Supplementary materials

S1 Supplementary results

S1.1 The difference in the percent accuracy within participants.

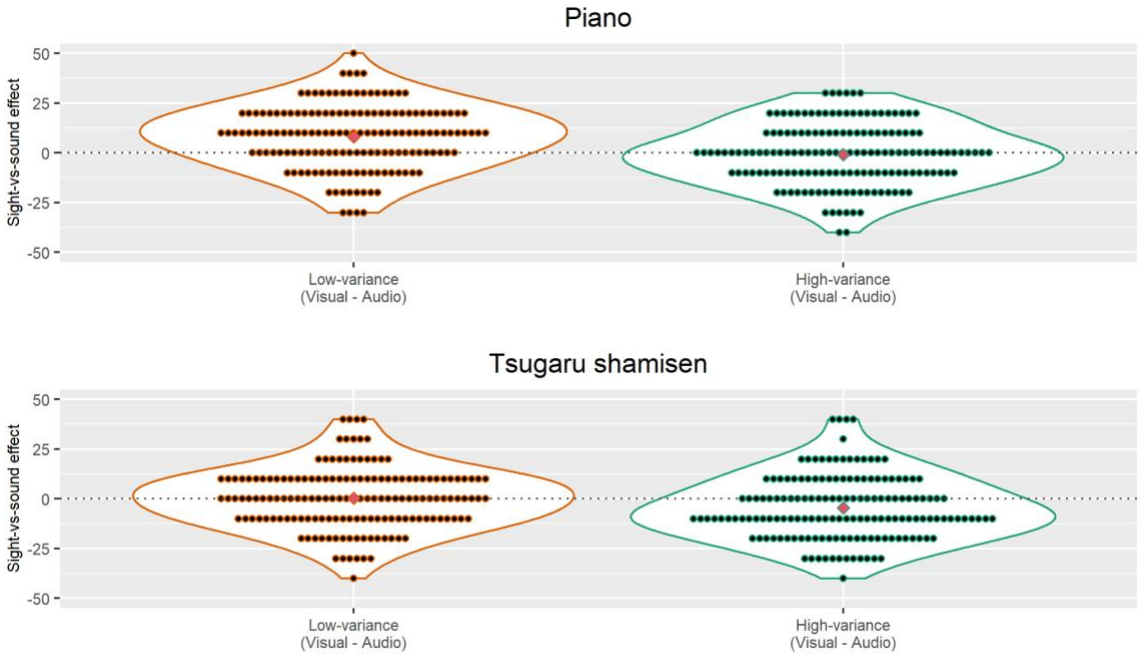

Figure S1. The difference in the percent accuracy between the visual-only condition and the audio-only condition at each variance condition within participants (n = 155). The difference is measured by subtracting the percent accuracy of the audio-only condition from the visual-only condition. Larger positive values indicate higher accuracy in the visual-only condition, and larger negative values indicate higher accuracy in the audio-only condition. Red diamonds indicate mean values.

S1.2 Alternative version of the confirmatory analysis (H1-6) including the exploratory stimuli data.

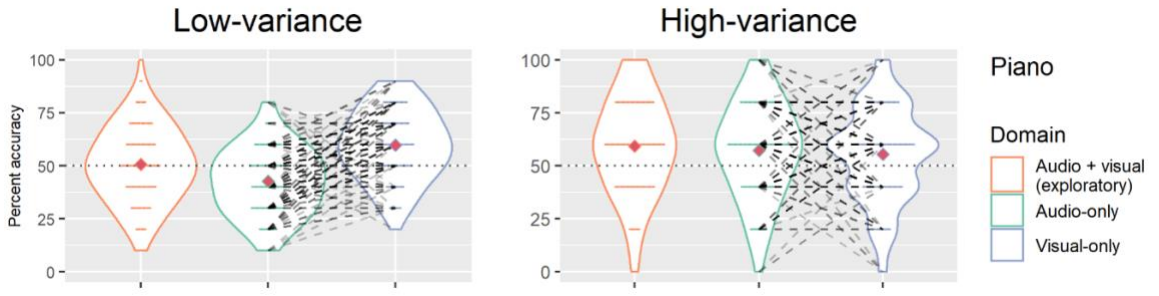

Figure S2. The violin plots of the full data (n = 155) including the 5 exploratory stimuli of the low-variance condition (see 2.1.1 Exploratory samples). Red diamonds indicate mean values. Dashed lines indicate paired data from the same participant.

