepared experiments, recruited participants, collected data]

Analysis: Yuto Ozaki, Gakuto Chiba, Patrick E. Savage

Writing – Stage 1 draft: Patrick E. Savage, Gakuto Chiba, Yuto Ozaki

Writing – Stage 2 draft: Yuto Ozaki, Patrick E. Savage

Writing –reviewing/editing: All authors

Project administration/supervision/funding acquisition: Patrick E. Savage, Shinya Fujii

Competing Interests

We declare we have no competing interests.

Acknowledgments

We thank Chia-Jung Tsay, Jonathan Wilbiks, Kyoshiro Sasaki, David Hughes, and Yuki Yamada for extensive feedback on earlier versions of the manuscript, Tomohiro Samma for discussion of ideas for testing the generality of sight vs. sound effects, and students of the Keio University CompMusic and NeuroMusic labs for assistance in collecting pilot data. We thank all participants who took the full experiments.

Funding

Funding for this study is provided by Grants-In-Aid from the Japan Society for the Promotion of Science to PES and SF (#19KK0064 and #20H04092), and by grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund) to PES.

Submitted: February 23, 2023 PDT, Accepted: February 23, 2023 PDT

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201–272). Academic Press. https://doi.org/10.1016/s0065-2601(00)80006-4

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-second sale: Using thin slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, *16*(1), 4–13. https://doi.org/10.1207/s15327663jcp1601_2

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441. https://doi.org/10.1037/0022-3514.64.3.431

Ballew, C. C., II, & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Bergeron, V., & Lopes, D. M. (2009). Hearing and seeing musical expression. *Philosophy and Phenomenological Research*, *78*(1), 1–16. https://doi.org/10.1111/j.1933-1592.2008.00230.x

Blacking, J. (1973). *How musical is man?* University of Washington Press.

Brunner, E., Bathke, A. C., & Konietschke, F. (2018). *Rank and pseudo-rank procedures for independent observations in factorial designs: Using R and SAS*. Springer. https://doi.org/10.1007/978-3-030-02914-2

Brunner, E., Dette, H., & Munk, A. (1997). Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association*, *92*(440), 1494–1502. https://doi.org/10.1080/01621459.1997.10473671

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Campbell, H., & Lakens, D. (2021). Can we disregard the whole model? Omnibus non-inferiority testing for R^2 in multi-variable linear regression and ⟐^{2} in ANOVA. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 64–89. https://doi.org/10.1111/bmsp.12201

Chiba, G., & Savage, P. E. (in press). Traditional folk music in contemporary Japan: Case studies of standardization and diversification in Tsugaru shamisen and folk song. In H. Johnson (Ed.), *Handbook of Japanese music in the modern era*. Brill. https://doi.org/10.31234/osf.io/mwb2z

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135. https://doi.org/10.1016/j.brainres.2008.04.023

Daijo, K. (1995). *津軽三味線の誕生：民俗芸能の生成と隆盛*. 新曜社.

de Gelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters*, *260*(2), 133–136. https://doi.org/10.1016/s0304-3940(98)00963-x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Friedrich, S., Konietschke, F., & Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, *113*, 38–52. https://doi.org/10.1016/j.csda.2016.06.016

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception*, *26*(5), 427–438. https://doi.org/10.1525/mp.2009.26.5.427

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *American Economic Review*, *90*(4), 715–741. https://doi.org/10.1257/aer.90.4.715

Griffiths, N. K. (2008). The Effects of Concert Dress and Physical Appearance on Perceptions of Female Solo Performers. *Musicae Scientiae*, *12*(2), 273–290. https://doi.org/10.1177/102986490801200205

Haan, M. A., Dijkstra, S. G., & Dijkstra, P. T. (2005). Expert judgment versus public opinion: Evidence from the Eurovision song contest. *Journal of Cultural Economics*, *29*(1), 59–78. https://doi.org/10.1007/s10824-005-6830-0

Hughes, D. W. (2008). Folk music: from local to national to global. In A. M. Tokita & D. W. Hughes (Eds.), *The Ashgate Research Companion to Japanese Music* (pp. 281–302). Ashgate.

Jacoby, N., Margulis, E. H., Clayton, M., Hannon, E., Honing, H., Iversen, J., Klein, T. R., Mehr, S. A., Pearson, L., Peretz, I., Perlman, M., Polak, R., Ravignani, A., Savage, P. E., Steingo, G., Stevens, C. J., Trainor, L., Trehub, S., Veal, M., & Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: Methodologies, pitfalls, and practices. *Music Perception*, *37*(3), 185–195. https://doi.org/10.1525/mp.2020.37.3.185

Kirk, R. E. (1996). Practical Significance: A Concept Whose Time Has Come. *Educational and Psychological Measurement*, *56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Konietschke, F., & Pauly, M. (2012). A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. *Electronic Journal of Statistics*, *6*(none), 1358–1372. https://doi.org/10.1214/12-ejs714

Konietschke, F., Placzek, M., Schaarschmidt, F., & Hothorn, L. A. (2015). **nparcomp**: An*R*Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software*, *64*(9), 1–17. https://doi.org/10.18637/jss.v064.i09

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

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

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Leman, M. (2008). *Embodied music cognition and mediation technology*. MIT Press.

