Published in final edited form as:

Curr Biol. 2018 February 05; 28(3): 356–368.e5. doi:10.1016/j.cub.2017.12.042.

Form and function in human song

Samuel A. Mehr[1,2,3,7,8,*], **Manvir Singh**[4,8,*], **Hunter York**[4], **Luke Glowacki**[5,6], and **Max M. Krasnow**[1]

[1]Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, Massachusetts 02138, United States of America [2]Data Science Initiative, Harvard University, 1350 Massachusetts Ave., Cambridge, Massachusetts 02138, United States of America [3]School of Psychology, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand [4]Department of Human Evolutionary Biology, Harvard University, Peabody Museum, 11 Divinity Ave., Cambridge, Massachusetts 02138, United States of America [5]Institute for Advanced Study in Toulouse, 21 Allée de Brienne, 31015 Toulouse, France [6]Department of Anthropology, Pennsylvania State University, 410 Carpenter Building, University Park, Pennsylvania 16802, United States of America

Summary

Humans use music for a wide variety of social functions: we sing to accompany dance, to soothe babies, to heal illness, to communicate love, and so on. Across animal taxa, vocalization forms are shaped by their functions, including in humans. Here we show that vocal music exhibits recurrent, distinct, and cross-culturally robust form-function relations detectable by listeners across the globe. In Experiment 1, internet users (N = 750) in 60 countries listened to brief excerpts of songs, rating each song's function on six dimensions (e.g., *used to soothe a baby*). Excerpts were drawn from a geographically-stratified pseudorandom sample of dance songs, lullabies, healing songs, and love songs recorded in 86 mostly small-scale societies, including hunter-gatherers, pastoralists, and subsistence farmers. Experiment 1 and its analysis plan were pre-registered. Despite participants' unfamiliarity with the societies represented, the random sampling of each excerpt, their very short duration (14 s), and the enormous diversity of this music, the ratings demonstrated accurate and cross-culturally reliable inferences about song functions on the basis of

[*]Correspondence: sam@wjh.harvard.edu and manvirsingh@fas.harvard.edu.
[7]Lead contact
[8]These authors contributed equally

Author Contributions

S.A.M., M.S., and L.G. conceived of the research. S.A.M., M.S., and M.M.K. created the experiments, designed their implementation, planned analyses, and wrote the pre-registration. S.A.M. and M.M.K. managed participant recruitment. H.W.Y. designed and ran the pilot study under the supervision of M.M.K. and S.A.M. S.A.M. conducted data analyses. S.A.M., L.G., and M.M.K. designed the survey of academics, M.M.K. and S.A.M. implemented it, and S.A.M. conducted data analyses. S.A.M., M.S., and M.M.K. wrote the paper and all authors edited it. The field recordings were used with permission from the *Natural History of Song* project, which is directed by S.A.M., M.S., and L.G.

Declaration of Interests

The authors declare no competing interests.

song forms alone. In Experiment 2, internet users (N = 1000) in the United States and India rated three "contextual" features (e.g., *gender of singer*) and seven "musical" features (e.g., *melodic complexity*) of each excerpt. The songs' contextual features were predictive of Experiment 1 function ratings, but musical features and the songs' actual functions explained more variability in function ratings. These findings are consistent with the existence of universal links between form and function in vocal music.

eTOC

Mehr et al. show form-function associations in vocal music detectable by listeners worldwide. People in 60 countries heard songs from 86 societies. They inferred the functions of dance, lullaby, and healing from song forms alone. Ratings were near-identical across listener cohorts and were guided by the contextual and musical features of the songs.

Introduction

Research from across the biological sciences demonstrates that the features of auditory signals and other communicative behaviors are shaped by their intended outcomes [1–3]. For instance, as a general principle, low-frequency, harsh vocal forms with nonlinearities are expected to function in signaling hostility, because those features are correlated with increases in body size and larger animals tend to defeat smaller animals in conflicts [1,4]. This form-function relation is found in many vertebrates, e.g., in the cricket frog [5], river bullhead [6], sparrow hawk [7], and red deer [8] and it is salient enough that naïve listeners identify arousal levels from vocalizations in mammals, amphibians, and reptiles [9].