Table S1| Summary of the hypotheses tests and obtained effect sizes: confirmatory analysis including the exploratory stimuli data

#	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	45.40	***1.6x10^{-11}**	-	Adjusted $\eta^2_{partial}$	0.070
H2	Relative effect	0.77	***< 1.0x10^{-15}**	-	Cohen's D	1.0
H3	Relative effect	0.48	.22	***0.48-0.53**	Cohen's D	-0.085

S1.3 Competition-wise average percent accuracy.

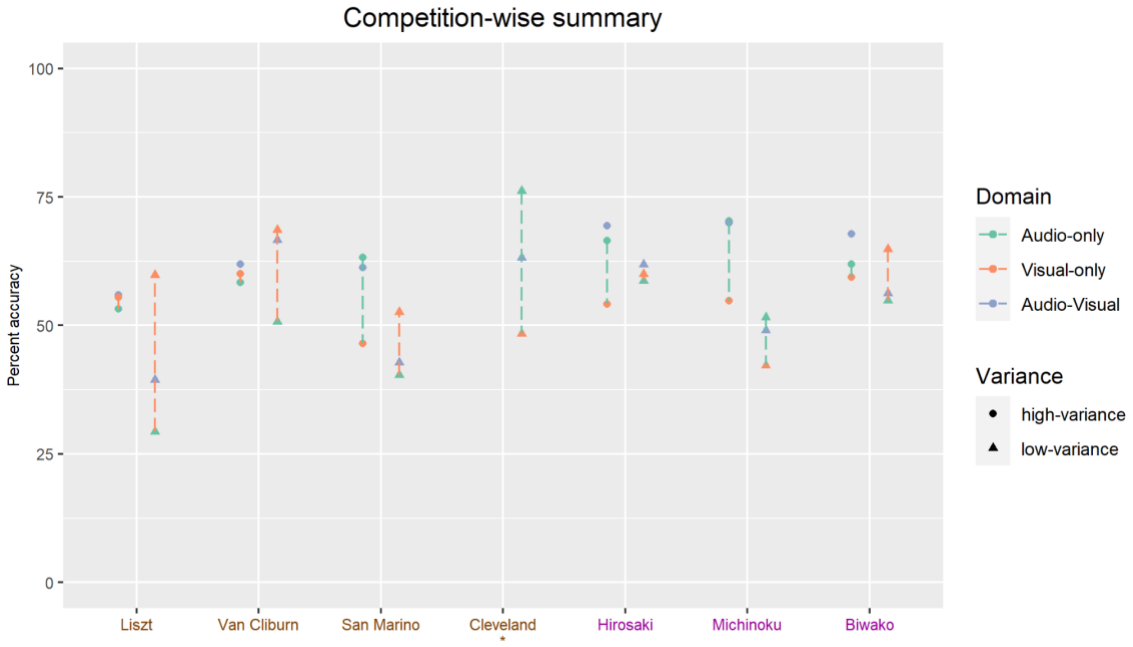

Figure S3. The average percent accuracy at each competition. The x-axis labels show the partial name of competitions and whether they are piano competitions (brown color) or Tsugaru-shamisen competitions (purple color). The asterisk indicates the stimuli only appearing in the exploratory analysis. The 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.

S1.4 Re-run of the confirmatory analysis (H2-3, H5-6) with paired t-tests.

Table S2 | Summary of the hypotheses tests and obtained effect sizes: confirmatory analysis with paired t-tests.

