Sight vs. sound in the judgment of music performance: Cross-cultural evidence from classical piano and Tsugaru shamisen competitions [Stage 1 Registered Report]

Gakuto Chiba&, Yuto Ozaki&, Shinya Fujii, Patrick E. Savage*
Keio University, Fujisawa, Japan

&Equal contribution
*Correspondence to: psavage@sfc.keio.ac.jp

Abstract

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 will test the following three predictions twice each (once for piano and once for shamisen): 1) an interaction is 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 are predicted to trump sound when variation in quality is low (1st vs. 2nd place); and 3) sound is predicted to trump visuals when variation in quality is high (1st vs. low-placings). Data from pilot experiments (*n* = 9 participants) suggest that participants are mostly able to correctly select the actual winning performers based on short excerpts at levels above chance. In Stage 2, we will collect a full sample of 155 participants in order to achieve 80% power

Formatted: Numbering: Continuous
Deleted: E
Deleted: in Japan
Deleted: live.shamisen@gmail.com;
Deleted: live.shamisen@gmail.com and
Deleted: E
Deleted: in Japan
Deleted: comparing
Deleted: ed
Deleted: .
Deleted: using the Tsugaru shamisen, a musical instrument that has a unique cultural setting and musical tradition in Japan
Deleted: which type of information might be most impactful in a unique cultural setting and musical tradition that has historically excluded the use of and dependence on visual information.
Deleted: We use 207 performances of "Tsugaru Jongara Bushi" from 109 categories of national competitions in performing on the Tsugaru *shamisen and the same piano performance data used in the previous studies*,
Deleted: a traditional Japanese musical instrument
Deleted: , to evaluate two hypotheses
Deleted: 1
Deleted: i.e.e.g., choosing between
Deleted: place
Deleted: ,
Deleted: 2
Deleted: e.g., choosing between
Deleted: /2nd
Deleted: place and those who did not place among the top finalist
Deleted: Importantly, these two hypotheses will be confirmed with both Tsugaru shamisen and piano to generalize the previous studies' findings cross-culturally. The above two hypotheses assume the interaction effect between modality (audio/visual) and the performance quality gap (high-variance/low-variance), so we will also formally test the interaction effect of those two factors.
Deleted: 11
Deleted: non-placing competition participants
Deleted: 60
Deleted: 97
Deleted: 95

to detect effects of at least Cohen's d[1] = 0.4. Our results will reveal the generalizability of sight vs. sound effects to non-Western participants and musical traditions, and may have practical applications to evaluation criteria for performers, judges, and organizers of competitions, concerts, and auditions.

Deleted: using our within-subjects design

Deleted: leverage the characteristics of a unique

Deleted: not previously empirically tested with such paradigms

Data and materials availability: Pilot data and videos are available at https://osf.io/p9fvs/. Analysis code is available at https://github.com/comp-music-lab/sight-vs-sound.git. The full experiment can be accessed at https://gakuto101207.github.io/.

Funding: Funding for this study is provided by a Grant-In-Aid from the Japan Society for the Promotion of Science (#19KK0064), and by grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund).
Competing interests: We declare we have no competing interests.

1. Introduction

Music is often defined primarily in auditory terms (e.g., "humanly organized sound"; Blacking, 1976). Indeed, sound is consistently reported to be the most important information for evaluating musical performance (Murnighan et al., 1991; Sloboda, Lamont, & Greasley, 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; Vines et al., 2006; Leman, 2008; Savage et al., 2021). For example, visuals play an important role in evaluating musical performance, with elaborate costumes, make-up, and dancing characteristic of both traditional and contemporary music performance (Nettl, 2015). The popular international song competition is called "Eurovision", not "Eurosound" (cf. Haan et al., 2005).

