Conceptual Boundaries of Music:

A Behavioral Study of Cross-Cultural Sound Classification

Carlos Y. Hsu Speck
Independent Researcher
Contact: speckcyh@musiccognition.org.

Author Note

Carlos Y. Hsu Speck ORCID: https://orcid.org/0009-0002-5940-6155

I gratefully acknowledge Trina Keil for meeting with me to discuss the general project; Professor Samuel Norman-Haignere for granting permission to use audio samples from (Norman-Haignere et al, 2015) in this study; and the Association for Cultural Equity for granting permission to use brief audio excerpts from The Global Jukebox in this study.

Correspondence concerning this article should be addressed to Carlos Hsu Speck, speckcyh@musiccognition.org.

Abstract

Music is a cultural universal, yet the conceptual boundary between “music” and “non-music” remains theoretically unsettled. This study used behavioral measures to examine the perceptual judgments of listeners from Brazil, China, and the United States as they classified short sound excerpts and rated their perceived musicality. Participants (N = 103) completed an online survey featuring 24 two-second stimuli drawn from 11 sound categories, including traditional music from 130 societies and non-musical sounds from established auditory taxonomies. Descriptively, the mean proportion of “music” classifications was 0.51, with substantial category-level variation: traditional vocal and non-vocal music received the highest endorsement rates, whereas speech and environmental sounds received the lowest. Musicality ratings showed a parallel pattern. Inter-participant agreement, quantified using Krippendorff’s α, was moderate at the aggregate level (α = .73) but generally low within categories.

To complement these descriptive patterns, mixed-effects models were used to assess whether sound category or country predicted responses while accounting for participant- and stimulus-level variability. Both models revealed large, reliable effects of sound category: the two traditional music categories were far more likely to be judged as music and received substantially higher musicality ratings than all other categories. No other category differed significantly from the reference, and country effects were small or nonsignificant. Together, the descriptive and inferential results indicate that listeners share strong intuitions about prototypical musical sounds, but judgments become highly heterogeneous for ambiguous cases, suggesting that the concept of music may be graded and context-dependent rather than sharply bounded.

Keywords: perceptual judgment, cognitive categorization, auditory perception, behavioral data, cross-cultural cognition, music perception

Across every known human society, music is an "absolute" cultural universal (Savage et al., 2015). From lullabies in the Amazon to ritual drumming in West Africa, musical expression appears in every ethnographic record, although its form varies widely within and between cultures (Mehr et al., 2019). Despite this universality, there remains no single, agreed-upon definition of what music actually is (Savage et al., 2015).

This conceptual ambiguity presents a profound challenge for researchers: how do they define music in a way that respects its global diversity while probing the cognitive and perceptual boundaries that shape its recognition? Prior work has examined categorization of environmental sounds (Bones et al, 2013) and speech–song continua (Patel, 2003), but the author could find no studies that directly tested how people classify culturally diverse sound samples as “music” or “non-music.”

The present study addresses this gap by asking: *Do individuals share a conceptual boundary for music across culturally diverse sound samples?* We designed a behavioral experiment in which participants judged short sound excerpts from traditional societies around the world, as well as other sounds.

Methods

Stimuli

The stimulus set comprised 240 auditory stimuli, each standardized to a duration of two seconds. The set included 110 non-musical sounds and 130 musical sounds, as labeled by the investigator.

Non-musical stimuli were drawn from the dataset used by Norman-Haignere et al. (2015). These publicly available WAV files, each originally two seconds in length, were obtained directly from the study’s GitHub repository and used without modification.

Musical stimuli were sourced from The Global Jukebox (globaljukebox.org), a large ethnomusicological archive founded by Alan Lomax that documents traditional music and expressive culture from more than 1,000 societies worldwide. Each musical stimulus represented a unique society randomly selected from the more than 1,070 societies indexed in the archive (Figure 1). The musical stimuli encompassed both canonical musical examples (e.g., drumming ensembles, flute melodies) and borderline cases (e.g., chant-like speech, ritual vocalizations, work songs). For each selected society, a two-second excerpt was randomly extracted from the first musical recording listed for that culture. Prior to excerpt selection, silent segments at the beginnings and ends of recordings were trimmed using the Python library Pydub; no additional preprocessing was applied.

Figure 1:

World Map Showing Geographic Distribution of Sampled Societies

All stimuli, musical and non-musical, were assigned to one of eleven sound categories. Non-musical stimuli were categorized according to the nine sound categories defined by

Norman-Haignere et al. (2015). Musical stimuli were assigned to one of two additional categories based on the presence or absence of human vocals. Table 1 summarizes the resulting category structure.