#	Test statistic	Obtained statistic	p-value (α=0.05/6)	Effect size translation	Obtained effect size
H2	t-statistic	6.4	***1.1x10^{-9}**	Cohen's D	0.51
H3	t-statistic	0.73	.23	Cohen's D	0.059
H5	t-statistic	0.21	.42	Cohen's D	0.017
H6	t-statistic	3.5	***2.8x10^{-4}**	Cohen's D	0.28

S1.5 Alternative version of the confirmatory analysis (H1-6) including data from 10 participants (total n = 155 + 10 = 165 participants) excluded based on criteria not explicitly stated in the Stage 1 protocol (repeated experiments by the same participants or participants completing the experiment in impossibly fast times).

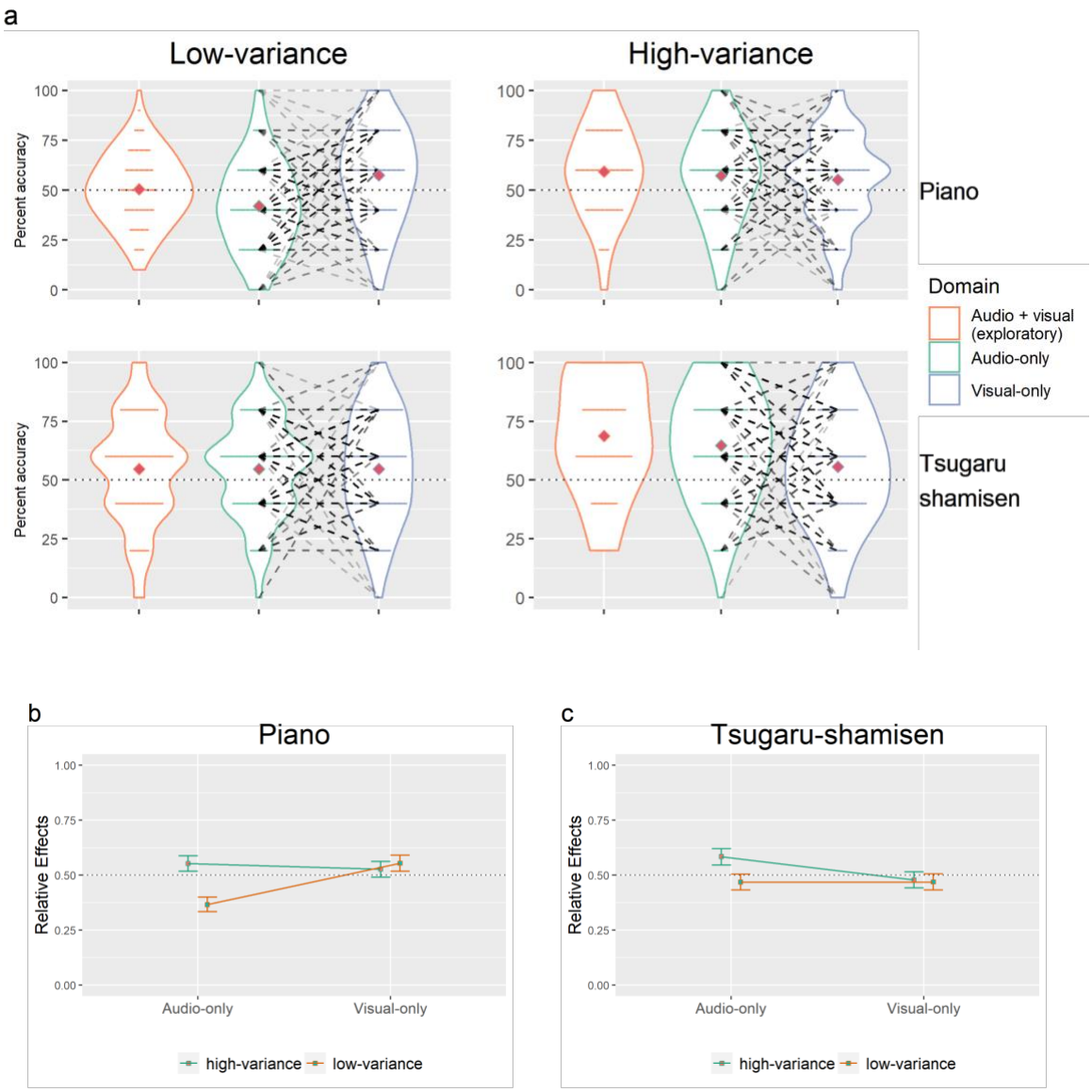

Figure S4. The top figure (a) shows violin plots of the data (n = 165 participants), including the participants excluded from the confirmatory analysis (cf. 3.1 for the details), for the dependent variable of % correctly choosing the 1st-placed performer in a two-choice forced choice task. Red diamonds 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 shamisen (c), and the bars are 95% confidence intervals based on the ANOVA-type statistics. Dashed lines (q = 0.5) indicate there is no effect.

Table S3 | Summary of the hypotheses tests and obtained effect sizes: confirmatory analysis including the excluded data.

#	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	24.92	***6.0x10^{-7}**	-	Adjusted η^2_{partial}	0.035
H2	Relative effect	0.68	***< 1.0x10^{-15}**	-	Cohen's D	0.68
H3	Relative effect	0.47	.18	***0.47-0.52**	Cohen's D	-0.10
H4	ANOVA-type statistic	7.58	***5.9x10^{-3}**	-	Adjusted η^2_{partial}	0.010
H5	Relative effect	0.50	.51	***0.45-0.55**	Cohen's D	-0.0035
H6	Relative effect	0.40	***< 1.0x10^{-15}**	-	Cohen's D	-0.37

S1.6 Belief about the judgment of musical performance.