Deleted: People rely on multiple senses to perceive and judge many aspects of daily life, and the importance of each sense depends on the situation. The social evaluation of other people can have important implications and consequences, impacting a range of outcomes from hiring decisions to political election results. Broadly, there has been interest in exploring what types of decision strategies may lead to better outcomes (Dane & Pratt, 2007; Rusou et al., 2013; Shapiro & Spence, 1997) and the conditions under which evaluations may be more influenced by one versus another channel of information (Harrigan, Wilson & Rosenthal, 2004; Tolsá-Caballero & Tsay, 2021). ¶
We focus on Mm

Deleted: S

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

[1] This Stage 1 Registered Report is a proposed protocol designed to be used for collecting full data after the initial protocol has been reviewed and approved. It includes a power analysis to determine what is a reasonable number of participants to recruit to appropriately balance logistical feasibility against the risks of false negative and false positive results. This involves terminology that may be unfamiliar to some readers without a background in statistics (e.g., "Cohen's d"; "80% power"). For accessible introductions to Registered Reports and power analysis, see Chambers (2019) and Braebart (2019), respectively.

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 (Wapnick et al., 1998; Bradley et al., 2006; Schutz et al., 2007; Goebl et al., 2009; Platz & Kopiez 2012, 2013; Tsay, 2013, 2014). 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).

Deleted: solo piano

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. 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%).

Formatted: Justified, Space After: 10 pt

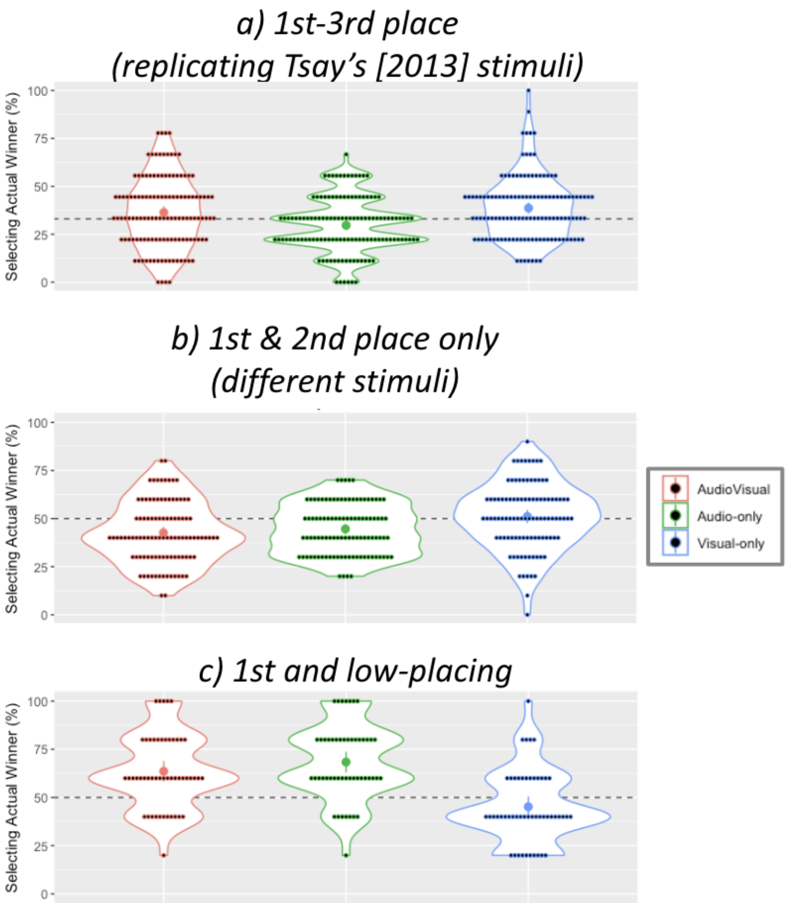

Figure 1. Violin plots visualizing Mehr et al.'s (2018) previous experimental results of sight vs. sound effects in judging piano performances (data were not visualized in the original publication). Panels a-c correspond to Studies 1-3 (see text for details). Dots indicate individual participants (a: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

1.1 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 performers 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.

Hypotheses

Based on previous findings from Western classical competitions described above (Tsay, 2013; Mehr et al., 2018), 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 visuals 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).

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). (see Methods).