Table 1:

Stimulus Sound Categories

Sound Category	Examples
Animal non-vocal	*Dog drinking, wings flapping*
Animal vocal	*Dog barking, puppy whining*
English speech	*Background speech, girl speaking*
Environmental sound	*Crumpling paper, dishes clanking*
Foreign speech	*Spanish, French*
Human non-vocal	*Finger tapping, door knocking*
Nonspeech human vocal	*Crying, baby crying*
Mechanical	*Cutting with scissors, cellphone vibrating*
Traditional music (non-vocal)	*Ibiza non-vocal music, Lucania non-vocal music*
Nature	*Wind, water splashing*
Traditional music (vocal)	*Kerala vocal music, Michoacan vocal music*

Survey Instrument

The survey was developed and administered online using the LimeSurvey platform (limesurvey.org). The instrument began with a disclosure statement, followed by twelve demographic items collectively labeled Preliminary Questions. Wording for these items was adapted from instruments created by Sam Mehr for The Music Lab (themusiclab.org). When applicable, demographic responses were cross-referenced with information supplied by Prolific (see Participants).

The disclosure statement described the study's purpose, procedures, potential risks, and confidentiality safeguards. Only individuals who provided informed consent were permitted to continue to the remainder of the survey. At the time the instrument was initially drafted, participant compensation had not yet been planned. Consequently, the disclosure statement incorrectly stated that participants "will not be compensated"; however, compensation was ultimately provided.

Participants were asked to provide a self-assessment of their listening skills using a 5-point scale. The item read: How good do you think your listening skills are? (This includes things like remembering melodies, hearing out-of-tune notes, or detecting a beat that is out of sync with the music.)

After completing the demographic section, participants proceeded through twenty-four question groups, each presented on a separate page. Each question group included:

- A two-second audio stimulus accompanied by the instruction, “Listen to Sample [Number]”;
- A binary judgment task (“Do you consider this sound to be music?”; Yes/No); and
- A Likert-type rating (“How musical do you consider this sound to be?”; 1–7, with 7 indicating “highly musical”).

The order of response options for the binary classification task (Yes/No) was randomized across participants.

Because LimeSurvey does not support randomization of question order, ten distinct survey versions (henceforth, “Surveys”) were created. Each version contained 24 stimuli drawn in fixed proportions from the eleven sound categories (Table 2), but the order of category presentation differed across versions

(Table 3). Table 4 provides examples of how specific stimuli were distributed across survey versions.

Table 2

Number of Stimuli per Survey by Sound Category and Type

Sound Category	Number of Stimuli Per Survey	Type	Number of Stimuli Per Survey
Animal non-vocal	1		
Animal vocal	1		
English speech	1		
Environmental sound	2		
Foreign speech	1	Non-music	11
Human non-vocal	1		
Nonspeech human vocal	1		
Mechanical	2		
Nature	1		
Traditional music (non-vocal)	4	Music	13
Traditional music (vocal)	9		
Total	24		24

Table 3

Order of Stimuli by Sound Category Across Surveys

Sound Category	Survey									
	Order	1	2	3	4	5	6	7	8	9	10
1 Animal Non-Vocal	1	3	8	9	8	11	11	2	9	4	11
2 Animal Vocal	2	2	7	11	8	1	11	11	5	10	11
3 English Speech	3	6	11	5	9	11	6	9	11	11	9
4 Environmental Sounds	4	8	9	11	9	8	10	11	11	11	4
5 Foreign Speech	5	11	8	4	4	9	9	9	10	11	9
6 Human Non-Vocal	6	1	1	11	11	3	9	3	11	8	5
7 Human Vocal	7	11	3	11	3	8	8	4	9	2	11
8 Mechanical	8	8	9	9	11	10	7	1	11	9	11
9 Traditional Music (non-vocal)	9	11	11	1	10	11	9	8	6	8	8
10 Nature	10	5	4	11	9	11	11	11	8	4	11
11 Traditional Music (vocal)	11	11	5	11	2	11	8	11	11	11	9
	12	4	11	8	1	5	11	11	2	11	7
	13	11	9	11	11	9	11	11	11	3	3
	14	11	10	10	9	11	4	11	1	9	11
	15	11	11	8	11	9	2	11	11	7	9
	16	4	11	4	11	4	3	6	11	11	10
	17	7	4	2	11	6	9	9	8	9	11
	18	9	11	9	11	11	11	10	4	9	2
	19	11	11	3	6	7	1	9	11	5	1
	20	11	9	11	4	4	4	8	9	11	11
	21	9	11	9	11	11	11	5	3	11	6
	22	9	2	11	5	11	5	4	4	1	4
	23	9	6	7	7	2	11	11	9	11	8
	24	10	11	6	11	9	11	7	7	6	11