Similar form-function relations are present in the hostile vocalizations of humans [10,11] and in other domains of human vocal communication. Across 24 societies, the sounds of co-laughter between friends and strangers are distinguishable by acoustic features of the voice associated with arousal [12]; relationships exist between sound and meaning in the word-forms of thousands of human languages [13]; and intention categories in both infant- and adult-directed speech are identifiable from their vocal forms alone [14].

Music has been predicted to show form-function relationships in the contexts of dance [15,16], infant care [17], and ceremonial healing [18]: music used for each of these social functions is expected to show regularities in its form across cultures. In the field of music theory, "form" typically refers to the organization of composed music (e.g., the exposition, development, and recapitulation of "sonata form"). This is not what we mean by "form". Here and throughout, we use "form" in a similar fashion to prior work concerning form and function in vocalization; that is, to refer to the acoustical properties of the vocalization. In vocal music, such forms include "contextual" features (e.g., *gender of singer*) and "musical" features (e.g., *melodic complexity*).

In the domain of emotion, listeners can accurately detect extra-musical information from music played in isolation. For instance, Canadians accurately detect intended emotions of joy, sadness, or anger in Hindustani ragas, despite being unfamiliar with the genre [19]. Similar effects are found with other music and with listeners from other societies [20,21],

including in one non-industrialized society, the Mafa of Cameroon [22] (for review, see [23]). Emotion recognition in music could help to inform form-function inferences about music, but it is unknown whether such inferences exist and, if they do, whether they extend across the music of all cultures.

Studies of a collection of lullabies and love songs [24,25] provide some evidence for regularities in infant-directed songs across cultures. However, the songs therein were selected in part on the basis of their acoustic features, were only sampled from two categories of a much wider musical repertoire, and were not sampled systematically across cultures, which undermines any general inferences of universality in the forms of infant-directed songs. The last issue is common among cross-cultural studies of music, which tend either to study a small number of cultures or use otherwise unrepresentative samples. For instance, a study examining cross-cultural regularities in music [26] used the *Garland Encyclopedia of Music*, which samples irregularly across geographic regions, ethnolinguistic histories, and, crucially, the many social contexts in which music is found. Infant-directed songs constitute less than 5% of the music studied, despite infant-directed music being a common and likely universal form of musical expression [17]. Uneven sampling has the clear potential to bias general inferences from cross-cultural datasets. In the case of [26], the under-sampling of infant-directed songs skews any estimate of gender bias in music away from female singers.

While researchers have proposed a number of potential universals in music and musical behavior [27–29], many of which pertain directly to the possibility of links between form and function in music, testing them requires representative samples of music that span geographic, linguistic, and cultural dimensions, along with the many social contexts in which music appears. Here, we present experiments that do so. We report the results of two experiments using the newly-created *Natural History of Song* Discography. We test for the existence of form-function links in the vocal music of 86 human cultures (Experiment 1) and investigate the forms of this music to explore the mechanisms by which listeners may infer form from function (Experiment 2).

Results

Views from the academy

Historically, the idea that there might be universals in music from many cultures has been met with considerable skepticism, especially among music scholars. This is unsurprising given the leeriness of human universals common across academic disciplines (see [30] for discussion), but the shaky state of evidence for universals in music and the inferential issues described above may in fact justify this skepticism.

Because intellectual trends on controversial topics can change rapidly, we quantified current views on the issue by surveying 940 academics at all career stages who self-reported affiliations in ethnomusicology (n = 206), music theory (n = 148), other areas of music scholarship (n = 299), and psychological and cognitive sciences (n = 302; in total, 15 scholars indicated multiple affiliations). The sample included respondents born in 56 different countries. We asked respondents to consider an imaginary experiment wherein

people listened to examples of vocal music from all cultures to ever exist and to predict two outcomes: (1) whether or not people would accurately identify the social function of each piece of music on the basis of its form alone, and (2) whether peoples' ratings would be consistent with one another (the full text of the questions are in STAR Methods and the dataset openly available at https://osf.io/xpbq2).