Deleted: role

Deleted: visuals and

Deleted: a

Deleted: al tradition

Deleted: in which significantly different characteristics may be prioritized

Deleted: depart from the bulk of prior research using

Deleted: and study

Deleted: familiar with such norms (Henrich et al., 2010; Jacoby et al., 2020),

Deleted: instead focus

Deleted: washas a long history of persecution and discrimination in Japan as an instrument

Deleted: highly

Deleted: with

Deleted: ing

Deleted: Tsugaru shamisen was even featured in the popular 2016 animated movie "Kubo and the Two Strings". Contemporary Tsugaru shamisen performance is no longer restricted by disability status, but musicians retain traditions of oral transmission, performing while closing their eyes and focusing on sound.
Indeed, this tradition should be noted for how the lack of sight in its original performers is not just an ordinary part of its origin story. In fact, blindness has even come to be seen as indicative of a more authentic musician: ... [1]

Formatted: Font: Not Italic, Font color: Auto

Deleted: Given the historical development of a traditio ... [2]

Deleted: Tsugaru shamisen competitions represents a d ... [3]

Deleted: solo piano

Deleted: comparing low-variance (1st-3rd place) and hig ... [4]

Deleted: ; cf. Appendix for more details

Deleted:

Deleted: e.g.,

Deleted: H2:

Deleted: e.g.,

Deleted: 8th place or

Deleted: er

Deleted: H3: We predict that there is an interaction eff ... [5]

Deleted: H1: We predict that visuals will dominate the ... [6]

Deleted: H4-6: We will predict the same effects for the ... [7]

Deleted: tradition, which depart from most other music ... [8]

Deleted: Support for the null hypothesis would be eval ... [9]

Deleted: i

Deleted: In either case of null results or statistically ... [10]

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; Tsay, 2014; Tsay 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 (Amabile, Krabbenhoft, & Hogan, 2006; Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1993), especially the studies of visuals vs. sound in music competition evaluation described above (Tsay, 2013; Mehr et al. 2018).

Deleted: B

Deleted: and

Deleted: similar

Deleted: ,

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).

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/

Deleted: - 5 for the high-variance condition and 5 pairscompetitions and We aim to replicate and generalize the previous studies' results in Tsugaru shamisen performances. Therefore, we chose the sample to consider both audio/visual and high-variance/low-variance in our experiment. 6-second excerpts of the 1st, 2nd, and 3rd-place performers were used in Exp.3 of Tsay's study but paired clips of different stimuli (variation in quality is high) were used in Exp.3 of Mehr et al.'s study. To examine the effects of variance conditions, we adapt Mehr's experimental design (paired clips) and use both Tsay's (low-variance) stimuli and Mehr et al.'s (high-variance) stimuli. Also, we focus on only 5 competitions of the 9 competitions used in Tsay's experiment. One reason why is that the pairs used in Mehr et al.'s experiment have only 5. The other reasons why are that the same performer overlapped in Mehr et al.'s experiment, 2 violin competitions and 4-second excerpts are contained in Tsay's experiment. Therefore we use 10 paired clips (5 from Tsay's low-variance experiment, 5 from Mehr's high-variance experiment) of previous studies in our experiment, randomly choose 10 paired clip (5 low-variance, 5 high-variance) from Tsugaru shamisen competitions to suit them
we randomly selected brief 6s excerpts of5 pairs (1st/-place, 2nd-placed), performers and 5 pairs(1st/ lowest-placed)ing performers from 109 different national Tsugaru shamisen competition categories (Table 1),. we used brief 6s excerpts of their performances. For the 46 categories that did not rank performers beyond a certain place, we manually randomly selected one of the non-placing performances to maximize variance in quality (similar to Study 3 in Mehr et al. 2018). because the ranking range that Tsugaru shamisen performers can be the lowest placed performer is too wide in case of automatic selection (e.g., the best rank is 8 place and the worst rank is 57 place in non-placing performances). In case the 8th placed performer is selected by automatic selection and there is almost no difference in performance quality compared to the 1st placed performer, we won't probably get the results by performance quality gap (high-variance / low-variance)

Deleted: because (estimated average placing across all lowest-placing performances: 20th).