Table 4

Examples of Stimuli Presented in Survey: Surveys 1 and 2

Survey 1				Survey 2			
1	background speech	13	Kerala vocal music	1	cellphone vibrating	13	Nova Scotia non-vocal music
2	dog barking	14	Nias vocal music	2	baby crying	14	water splashing
3	finger tapping	15	Haya vocal music	3	Hehe vocal music	15	Paiwan vocal music
4	coin in vending machine	16	coin dropping	4	Rade non-vocal music	16	Mbendjele vocal music
5	Savo vocal music	17	crying	5	telephone dialing	17	chimes in the wind
6	dog drinking	18	Alur non-vocal music	6	wings flapping	18	Michoacan vocal music
7	Central Polish Folk vocal music	19	Asturias vocal music	7	girl speaking	19	Gargano vocal music
8	cutting with scissors	20	Trinidad vocal music	8	Fon non-vocal music	20	Lucania non-vocal music
9	Amhara vocal music	21	Thao-Ngan non-vocal music	9	Spanish Basques vocal music	21	Kerala vocal music
10	Spanish speech	22	Ibiza non-vocal music	10	dishes clanking	22	puppy whining
11	Croatian Istria vocal music	23	Portuguese Goa non-vocal music	11	French speech	23	door knocking
12	crumpling paper	24	wind	12	Siassi-Umboi vocal music	24	Shetlands vocal music

Participants

A total of 103 participants (ages 10–80; M = 40.0) were recruited through Prolific (prolific.org). Because Prolific does not support recruitment by geographic region, participants were sampled from three countries, namely, Brazil, China, and the United States, each representing a major population center within a distinct world region. Prolific was instructed to approximate an equal distribution across these countries. The final sample included 31 participants from Brazil, 22 from China, and 49 from the United States.

When asked, “How good do you think your listening skills are? (This includes things like remembering melodies, hearing out of tune notes, or hearing a beat that is out of sync with the music.)” participant answers ranged from 1 or 2 (7% of respondents) to the maximum rating of 5 (17%). All but one participant reported normal hearing. No exclusion criteria were applied prior to enrollment.

Procedure

The survey was administered over two consecutive days in early January 2026. Participants were compensated at a rate of $15.00 per hour, based on an estimated completion time of six minutes.

Data Preparation

Prolific automatically rejected implausibly fast submissions during administration. The remaining 144 survey responses were exported from LimeSurvey in CSV format and processed in Excel prior to analysis. Ten responses were removed because the associated participants did not provide a valid Prolific ID, and an additional 30 responses were excluded because the participants had already taken part in the study. One further response was removed because the session timed out and the survey was returned incomplete. After applying these criteria, the final analytic dataset consisted of 103 completed surveys, yielding a total of 4,944 individual judgments.

All stimulus-level ratings were screened for missing values; none were present because the survey required a response on every trial. Stimulus-level metadata, including vocal versus non-vocal status, source dataset, and assigned sound category, were merged with participant-level responses to produce a single analysis-ready dataset. No transformations, normalizations, or outlier adjustments were applied to the rating data.

Statistical Analysis

All statistical analysis was carried out using R. For each of the two response questions, sample proportions ($\hat{p}$) were calculated across all participants and separately within each sound category and country. Inter-participant agreement was assessed using Krippendorff's α, computed both at the aggregate level and within the same subgroupings.

In addition to the descriptive analyses, exploratory inferential models were fit to assess whether sound category or participant-reported country of residence systematically predicted participants' responses. For the binary "music/not-music" judgments, a mixed-effects logistic regression was fit with Category and Country entered as fixed effects and with random intercepts specified for participants and stimuli. For the 1–7 musicality ratings, a linear mixed-effects model with the same fixed- and

random-effects structure was fit. All models were implemented in R using the *lme4* and *lmerTest* packages, and p-values for fixed effects were obtained using Satterthwaite's approximation. These inferential analyses were conducted as a complement to the descriptive results, allowing the extent of category- and country-related variation to be quantified while accounting for variability across participants and stimuli.

Data and Code Availability

All behavioral data and analysis code required to reproduce the inferential models are available in the PsyArXiv project repository for this study: https://osf.io/7gqe9/.