The responses differed strikingly across academic fields. Among academics who self-identified as cognitive scientists, 72.9% predicted that listeners would make accurate form-function inferences and 73.2% predicted that those inferences would be mutually consistent. In contrast, only 28.8% of ethnomusicologists predicted accurate form-function inferences and 27.8% predicted mutually consistent ratings. Music theorists were more equivocal (50.7% and 52.0%), as were academics in other music disciplines (e.g., composition, music performance, music technology; 59.2% and 52.8%). When restricting the sample to tenure-track, tenured, and retired academics (n = 539), the results were comparable, with a gap of over 50 percentage points between cognitive scientists and ethnomusicologists on both measures. In sum, there is substantial disagreement among scholars about the possibility of a form-function link in human song.

Experiment 1

We used the *Natural History of Song* Discography to conduct a real version of the imaginary experiment we presented to survey respondents. This collection includes vocal music drawn pseudo-randomly from 86 predominantly small-scale societies, including hunter-gatherers, pastoralists, and subsistence farmers and spanning all 30 world regions defined by the Probability Sample Files of the Human Relations Area Files [31,32] (see Figure 1A and Table 1). Over 75 languages are represented. The discography was assembled by sampling four recordings from each region, with each recording representing a specific social function: dance, healing, love, or lullaby (see Figure 1A for details on the selection criteria). These four functions were chosen because they are known to be found across many cultures [26–29,33,34] and are relevant to the biological and cultural evolution of music [15,17,18,35]. Recordings were selected on the basis of ethnographic information alone: the only auditory criterion for inclusion was that the recording included audible singing, circumventing researcher biases concerning the prototypical musical features of song forms. As such, the *Natural History of Song* discography is a representative sample of human music, the analyses of which can help to answer questions about universality.

If music exhibits universal form-function associations, then (1) listeners who are unfamiliar with a given culture's music should nonetheless accurately identify the functions of songs from that culture based on their forms alone; and (2) listeners should demonstrate comparable form-function inferences regardless of their cultural background. We pre-registered the form-function hypothesis (see https://osf.io/xpbq2) and tested it here, in Experiment 1. We presented the 118 song excerpts to 750 internet users in 60 countries (see Figure 1B and Figure S1). To ensure that listeners could hear the songs, we required them to pass a headphone screening task [36]; we also included a variety of manipulation checks designed to remove inattentive participants (see STAR Methods). Participants listened to a random sample of 36 song excerpts, yielding an average of 225 independent listens (*SD* =

13.9, range: 175–254) for each of the 118 songs (26,580 in total). The broad range of cultures and languages represented in the *Natural History of Song* Discography, combined with the broad range of countries of origin of the participants, makes it likely that participants were both unfamiliar with the music they heard and unable to understand the lyrics.

After each excerpt, participants answered 6 questions indicating their perceptions of the function of each song: on 6-point scales, the degree to which they believed that each song was used (1) *for dancing;* (2) *to soothe a baby;* (3) *to heal illness;* (4) *to express love for another person;* (5) *to mourn the dead;* and (6) *to tell a story*. In total, participants provided 159,480 ratings (26,580 total listens × 6 ratings/song). The first four questions correspond to *actual* functions of the songs, while the last two do not: they were included as foils, to dissuade listeners from an assumption that only four song types were actually present, which could have influenced their responses toward the study's hypothesis. However, because storytelling and mourning are common functions of music in small-scale societies worldwide [33,34], we also analyzed responses on these dimensions; the songs in the *Natural History of Song* discography are not explicitly used for storytelling or mourning, but they may nevertheless share features in reliable patterns with songs that are. A demonstration experiment is at https://harvard.az1.qualtrics.com/jfe/form/SV_e8M5XpwzWS7A0Nn and all data and song excerpts are at https://osf.io/xpbq2.