Deleted: 2

Deleted: national

Formatted: Font: Bold

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/

2.1.1 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 different 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. Participants will be randomly assigned 9 tasks, 3 from each of these types. In the Audio-only condition, only the sounds of "Tsugaru Jongara Bushi" by the three players are heard in succession, with no visual input. In the Visual-only condition, only

Deleted: ¶

We use a within-subjects design in which experimental participants all rated 72 paired clips, 10 Tsugaru shamisen competition categories divided into 5 categories (high-variance) and 5 categories (low-variance) in 3 domains (AudioVisual, Audio-only, Visual-only), 10 Piano competition categories (used in previous study) divided into high-variance/low-variance in 3 domains, and 4 Violin & Piano competition categories (used in previous study) of low-variance in 3 domains.We use a within-subjects design in which experimental participants all rated 3 audio-only, 3 video-only, and 3 audiovisual competitions, Thesethese choiceschoice and order of which were randomly assigned but AudioVisual condition tasks were assigned after Audio-only and Visual-only condition tasks because they were used for exploratory analysis.

Deleted: (see Fig. 2 for an overview)

Deleted: ¶
Figure 2. Overview of the experimental paradigm for rating 7227 performances from 9Tsugaru shamisen, Piano and Violin competition categories based on 6s excerpts of audio-only, video-only, or audiovisual information.¶

Deleted: *1*

Deleted: .

Deleted: *One*The independent variable is the

Deleted: ,

Deleted: and

Deleted: another is

Deleted: The Audio-Visual condition is also applied during experiments but this data is only used for exploratory analysis purposes and is excluded at the hypothesis testing.

Deleted: that

Deleted: s

Deleted: and

Deleted: ana

the three players are displayed on the video screen in succession, with no auditory input. In the AudioVisual condition, three performance videos with sound are presented. In these three conditions, participants are asked to evaluate all performances.

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 (Noguchi et al., 2012; Friedrich et al., 2017; Brunner et al., 2018). 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-

Deleted: "Tsugaru Jongara Bushi" is chosen because it is the most famous song among Tsugaru shamisen players, and it is a compulsory component of all competitions, which allows us to collect a large number of comparable samples from limited performance videos.

Deleted: ¶

Deleted: *2*

Deleted: *s*

Deleted: two

Deleted: s

Deleted: 1)

Deleted: , and 2) the lowest-placed performer

Deleted: We consider t

Deleted: After judging 3 performers who are displayed in random order, participants will select who they believe was the highest- and lowest-ranking performer.

Deleted: 9 such

Deleted: *3*

Deleted: *3*

Deleted: *and H4*

Deleted: *per instrument*

Deleted: we consider

Deleted: does

Deleted: matter

Deleted: 3

Deleted: and H5-H6

Deleted: per

Deleted: and instrument

Formatted: Font: Bold, Not Italic

Deleted: p

Deleted: and H4

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 (Schuirmann, 1987; Lakens, 2017) 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 prespecified 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 (Schuirmann, 1987; Lakens, 2017). 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 (Smithson, 2001; Campbell & Lakens, 2021). 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. $F(df_1, \infty)$). 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

Deleted: and H5

Deleted: To test our predictions, we will follow previous analyses of similar paradigms (Tsay, 2013; Mehr et al., 2018) by performing paired t-tests of both dependent variables between the two key conditions (audio-only and visual-only; the audiovisual condition is used as a control for interpreting data, but is not specifically relevant for hypothesis testing).

Deleted: 3

Deleted: two

Deleted: *t*-

Deleted: 25

Deleted: 3

Deleted: of H2-H3 and H5-H6

Deleted: W

Deleted: equivalent

Deleted: the

Deleted: when the null hypothesis is not rejected

Deleted: Note that methods analogous to equivalence testing for the proposed interaction effect (H1) are not yet available (

Deleted: et al.