Participants of Part 2 were asked to choose which audio and visual more important is to evaluate the musical performances after the whole experiment. Amongst the 160 participants, 146 participants (around 91% of all participants) answered that audio information is more important than visual information. In addition, based on the authors' impression, we observed that the free description comments regarding the factors being weighed during the experiments collected by the Part 2 participants also reflected that they put more value on the audio information rather than visual information. Please note that the same exclusion criteria of the confirmatory analysis explained in 3.1 were applied to this analysis (see Section 3.2 for explanation of the discrepancy between n=160 participants for exploratory analysis vs. n=155 participants for confirmatory analysis).

S1.7 Testing for deviation from chance.

Following previous studies (Tsay, 2013; Mehr et al., 2018; Wilbiks & Yi, 2022), we ran two-sided one-sample t-tests to check whether the percent accuracy of each pattern is significantly different from the chance level (50%). Before running the tests, we checked whether two-sided one-sample t-tests can appropriately control type I error rate with our data by simulation. In the simulation, we first assign probability mass to realizable percent accuracy (0.0, 0.2, 0.4, 0.6, 0.8, and 1.0) and the assignment is randomly drawn from the Dirichlet distribution (Dir(α_1, α_2, α_3, α_4, α_5, α_6), α_k=1). For each randomly generated distribution of percent accuracy, we obtain the mean value of the distribution and run Monte Carlo simulations (n = 4,096) that samples 155 percent accuracy scores (same as our sample size) from the PMF to check whether two-sided one-sample t-test can maintain the nominal type error rate under the null hypothesis being true (i.e. the confidence interval constructed on the t-value based on i.i.d. random samples includes the mean value). We confirmed one-sample t-tests could control type I error rate appropriately including the cases that the distribution of percent accuracy radically differs from the bell-shaped curve (cf. Figure S5).

Table S4 | Summary of one-sample t-tests (n = 155) testing whether the percent accuracy of each pattern is significantly different from chance level (50%).