The analysis strategy had two parts. First, we tested the accuracy of listeners' function inferences via no-constant multiple regressions of the average rating on each of the six questions, with binary predictors for each of the four song functions. We compared perceived song functions to actual song functions via post-hoc general linear hypothesis tests of two types: (1) comparisons of perceived function across known song functions (e.g., "are lullabies rated higher on '…to soothe a baby' than dance songs are?"), and (2) comparisons of each song form to the base rate for a perceived function across all songs (e.g., "are lullabies rated lower on '…for dancing' than the average song is?"). The latter analysis is informative in both positive and negative directions: response patterns reveal listeners' intuitions both for whether a song form has a given function and whether it does *not*. For all analyses, we report results both in raw units (a song type's average rating from "Definitely not used…" [1] to "Definitely used…" [6]) and in standardized units (*z*-scores). Full reporting is in Tables 2 and 3.

Second, to investigate the uniformity of form-function inferences across participants, we split our sample into three cohorts ($N = 250$ each: United States, India, and a "World" cohort of 58 other countries with relatively low Human Development Index scores; see STAR Methods and Figure S1) and examined the degree of cohort-wise agreement for each function rating. For each question, we ran three multiple regressions, each predicting one cohort's average ratings for each song from those of the other two cohorts; we report the best-fitting regression. Listeners' *perceptions* of song functions were in reliable agreement with the songs' *actual* functions. When listening to dance songs, participants rated them as used "for dancing" higher than they did for any other song type (Figure 2A), with the mean difference (M_{diff}) in raw scores ranging from 1.09 to 2.18 (on a 6-point scale). These effects correspond to *z*-scores of 0.85 to 1.70 (Table 2). Dance songs were also rated substantially

higher than the base rate of "used for dancing" across all songs (M_{diff} = 1.16, 95% CI = [0.79, 1.53], F(1,114) = 39.1, p = 7.23×10^{-9}, z-score = 0.91), while lullabies were rated substantially lower than the base rate (M_{diff} = −1.01, 95% CI = [−1.38, −0.65], F(1,114) = 29.7, p = 2.98×10^{-7}, z-score = −0.80). Moreover, these ratings were reliable across listeners: listeners' ratings of "…for dancing" were tightly related to one another between the USA, India, and World cohorts (Figure 2B; F(2,115) = 1877.5, p = 4.67×10^{-90}, R^2 = .970). Listeners thus intuited that dance songs are the most "for dancing" of all song forms, whereas lullabies are *not* for dancing. Moreover, despite their near-complete unfamiliarity with the music they heard, listeners at opposite ends of the world shared intuitions for the musical forms of dance songs.

These effects are large. The raw difference in ratings between lullabies and dance songs (M_{diff} = 2.18) covers more than one third of the entire scale available. The same comparison in units of standard deviation (z-score = 1.70) is roughly the size of the average difference in height between men and women worldwide [37] and over three times the size of typical effects in psychology [38].

In results of similar sizes and patterns, listeners rated lullabies as used "to soothe a baby" higher than any other song type (Figure 2C and Table 2). Their ratings were far higher than the base rate across all songs (M_{diff} = 1.03, 95% CI = [0.76, 1.30], F(1,114) = 57.0, p = 1.16×10^{-11}, z-score = 1.07). Further, dance and healing excerpts were rated lower than the base rate, indicating that listeners felt dance and healing songs are *not* for soothing babies (dance songs: M_{diff} = −0.50, 95% CI = [−0.77, −0.23], F(1,114) = 13.7, p = .0003, z-score = −0.52; healing songs: M_{diff} = −0.39, 95% CI = [−0.67, −0.11], F(1,114) = 7.69, p = .006, z-score = −0.41). As with dance songs, listeners' ratings of "…to soothe a baby" were nearly identical across cohorts (Figure 2D; F(2,115) = 2188.2, p = 7.70×10^{-94}, R^2 = .974). Thus, lullabies found worldwide share enough features to elicit large and distinctive profiles of function ratings from naïve listeners. These results confirm predictions from a theoretical account of infant-directed music [17].