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 and 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 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

Deleted: *4*

Deleted: *2.54.1 Presumptions*

Deleted: two sources of evidencecalculate power analyses assuming an to estimate effect size: 1) previous guidelines suggesting of Cohen's d of 0.4 as an approximation of the "smallest effect size of interest" (Lakens, 2017); 2

Deleted: To achieve

Deleted: 95

Deleted: given

Deleted: n

Deleted: 25

Deleted: , the two-tailed nature of the hypotheses, and our within-subject design, we will require data from at least 97 participants

Deleted: *Since our experimental design is similar to Mehr et al. (2018)'s study, we will use their data to inform the power analysis and we will aim to set a power of 80% (N.B. α-level is 0.0083 explained previously). Specifically, we use their Experiment 2 and Experiment 3 data for our power analysis since our experimental design follows those experimental designs. However, we will use the stimuli used in Experiment 1 which was originally used in Tsay (2013)'s experiments for our low-variance piano stimuli so a difference in the actual effect and the estimated effect would potentially exist due to the difference of the stimuli. In addition, their experiments are conducted by between-subjects designs which differs from our study, but we will estimate necessary sample sizes for our study as if we will conduct non-repeated measurements between-subjects designs since there is no other information currently we can rely on. Using the sample size based on the between-subjects design assumption would possibly raise the power of our study higher than 80% because within-subjects designs are considered to potentially have higher power than between-subjects designs (Lakens, 2013). Lastly, we put an additional assumption for the sample size estimation that the effect size to be observed in Tsugaru shamisen would be the same with piano, which is an instrument studied in Mehr et al. (2018)'s analysis.*
Though we collected the pilot data which used the planned experimental design, the number of samples was only 9. Therefore we decided to rely on larger data even though the experimental design is not compatible. However, as mentioned above, we consider the sample size estimation based on the assumption that the between-subjects des ... [11]

Deleted: A priori power analysis requires estimating the effect size before collecting data, which is notoriously difficult. In this paper, we calculate power analyses as ... [12]

Deleted: ***2.5 Sample choice***
We aim to replicate and generalize the previous studies' results in Tsugaru shamisen performances. Therefore, ... [13]

Deleted: The number of the uploaded videos in Tsugaru shamisen competitions is small, and the public availability of videos is also limited. Among them, we selected the th ... [14]

Formatted: Font color: Auto

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 3 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.

Formatted: Font: Not Italic

Deleted: s

Deleted: 11

Deleted: before we updated the stimuli to increase the variance of trials…

Deleted: and non-winner

Deleted: the three

Deleted: s

Deleted: ,

Deleted: ,

Deleted: AudioVisual

Deleted: though

Deleted: demonstrates

Deleted: can also

Deleted: at

Deleted: show

Deleted: of

Deleted: but

Deleted: was much

Deleted: This would potentially violate the above assumption of the same level of effect between piano and Tsugaru-shamisen, though it would not affect the judgment that effects of less than approximately Cohen's d = 0.4 may be considered too small to be of practical interest.

Deleted: Differences between the audio-only and visual-only conditions are currently too small to be able to draw meaningful conclusions regarding our hypotheses from this small pilot, though they seem suggestive that at least when selecting the lowest-placing performer, participants appear to be able to achieve this at levels above chance (33%, indicated by the dashed line in Fig. 3).