Results

Descriptive Analyses

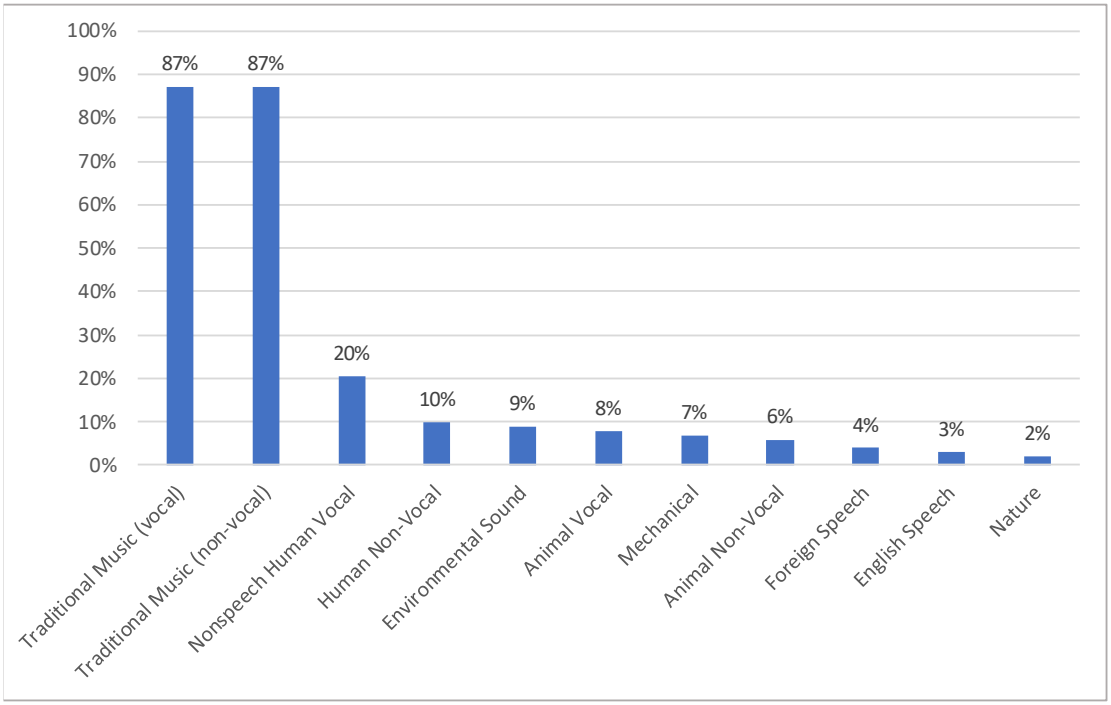
Stimulus Category Distribution

Across all participants and stimuli, the mean proportion of "music" classifications was $\hat{p} = 0.51$, indicating that the likelihood that stimuli would be classified as music on the binary classification task was practically indistinguishable from chance.

On the other hand, classification rates varied substantially across the eleven sound categories (Figure 2). Non-vocal and vocal traditional music received the highest proportions of "music" judgments ($\hat{p}$s equal to 0.87 in both cases), while non-musical sound categories such as English speech, foreign speech and nature sounds received the lowest ($\hat{p}$s ranging from 0.02 to 0.04). In the middle, the borderline category of nonspeech human vocalizations (e.g., crying, sighing) fell between these extremes ($\hat{p}$ equal to 0.20).

Figure 2

Proportion of Respondents Classifying Examples of Each Sound Category as Music (All Countries), p̂