Inferences about healing songs showed similar patterns, though listeners were less confident in their appraisals, as indicated by smaller effect sizes (Figure 2E). They rated healing songs significantly above the base rate of the dimension "to heal illness" (M_{diff} = 0.26, 95% CI = [0.07, 0.45], F(1,114) = 7.21, p = .008, z-score = 0.49) and significantly higher than dance songs and love songs, with a nonsignificant difference from lullabies (Table 2). Only dance songs were rated significantly below the base rate (M_{diff} = −0.20, 95% CI = [−0.39, −0.02], F(1,114) = 4.69, p = .032, z-score = −0.38). Listeners around the world shared notions of which songs were used "to heal illness", although cohort-wise agreement was lower than for dance songs or lullabies (Figure 2F; F(2,115) = 352.3, p = 1.27×10^{-50}, R^2 = .860).

Further, listener ratings exhibited a modest relation between healing songs and the foil dimension "to mourn the dead" (Figure 3A), with healing songs rated significantly higher than the base rate (M_{diff} = 0.36, 95% CI = [0.07, 0.64], F(1,114) = 6.27, p = .014, z-score = 0.46). Healing songs were also rated higher than dance songs and marginally higher than lullabies and love songs (Table 3). Dance songs were rated significantly lower than the base rate (M_{diff} = −0.38, 95% CI = [−0.65, −0.11], F(1,114) = 7.57, p = .007, z-score = 0.48). The

ratings also exhibited high cohort-wise agreement, though lower than the non-foil dimensions (Figure 3B; $F(2,115) = 620.4$, $p = 2.08 \times 10^{-63}$, $R^2 = .915$). Thus, not only are cross-cultural regularities in the forms of healing song detectable by listeners from industrialized societies, but these listeners share conceptualizations of what constitutes a healing song, despite the fact that healing songs are rare in many developed nations [18].

Listeners' form-function inferences about love songs were the weakest of the four song types (Figure 2G). In contrast to the other three song types, love songs were not rated significantly higher than the base rate ($M_{diff} = 0.15$, 95% CI = [−.04, 0.35], $F(1,114) = 2.45$, $p = .120$, z-score = 0.27) and only healing songs were rated significantly below it ($M_{diff} = −0.31$, 95% CI = [−0.51, −0.11], $F(1,114) = 9.60$, $p = .002$, z-score = −0.56). Listeners rated love songs as used "to express love to another person" higher than healing songs only ($M_{diff} = 0.46$, 95% CI = [0.19, 0.74], $F(1,114) = 11.0$, $p = .001$, z-score = 0.83), but not the other two song types (Table 2). Listeners did, however, make reliable assessments in their ratings of love songs across cohorts (Figure 2H; $F(2,115) = 283.6$, $p = 5.85 \times 10^{-46}$, $R^2 = .831$). They also judged love songs to be higher-than-average on the foil dimension "to tell a story" (Figure 3C; $M_{diff} = 0.19$, 95% CI = [0.04, 0.35], $F(1,114) = 6.18$, $p = .014$, z-score = 0.43), higher than both healing songs and lullabies, but not dance songs (Table 3). Ratings for "to tell a story" were highly similar across study populations (Figure 3D; $F(2,115) = 235.2$, $p = 4.52 \times 10^{-42}$, $R^2 = .804$). Listeners thus do make some form-function inferences about love songs, but they are not nearly as clear as those of the other song types we studied.

To investigate the variability of these findings across the geographic regions from which songs were recorded, we took advantage of the geographic stratification used in the construction of the *Natural History of Song* discography. Songs in the discography were gathered by obtaining one example of each of the four song types across 30 geographic regions (see STAR Methods), which enables a simple test of the geographic variability of the form-function inferences described above. For each of the three high-accuracy form-function inferences (i.e., dance songs used "for dancing", lullabies used "to soothe a baby", and healing songs used "to heal illness") we took the region-wise average function rating across each region and counted the number of regions in which the target song type had a higher-than-average function rating.