Deleted: This

Deleted: results

Deleted: functions

Deleted: as a positive control to

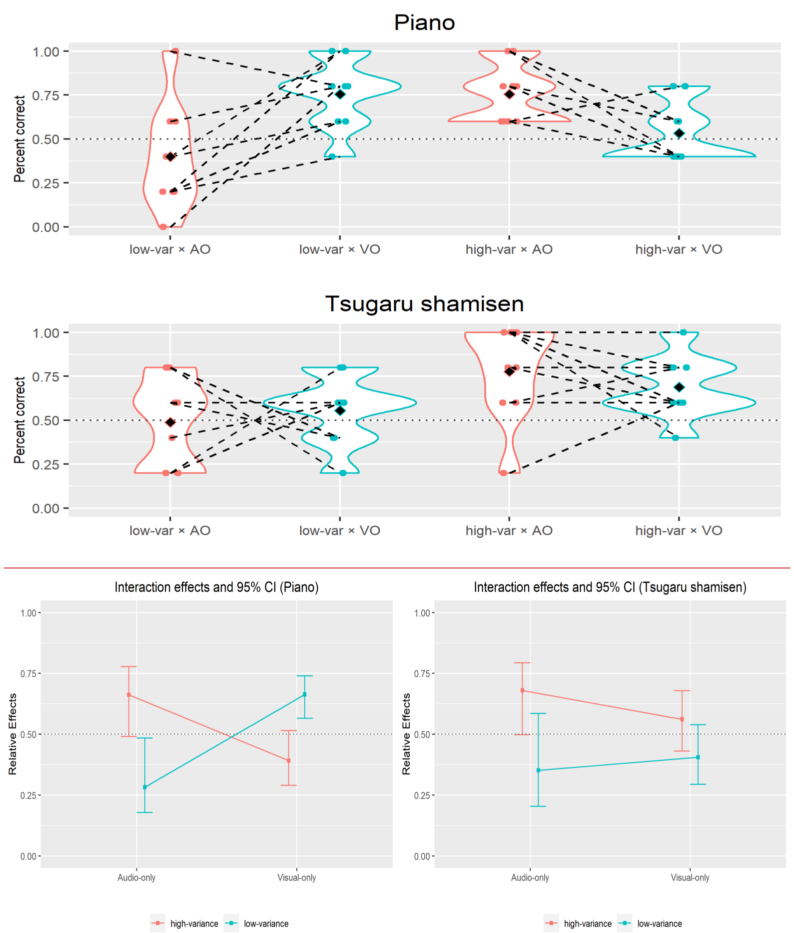

Figure 3. The top figure is the violin plots of the pilot data (n = 9). Black diamonds indicate mean values. Dashed lines indicate paired data from the same participant. The bottom two figures show the relative effects of piano (left) and shamisen (right), and the bars are 95% confidence intervals based 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.

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).

Table 2 | Registered Report design planner

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given 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.
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.
(Same as above)	H3: Sound dominates the judgment of piano performance between upper and lower ranks		Same as above	

Deleted: audio-visual dominance appears in

Deleted: are interested in

Deleted: is a

Deleted: so we will analyze that

Deleted: 60

Deleted: piano

Deleted: piano

Deleted: e.g.,

Deleted: , but for rate correctly selecting the lowest-placed performer

Deleted: Tsugaru Shamisen

(1st place vs. low-placing), due to the high variance in trials.

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

Deleted: e.g.,

Deleted: or 2nd

Deleted: 8th place or lower

Deleted: Same as

Deleted: but the target music is

Ethics:
We have approval of the Keio University Shonan Fujisawa Campus Institutional Review Board to PES (approval #298). All pilot participants provided informed consent and all future participants will provide informed consent.

Data/code availability:
Pilot data and videos are available at https://osf.io/p9fvs/
Analysis code is 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, collected pilot data]
Analysis: Yuto Ozaki, Gakuto Chiba, Patrick E. Savage
Writing –original draft: Patrick E. Savage, Gakuto Chiba, Yuto Ozaki
Writing –reviewing/editing: Shinya Fujii
Project administration/supervision/funding acquisition: Patrick E. Savage

Deleted: , Yuto Ozaki

Deleted: Gakuto Chiba,

Competing interests. We declare we have no competing interests.

Acknowledgments. We thank Chia-Jung Tsay, Kyoshiro Sasaki, and David Hughes 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.

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

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*, 32, 201-272. San Diego, CA: Academic Press.

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.

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.

Ballew, C. C., Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences,* 104(46), 17948-17953.

Bergeron, V., Lopes, D. M. (2009). Hearing and seeing musical expression. *Philosophy and Phenomenological Research*, 78(1), 1-16.

Booth, A. (1995). *Looking for the Lost: Journeys Through a Vanishing Japan*. New York: Kodansha America. 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

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://ci.nii.ac.jp/ncid/BB28708839

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

Campanella, S., Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences*, 11(12), 535-543.

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

Chambers, C. (2019). What's next for Registered Reports? Nature, 573(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Collignon, O., et al. (2008). Audio-visual integration of emotion expression. *Brain research*, 1242, 126-135.

de Gelder, B., Böcker. K, B., 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.