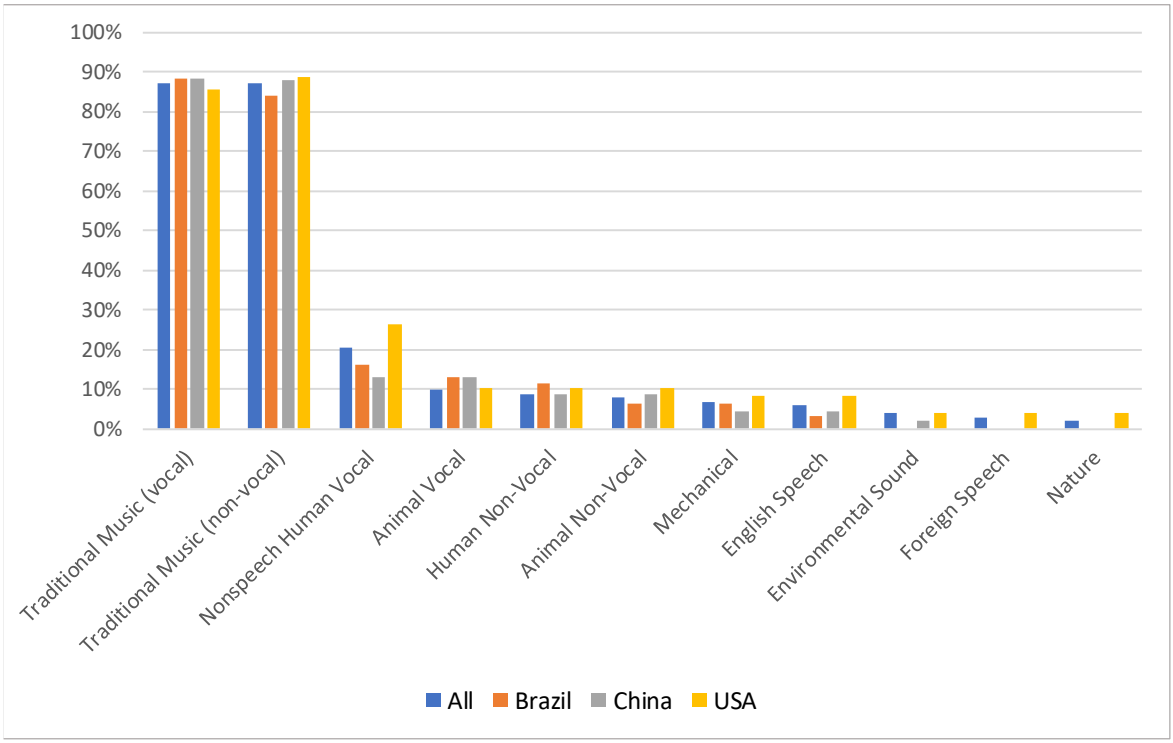


Country-Level Patterns

Similar results were observed at the country of residence level, where classification patterns were broadly similar across participants from Brazil, China, and the United States, and these patterns mimicked the all-countries results (Figure 3). Mean proportions of "music" judgments were $\hat{p}_{\text{Brazil}} = 0.50$, $\hat{p}_{\text{China}} = 0.50$, and $\hat{p}_{\text{United States}} = 0.51$. Although minor differences emerged for specific categories (most notably slightly higher endorsement of vocal music among Brazilian participants), no category exhibited a divergence large enough to suggest systematic cross-country differences.

Figure 3

Proportion of Respondents Classifying Examples of Each Sound Category as Music (by Country), p̂



Musicality Ratings

Participants' Likert-type ratings of "how musical" each sound was showed a pattern consistent with the binary judgments (Table 4). The two traditional musical categories received the highest mean ratings (Ms = 4.5 – 5.1) and non-musical categories the lowest (Ms = 1.2 – 1.3). Ratings and binary judgments were strongly aligned: stimuli classified as "music" received higher musicality ratings on average than those classified as "not music" (mean difference = 3 .2 points). This difference was effectively the same (M = 3.1 to 3.2) for each country of residence as it was for all countries overall.

Table 4

Mean musicality ratings for each sound category (by country), with corresponding Likert scale. "How musical do you consider this sound to be?"