The results show near-uniformity of form-function inferences for dance songs and lullabies across the geographic regions from which songs were sampled, with weaker results for healing songs. In 27 of 30 world regions (90.0%), dance songs were rated higher as "for dancing" than the other three song types; in 29 of 30 regions (96.7%), lullabies were rated higher as "to soothe a baby" than the other three song types; and in 20 of 28 regions (71.4%; n.b., the *Natural History of Song* discography lacks healing songs from two regions) were healing songs rated higher as "to heal illness" than the other three song types. Thus, not only are listeners' form-function inferences accurate and reliable, but they show a strong degree of uniformity across the cultures studied (especially for dance songs and lullabies).

In sum, three common types of songs found worldwide — dance songs, lullabies, and healing songs — appear to elicit accurate and reliable form-function inferences from a

diverse body of listeners. These findings are consistent with the existence of universal form-function links in human song.

Experiment 2

What features of song forms enable naïve listeners to accurately and reliably identify song functions? In Experiment 2, we conducted an exploratory investigation of the features listeners used to discriminate song functions, focusing on general traits of the recordings that are detectable by naïve listeners. We presented the same 118 excerpts from Experiment 1 to 1000 internet users in India (n = 500) and the United States (n = 500). No listeners participated in both experiments. As in Experiment 1, we required listeners to pass a headphone screening task and filtered out inattentive participants with a series of manipulation checks (see STAR Methods). Each participant listened to 18 song excerpts, yielding an average of 149 independent listens (*SD* = 11.3, range: 123–176) per song (17,527 in total).

For each excerpt, participants answered a random set of 5 questions drawn from a set of 10. Three corresponded with participants' ratings of contextual aspects of the performance: (1) *number of singers*; (2) *gender of singer(s)*; and (3) *number of instruments*. Seven corresponded with subjective musical features of the song: (1) *melodic complexity*; (2) *rhythmic complexity*; (3) *tempo*; (4) *steady beat;* (5) *arousal*; (6) *valence*; and (7) *pleasantness*. Listeners provided a total of 87,142 ratings (17,527 total listens × 5 ratings/song – 493 listener/song/feature combinations where no answer was provided) and split-half reliability of the items was acceptable (rs = .81–.99; see STAR Methods for more information along with the full text of the 10 items).

To assess whether and how the contextual and musical features of song forms predicted listeners' ratings of social function, we conducted 3 sets of exploratory analyses. First, we examined the degree of variation on each of the 10 features across each of the song forms and tested whether or not song forms differed on those features. Second, we summarized the musical features via a principal components analysis. Third, we examined the influence of the songs' contextual features and musical features on listeners' function ratings with a series of regressions. Given the high degree of subjectivity of the ratings, the very brief excerpts, and the complete lack of context provided to the listeners, we consider these analyses to be exploratory and not exhaustive: they are intended to help explain the findings of Experiment 1, not to provide a comprehensive feature analysis of *Natural History of Song* recordings.

The four song types showed clear differences in both contextual and musical features (Figure S2). Unsurprisingly, the forms of dance songs and lullabies differed most from other song types, both for contextual and musical features (full reporting is in Table S1). Relative to the other three song types, listeners rated dance songs as having more singers (z-score = 0.86), more instruments (z-score = 0.76), higher melodic complexity (z-score = 0.79), higher rhythmic complexity (z-score = 0.87), faster tempo (z-score = 1.09), steadier beat (z-score = 0.84), higher arousal (z-score = 1.17), higher valence (z-score = 1.09), and higher pleasantness (z-score = 0.72). Effects for lullabies were comparably large, but in the opposite direction: relative to the other song types, lullabies were rated as having fewer