Matsuki, H. (2011). *津軽三味線 まんだら 津軽から世界へ - 奏者たちの苦闘とその歴史*. 邦楽ジャーナル.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mehr, S. A., Scannell, D. A., & Winner, E. (2018). Sight-over-sound judgments of music performances are replicable effects with limited interpretability. *PLOS ONE*, *13*(9), e0202075. https://doi.org/10.1371/journal.pone.0202075

Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226520452.001.0001

Mordkoff, J. T. (2019). A Simple Method for Removing Bias From a Popular Measure of Standardized Effect Size: Adjusted Partial Eta Squared. *Advances in Methods and Practices in Psychological Science*, *2*(3), 228–232. https://doi.org/10.1177/2515245919855053

Murnighan, J. K., & Conlon, D. E. (1991). The dynamics of intense work groups: A study of British string quartets. *Administrative Science Quarterly*, *36*(2), 165–186. https://doi.org/10.2307/2393352

Nettl, B. (2015). *The study of ethnomusicology: Thirty-three discussions* (3rd ed.). University of Illinois Press.

Noguchi, K., Abel, R. S., Marmolejo-Ramos, F., & Konietschke, F. (2020). Nonparametric multiple comparisons. *Behavior Research Methods*, *52*(2), 489–502. https://doi.org/10.3758/s13428-019-01247-9

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). **nparLD**: An*R*Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *Journal of Statistical Software*, *50*(12), 1–23. https://doi.org/10.18637/jss.v050.i12

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orquin, J. L., Perkovic, S., & Grunert, K. G. (2018). Visual Biases in Decision Making. *Applied Economic Perspectives and Policy*, *40*(4), 523–537. https://doi.org/10.1093/aepp/ppy020

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, *30*(1), 71–83. https://doi.org/10.1525/mp.2012.30.1.71

Platz, F., & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behavior. *Musicae Scientiae*, *17*(2), 167–197. https://doi.org/10.1177/1029864913486369

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*(2), 109–111. https://doi.org/10.1111/j.1467-9280.2008.02054.x

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989x.13.1.19

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Authors' response: Toward inclusive theories of the evolution of musicality. *Behavioral and Brain Sciences*, *44*(e121), 132–140. https://doi.org/10.1017/s0140525x21000042

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/bf01068419

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, *36*(6), 888–897. https://doi.org/10.1068/p5635

Sloboda, J. A., Lamont, A., & Greasley, A. E. (2008). *The Oxford Handbook of Music Psychology* (S. Hallam, I. Cross, & M. Thaut, Eds.). Oxford Univ Press.

Smithson, M. (2001). Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals. *Educational and Psychological Measurement*, *61*(4), 605–632. https://doi.org/10.1177/00131640121971392

Sommers, C. H. (2019). Blind spots in the 'blind audition' study. *Wall Street Journal*. https://www.wsj.com/articles/blind-spots-in-the-blind-audition-study-11571599303

Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science*, *18*(9), 756–757. https://doi.org/10.1111/j.1467-9280.2007.01973.x

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Tolsá-Caballero, N., & Tsay, C.-J. (2021). Blinded by our sight: Understanding the prominence of visual information in judgments of competence and performance. *Current Opinion in Psychology*, *43*, 219–225. https://doi.org/10.1016/j.copsyc.2021.07.003

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, *110*(36), 14580–14585. https://doi.org/10.1073/pnas.1221454110

Tsay, C.-J. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, *124*(1), 24–33. https://doi.org/10.1016/j.obhdp.2013.10.003

Tsay, C.-J. (2021). Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries*, *7*(3), 343–366. https://doi.org/10.5465/amd.2019.0234

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Umlauft, M., Placzek, M., Konietschke, F., & Pauly, M. (2019). Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *Journal of Multivariate Analysis*, *171*, 176–192. https://doi.org/10.1016/j.jmva.2018.12.005

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, *101*(1), 80–113. https://doi.org/10.1016/j.cognition.2005.09.003

Wapnick, J., Mazza, J. K., & Darrow, A.-A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education*, *46*(4), 510–521. https://doi.org/10.2307/3345347

Wilbiks, J. M. P., & Yi, S. M. (2022). Musical novices are unable to judge musical quality from brief video clips: A failed replication of Tsay (2014). *Vision*, *6*(4), 65. https://doi.org/10.3390/vision6040065

Yamada, Y. (2021). Understanding the role of visual and auditory information in evaluating musical performance [Recommendation of Chiba, Ozaki, et al. Stage 1 Registered Report, 2021, PsyArXiv]. *Peer Community In Registered Reports*, 1–4. https://doi.org/10.24072/pci.rr.100003

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/73641-sight-vs-sound-judgments-of-music-performance-depend-on-relative-performer-quality-cross-cultural-evidence-from-classical-piano-and-tsugaru-shamisen/attachment/153763.docx?auth_token=YEaiNtmC5aH4m4Aqexh3

Supplemental Material

Download: https://collabra.scholasticahq.com/article/73641-sight-vs-sound-judgments-of-music-performance-depend-on-relative-performer-quality-cross-cultural-evidence-from-classical-piano-and-tsugaru-shamisen/attachment/153764.pdf?auth_token=YEaiNtmC5aH4m4Aqexh3