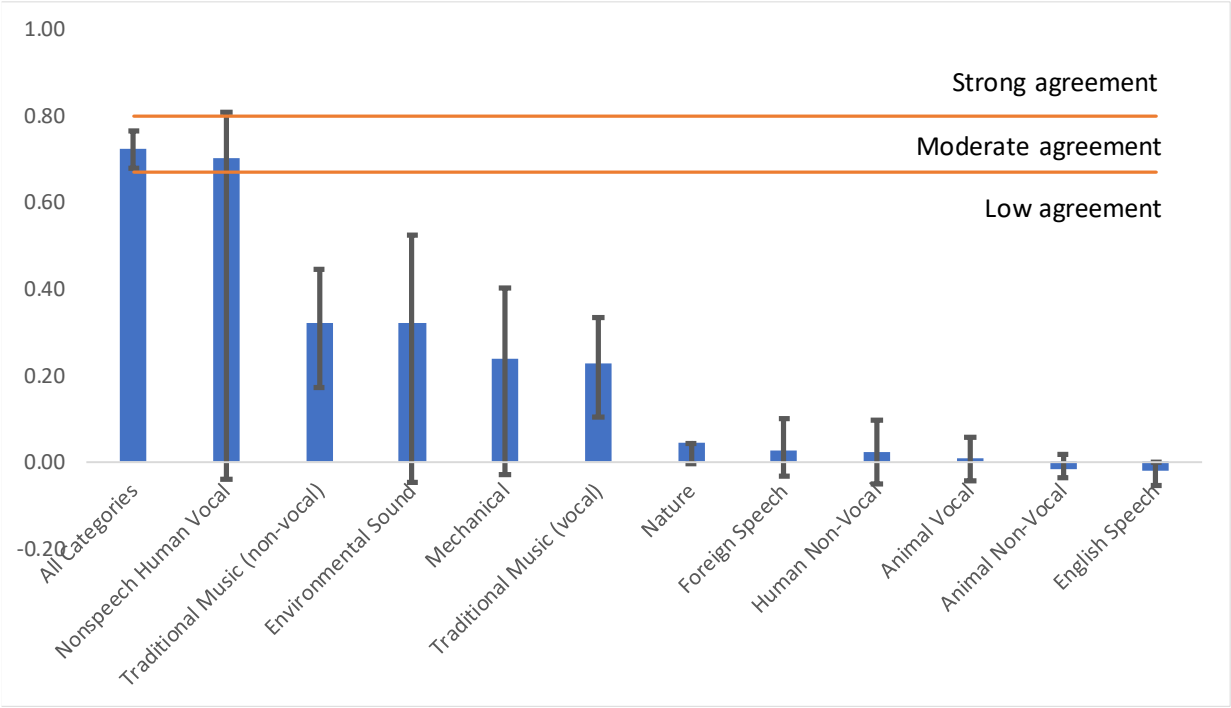
Likert Scale	Sound Category	All	Brazil	USA	China
7 Completely	Traditional Music (non-vocal)	5.1	5.0	5.1	5.2
6 Highly	Traditional Music (vocal)	4.5	4.4	4.5	4.7
5 Very	Nonspeech Human Vocal	1.9	1.5	2.3	1.7
4 Moderately	Human Non-Vocal	1.6	1.4	1.5	2.1
3 Somewhat	Environmental Sound	1.5	1.4	1.5	1.8
2 Slightly	Animal Vocal	1.6	1.4	1.6	2.0
1 Not at all	Mechanical	1.6	1.4	1.6	1.8
	Animal Non-Vocal	1.6	1.5	1.4	2.1
	Foreign Speech	1.2	1.0	1.3	1.4
	English Speech	1.3	1.1	1.2	1.7
	Nature	1.3	1.2	1.3	1.6
	Traditional music categories	4.7	4.6	4.7	4.9
	Other categories	1.5	1.3	1.5	1.8
	Difference	3.2	3.2	3.1	3.1

Consensus Analysis

Inter-participant agreement, quantified using Krippendorff's α, was moderate at the aggregate level for the binary classification task ($\alpha = 0.73$, 95% CI $[0.68, 0.77]$). All confidence intervals were generated using 1000 bootstrap resamples. Agreement varied across categories, with the highest point estimate observed for nonspeech human vocal sounds ($\alpha = 0.70$, 95% CI $[-0.04, 0.81]$) and the lowest for English speech ($\alpha = -0.02$, 95% CI $[-0.05, 0.00]$). Agreement for nearly all categories was low (αs = –0.02–0.32), and many estimates were accompanied by wide confidence intervals that spanned interpretive ranges, indicating substantial uncertainty. Only three categories yielded agreement estimates with reasonably narrow confidence intervals that supported clear interpretation: the aggregate "all categories" estimate ($\alpha = 0.73$, 95% CI $[0.68, 0.77]$), "traditional music (vocal)" ($\alpha = 0.23$, 95% CI $[0.10, 0.33]$), and "traditional music (non-vocal)" ($\alpha = 0.32$, 95% CI $[0.17, 0.45]$). These values indicate moderate agreement for the full set of stimuli and low but measurable agreement for the two traditional-music categories.

Figure 5

Inter-participant agreement of musicality judgments by sound category, measured by Krippendorff's α



Note. 95% confidence intervals computed using 1000 bootstrap resamples.

Inferential Analyses

Binary "Music/Not-Music" Judgments

Participants classified each sound as either "music" or "not music." As a complement to the descriptive analyses, we fit a mixed-effects logistic regression with fixed effects of Category and Country and random intercepts for participants and stimuli. This model assessed whether classification behavior varied systematically across sound categories or across the three countries. Table 5 presents the results.

Sound category strongly predicted the likelihood of classifying a stimulus as music. Relative to the reference category (Animal Non-Vocal), both Traditional music categories were overwhelmingly more likely to be judged as music: Traditional Music (non-vocal), $b = 7.34$, $SE = 1.10$, $z = 6.67$, $p < .001$, and Traditional Music (vocal), $b = 7.13$, $SE = 1.06$, $z = 6.74$, $p < .001$. These coefficients correspond to odds ratios of approximately 1,546 and 1,243, respectively. No other category differed significantly from the reference (all $p > .26$), consistent with the descriptive pattern in which ambiguous or borderline categories elicited highly variable judgments.