Instrument	Domain	Variance	t-statistic	p-value	Estimated mean percent accuracy
Piano	Audio-only	High	4.1	7.7×10^{-5}	57%
		Low	-4.0	2.0×10^{-5}	42%
	Visual-only	High	3.1	2.2×10^{-3}	55%
		Low	4.5	1.4×10^{-5}	58%
	Audio and visual	High	5.2	6.1×10^{-7}	59%
		Low	0.45	.65	51%
Tsugaru shamisen	Audio-only	High	8.1	2.0×10^{-13}	65%
		Low	2.6	9.4×10^{-3}	54%
	Visual-only	High	3.3	1.1×10^{-3}	56%
		Low	2.7	7.4×10^{-3}	55%
	Audio and visual	High	9.7	$< 2.2 \times 10^{-16}$	69%
		Low	2.5	1.4×10^{-2}	55%

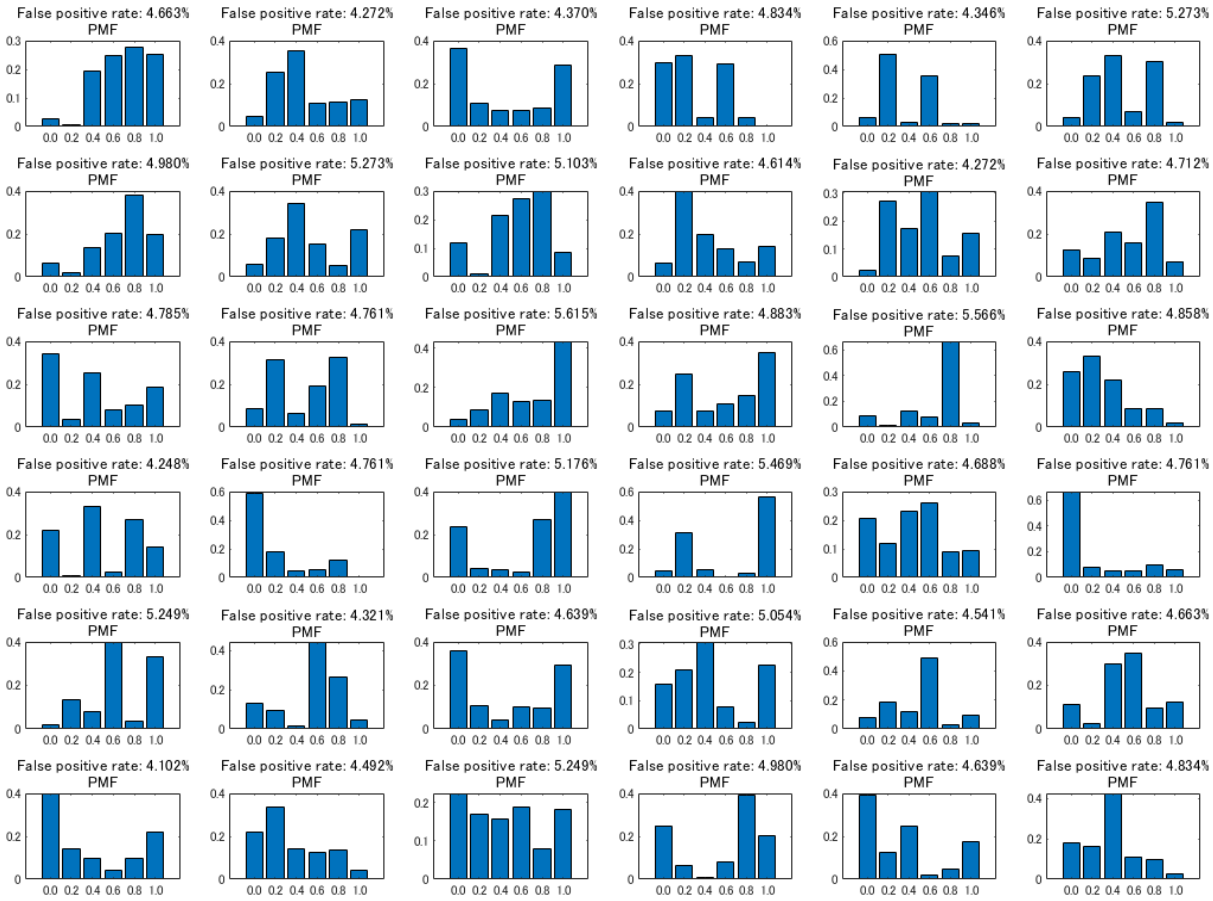

Figure S5. Validation of the use of two-sided one-sample t-tests. PMF stands for probability mass function which represents the randomly generated distribution of percent accuracy, and the title labels show simulated type I error rates for each PMF. Clearly, the distributions of percent accuracy can differ from the normal distribution, but our simulation indicates t-test can maintain the nominal type I error rate in general.