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

Dane, E., Pratt, M. G. (2007). Exploring Intuition and Its Role in Managerial Decision Making. *The Academy of Management Review*, 32(1), 33-54. https://doi.org/10.2307/20159279

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

Deleted: ¶

Deleted: arlan

Deleted: aniël

Formatted: Font: Italic

Deleted: ¶

Ginsburgh, V. A., Van Ours, J. C. (2003). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review*, 93(1), 289-296.

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.

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

Harrigan, J. A., Wilson, K., & Rosenthal, R. (2004). Detecting state and trait anxiety from auditory and visual cues: A meta-analysis. *Personality and Social Psychology Bulletin*, 30(1), 56-66.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-135. https://doi.org/10.1017/S0140525X0999152X

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*, 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., 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, 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.

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

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Deleted:

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.

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 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, K., Conlon, D. (1991). The dynamics of intense work groups: A study of British string quartets, *Administrative Science Quarterly,* 36(2), 165-186.

Nettl, B. (2015). *The study of ethnomusicology: Thirty-three discussions,* (3rd ed.). University of Illinois Press.Noether, G. E. (1987). Sample Size Determination for Some Common Nonparametric Tests. Journal of the American Statistical Association, 82(398), 645–647. https://doi.org/10.1080/01621459.1987.10478478

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, 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

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.

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.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

Rusou, Z., Zakay, D., & Usher, M. (2013). Pitting intuitive and analytical thinking against each other: The case of transitivity. *Psychonomic Bulletin & Review*, 20(3), 608-614. https://doi.org/10.3758/s13423-013-0382-7Savage, 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/S0140525X21000042Schuirmann, 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

Shapiro, S., & Spence, M. T. (1997). Managerial intuition: A conceptual and operational framework. *Business Horizons*, 40(1), 63-68. https://doi.org/10.1016/S0007-6813(97)90027-6

Deleted: ¶

Deleted: ¶

Formatted: Font color: Gray-87.5%, Highlight

Formatted: Font color: Gray-87.5%, Highlight

Formatted: Font color: Gray-87.5%, Highlight

Formatted: Font color: Gray-87.5%, Highlight

Deleted: ¶

Deleted: ¶

Deleted: ¶

Sloboda, J. A., Lamont, A., Greasley, A. E. (2008). *The Oxford Handbook of Music Psychology*, eds Hallam S, Cross I, Thaut M (Oxford Univ Press, Oxford), 431-440. 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

Thompson, W. F., Russo, F. A. (2007). Facing the music. Psychological Science, 18(9), 756-757.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626.

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

Tsay, C. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences,* pmid:23959902.

Tsay, C. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes,* 124(1), 24-33.

Tsay, C. (2021). Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries*, 7(3), 1-23. 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. doi: 10.2307/3345347.

Deleted: ¶

Deleted: ¶

Page 6: [1] Deleted Patrick Savage 12/8/21 8:30:00 AM

Page 6: [2] Deleted Gakuto Chiba 11/17/21 2:55:00 PM

Page 6: [3] Deleted Patrick Savage 12/8/21 8:41:00 AM

Page 6: [4] Deleted Patrick Savage 12/8/21 9:44:00 AM

Page 6: [5] Deleted Patrick Savage 12/8/21 9:44:00 AM

Page 6: [6] Deleted Yuto Ozaki 11/26/21 9:26:00 AM

Page 6: [7] Deleted Patrick Savage 12/10/21 1:14:00 AM

Page 6: [8] Deleted Gakuto Chiba 11/17/21 11:04:00 PM

Page 6: [9] Deleted Patrick Savage 12/10/21 1:21:00 AM

Page 6: [10] Deleted Patrick Savage 12/10/21 1:21:00 AM

Page 13: [11] Deleted Patrick Savage 12/10/21 4:54:00 AM

Page 13: [12] Deleted Yuto Ozaki 11/26/21 4:10:00 PM

Page 13: [13] Deleted Patrick Savage 12/10/21 1:25:00 AM

Page 13: [14] Deleted Gakuto Chiba 12/7/21 4:55:00 AM