Country of residence did not significantly predict classification behavior. Using Brazil as the reference category, participants residing in China (b=0.14, SE=0.43, z=0.32, p=.75) and the United States (b=0.21, SE=0.35, z=0.60, p=.55) showed no meaningful differences in the odds of classifying a sound as music. Random-effects estimates indicated substantial variability across both participants (SD=1.26) and stimuli (SD=1.85), reflecting individual differences and heterogeneity among the sound excerpts.

Overall, the inferential results reinforce the descriptive pattern: listeners strongly agreed that prototypical musical sounds were "music," but judgments for all other categories were inconsistent, with no reliable differences across countries.

Table 5

Fixed Effects From the Logistic Mixed-Effects Model Predicting "Music" Judgments

Predictor	Estimate (b)	SE	z	p	Odds Ratio	95% CI (OR)
Intercept	−4.07	1.04	−3.92		0.017	[0.002, 0.131]
Animal Vocal	0.06	1.26	0.05	0.959	1.07	[0.090, 12.67]
English Speech	−1.30	1.37	−0.95	0.343	0.27	[0.018, 4.01]
Environmental Sound	−0.15	1.15	−0.13	0.898	0.86	[0.091, 8.19]
Foreign Speech	−0.77	1.4	−0.55	0.582	0.46	[0.030, 7.19]
Traditional Music (non-vocal)	7.34	1.1	6.67		1546.28	[178.57, 13,389.68]
Traditional Music (vocal)	7.13	1.06	6.74		1242.83	[156.57, 9,865.53]
Human Non-Vocal	0.47	1.24	0.38	0.706	1.6	[0.140, 18.31]
Human Vocal	1.16	1.25	0.93	0.352	3.19	[0.278, 36.69]
Mechanical	−0.50	1.16	−0.43	0.668	0.61	[0.062, 5.92]
Nature	−1.91	1.7	−1.12	0.261	0.15	[0.005, 4.13]
China	0.14	0.43	0.32	0.751	1.14	[0.50, 2.64]
United States	0.21	0.35	0.6	0.55	1.23	[0.62, 2.46]

Note. Reference levels: Category = Animal Non-Vocal; Country = Brazil. Odds ratios and confidence intervals are exponentiated estimates.

Musicality Ratings

Participants also rated each sound's musicality on a 1–7 scale. To examine whether these ratings varied across categories or countries, we fit a linear mixed-effects model with fixed effects of Category and Country and random intercepts for participants and stimuli. Table 6 presents the results.

Sound category strongly predicted musicality ratings. Using Animal Non-Vocal as the reference category, both traditional music categories received substantially higher ratings: Traditional Music (Non-Vocal), $b = 3.49$, $SE = 0.45$, $t(189.5) = 7.79$, $p < .001$, and Traditional Music (Vocal), $b = 2.90$, $SE = 0.43$, $t(188.2) = 6.68$, $p < .001$. No other category differed significantly from the reference (all $p > .48$), mirroring the descriptive pattern in which non-musical and ambiguous categories clustered at relatively low musicality.

Country effects were modest. Relative to Brazil, China showed slightly higher overall ratings ($b = 0.38$, $SE = 0.17$, $t(97.9) = 2.21$, $p = .030$), whereas the United States did not differ reliably ($b = 0.14$, $SE = 0.14$, $t(96.8) = 0.99$, $p = .32$). Random-effects estimates indicated meaningful variability across both participants ($SD = 0.58$) and stimuli ($SD = 0.91$), as well as residual variability ($SD = 1.04$).

Together, the inferential analyses converge with the descriptive results: listeners across countries consistently rated prototypical musical excerpts as highly musical, while ratings for all other categories were lower and more variable.

Table 6

Fixed Effects From the Linear Mixed-Effects Model Predicting Musicality Ratings

Predictor	Estimate (b)	SE	df	t	p
Intercept	1.45	0.43	211.5	3.33	0.001
Animal Vocal	0.04	0.52	194.5	0.07	0.944
English Speech	−0.33	0.52	194.5	−0.63	0.529
Environmental Sound	−0.09	0.47	191.5	−0.19	0.849
Foreign Speech	−0.38	0.56	192.4	−0.69	0.491
Traditional Music (non-vocal)	3.49	0.45	189.5	7.79	
Traditional Music (vocal)	2.9	0.43	188.2	6.68	
Human Non-Vocal	0	0.52	194.5	0	0.999
Human Vocal	0.33	0.52	194.5	0.63	0.527
Mechanical	−0.02	0.47	191.5	−0.04	0.971
Nature	−0.25	0.63	185	−0.40	0.687
China	0.38	0.17	97.9	2.21	0.03
United States	0.14	0.14	96.8	0.99	0.324

Note. Reference levels: Category = Animal Non-Vocal; Country = Brazil. Ratings were on a 1–7 scale.