S1.8 Nonparametric multiple comparisons controlling artifacts by nontransitive paradox.

Our original testing plan stipulated (1) testing interaction effects and performing two pairwise nonparametric tests for the experiments with piano stimuli and Tsugaru-shamisen stimuli, (2) grouping six hypotheses as a single family, and (3) employing Bonferroni's correction to control family-wise error rate. However, considering the hypotheses of interaction effects between the domain and variance (H1/H4), audio-only vs. visual-only under the low-variance condition (H2/H5), and audio-only vs. visual-only under the high-variance condition (H3/H6) as a family of multiple hypotheses testing, a more nuanced test can be conducted with control of potential nontransitive paradox (Noguchi et al., 2020). A nontransitive paradox is a paradox that does not preserve transitivity among multiple comparisons (e.g. A < B, B < C, but C < A, Blyth, 1972; Brown & Hettmansperger, 2002; Noguchi et al., 2020). This paradox can arise when multiple comparisons are executed with nonparametric statistics, especially when test statistics only measures pairwise relative superiority in each comparison. Noguchi et al.'s method of multiple

comparisons controlling for the nontransitivity paradox can inform the consistent and overall superiority of samples among all pairs, and we re-analyzed our data with their multiple nonparametric comparisons with the R package nparcomp (Konietschke et al., 2015) as an exploratory analysis. In our case, this approach allows us to measure the superiority of percent accuracy among audio/visual × high/low-variance conditions, which more accurately demonstrates the effects of domain and variance. Moreover, we divide the original family into two families by instruments since we are not interested in performing multiple comparisons beyond instruments (e.g. piano audio-only under high-variance vs.Tsugaru-shamisen visual-only under high-variance). In other words, we define families in this exploratory analysis according to the unit of multiple comparisons and not by the set of multiple inferences as in the original text. Noguchi et al.'s method models p-values of multiple hypotheses simultaneously using multivariate distributions, so family-wise error control such as Bonferroni's correction is not necessary and the nominal alpha-level (α = .05 in our case) can be directly applied to interpret the result. This exploratory analysis confirmed the same results obtained in our main analysis which eliminates the possibility that our observation of the dependency of domain and variance via multiple comparisons is an artifact of a nontransitivity paradox.

Table S5-6 | Summary of nonparametric multiple comparisons (Noguchi et al., 2020). Please note this method uses test statistics and effect sizes different from the method used in the main analysis. The log odds ratio is used for effect sizes but this value is adjusted to approximate Cohen's *d* (Chinn, 2000; Noguchi et al., 2020; Sánchez-Meca et al., 2003). Therefore the rejection region of equivalence testing is also based on quantification by Cohen's *d*.

(Piano)

#	Test statistic	Obtained statistic	p-value (α=0.05)	Effect size	Obtained effect size	90% CI for equivalence testing (rejection region -0.4~0.4)
H2	Student's t-type test statistics	6.44	***4.7x10^{-10}**	Log odds ratio of relative effects	0.48	-
H3	Student's t-type test statistics	0.72	.42	Log odds ratio of relative effects	0.052	***-0.089 - 0.19**

(Tsugaru-shamisen)

#	Test statistic	Obtained statistic	p-value (α=0.05)	Effect size	Obtained effect size	90% CI for equivalence testing (rejection region -0.4~0.4)
H5	Student's t-type test statistics	0.24	.65	Log odds ratio of relative effects	0.017	***-0.12 - 0.16**
H6	Student's t-type test statistics	3.26	***1.2x10^{-3}**	Log odds ratio of relative effects	0.25	-