Convergence of Descriptive and Inferential Findings

Across both analytic approaches, the same core pattern emerged. The descriptive statistics showed that listeners consistently identified the two traditional music categories as “music” at very high rates and rated them as highly musical, with strong agreement reflected in elevated $\hat{p}$values and higher Krippendorff’s $\alpha$for these categories. In contrast, all other sound categories elicited low and highly variable judgments, with agreement values near chance and substantial within-category dispersion. The inferential models converged with this pattern: both the logistic regression for binary judgments and the linear mixed-effects model for musicality ratings revealed large, reliable effects of sound category, driven almost entirely by the two prototypical musical categories. No other category differed significantly from the reference category in either model, and country effects were small or nonsignificant throughout. Together, these results indicate that listeners across countries strongly agree

on what counts as prototypical music, while judgments for all other sound types remain inconsistent and heterogeneous.

Discussion

The present study examined how listeners from three countries classified a diverse set of sounds as "music" and rated their perceived musicality. Descriptively, participants judged only about half of all stimuli as music, and responses varied widely across sound categories. Traditional vocal and non-vocal music received the highest proportions of "music" classifications and the highest musicality ratings, whereas speech, environmental sounds, and other non-musical categories were rarely judged to be music. These patterns suggest that listeners share broad intuitions about the most prototypical musical sounds, even when the excerpts are culturally unfamiliar.

At the same time, inter-participant agreement was generally low. Although the aggregate reliability estimate indicated moderate agreement, most category-level estimates were accompanied by wide confidence intervals, reflecting substantial uncertainty in how consistently participants applied the concept of "music." Only the two traditional-music categories yielded agreement estimates with sufficient precision to support clear interpretation, and even these values indicated only modest consensus. The low and unstable agreement observed for many categories suggests that judgments about musicality are highly variable across individuals, even within the same cultural group.

The inferential analyses reinforce and clarify these descriptive patterns. The mixed-effects logistic regression revealed that sound category was the dominant predictor of "music/not-music" judgments: both traditional music categories were overwhelmingly more likely to be classified as music than the reference category, whereas no other category differed significantly. Similarly, the linear mixed-effects model showed that these same categories received substantially higher musicality ratings than all others. In both models, country effects were small or nonsignificant, indicating that the large

individual-level variability observed in the descriptive analyses was not strongly patterned by national grouping. The random-effects estimates further underscored the substantial heterogeneity across both participants and stimuli.

Taken together, the descriptive and inferential results converge on a consistent conclusion: listeners across countries share strong intuitions about prototypical musical sounds, but judgments become highly heterogeneous for ambiguous or borderline cases. This pattern suggests that the conceptual boundary of music may be fuzzy rather than categorical. Instead of a single, universally applied criterion, listeners may rely on a constellation of cues—some acoustic, some cultural, some experiential—that vary in salience across individuals.

Several limitations should be noted. First, the use of short, isolated audio excerpts may have reduced the contextual information that often guides everyday musical interpretation. Second, the low reliability estimates for many categories limit the strength of conclusions that can be drawn about those stimuli; additional data or alternative methodological approaches may be needed to estimate agreement with greater precision. Finally, although the sample included participants from three countries, it did not capture the full range of cultural diversity relevant to global musical perception.

Despite these limitations, the findings highlight both the shared and idiosyncratic aspects of how listeners classify sounds as music or non-music. Participants showed broad consensus for traditional musical sounds but substantial variability for ambiguous cases, and this variability was not strongly patterned by country. These results underscore the complexity of the concept of music and point toward the value of future work examining how people answer the question: *What is music?*

References

Bones, O., Cox, T. J., Davies, W. J. (2018). Sound Categories: Category Formation and Evidence-Based Taxonomies. *Frontiers in Psychology 9*(1277). https://doi.org/10.3389/fpsyg.2018.01277

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., Glowacki, L. (2019). Universality and diversity in human song. *Science, 366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*(6), 1281-1296.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674-681. https://doi.org/10.1038/nn1082

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Wood, A. L. C., Kirby, K. R., Ember, C. R., Silbert, S., Passmore, S., Daikoku, H., McBride, J., Paulay, F., Flory, M. J., Szinger, J., D'Arcangelo, G., Bradley, K. J., Guarino, M., Atayeva, M., Rifkin, J., Barron, V., El Hajli, M., Szinger, M., Savage, P. E. (2022). The Global Jukebox: A public database of performing arts and culture. *PLOS ONE*, 17(11), e0275469. https://doi.org/10.1371/journal.pone.0275469