singers (z-score = −0.76), fewer instruments (z-score = −0.92), lower melodic complexity (z-score = −1.12), lower rhythmic complexity (z-score = −1.06), slower tempo (z-score = −1.04), less steady beat (z-score = −0.63), lower arousal (z-score = −0.90), lower valence (z-score = −0.74), and lower pleasantness (z-score = −0.45). Lullabies were also rated substantially more likely than the other song types to have a female singer (z-score = 0.93). As in Experiment 1, results with healing songs and love songs were mostly inconclusive (see Table S1). In sum, listeners heard substantial differences between the forms of lullabies and dance songs, with modest results for healing and love songs.

Because the 7 musical ratings were highly correlated with one another (Table S2) we conducted a principal components analysis to summarize them. This yielded two components with eigenvalues greater than 1, explaining 88.1% of item variance. We report unrotated components here and throughout. Component 1 correlated moderately and positively with all 7 features, while Component 2 correlated negatively with melodic and rhythmic complexity, positively with pleasantness and steady beat, and did not correlate with valence or arousal (full reporting is in Table S3).

Because listeners in Experiment 1 did not provide mutually exclusive ratings for song function, as they did in previous work (e.g., [24], where listeners rated songs as either “lullaby” or “love song”), listener “errors” in ratings can be captured here on continuous scales. To explore cases where different song types were highly rated on the same function (e.g., a healing song and a dance song both rated highly — and erroneously — as “to soothe a baby”), we plotted each song’s function rating against its location in principal components space. This analysis, visualized in Figure 4, demonstrates the relation between the strength of each song’s function rating (from Experiment 1) and a two-dimensional summary of each song’s form (from Experiment 2).

There were two main results. First, songs of different types overlapped substantially in principal components space. Second, incorrect ratings occur non-randomly: songs rated erroneously high on a given function tend to share similar forms with songs that *do* have that function. This pattern is evident for all song types, including those with accurate, reliable form-function inferences: while lullabies and dance songs were clearly distinguished from one another in Experiment 1, in principal components space, some lullabies appear alongside dance songs and are rated correspondingly high on the dimension “for dancing”. The converse is also true.

Last, we examined the extent to which the feature ratings in Experiment 2 explained the form-function inferences in Experiment 1. If function inferences are determined by contextual features alone, the findings of Experiment 1 may simply reflect broad patterns in how music is used across cultures — e.g., “lullabies usually have only one singer, who is usually female” — rather than supporting the hypothesis that song forms themselves inform listeners’ function inferences. To test this question, we built four series of regression models (one series per function rating). Within each series, we examined the degree to which their variance was explained by the contextual feature ratings alone (Model 1), the principal-components reduction of musical feature ratings alone (Model 2), both sets of features

(Model 3), and both sets with an indicator variable for the target song type (Model 4; full reporting is in Tables S4–S7).

Relative to models predicting perceived song function from contextual features alone, the inclusion of the two principal components and the target song form as covariates substantially increased model fit. A model with only the contextual features predicted 74.6% of variance in the function rating "for dancing" (Table S4; $F(3,114) = 112$, $p = 8.07 \times 10^{-34}$), whereas the inclusion of the principal components and an indicator variable for dance songs increased explanatory power by 14.8 percentage points ($R^2 = .895$; nested test: $F(3,111) = 52.0$, $p = 4.59 \times 10^{-21}$). Even with these covariates, the indicator for dance songs explained unique variance (partial $R^2 = .0846$, $p = .002$). For lullabies (Table S5), a model with contextual features, principal components, and an indicator variable for lullabies explained 9.7 percentage points more variance in the function rating "to soothe a baby" ($R^2 = .683$) than did a model with only contextual features ($R^2 = .586$), a significant difference (nested test: $F(3,111) = 11.3$, $p = 1.55 \times 10^{-6}$). As with dance songs, the indicator for lullabies explained unique variance (partial $R^2 = .094$, $p = .0009$). Similar results were present in healing songs (Table S6) and love songs (Table S7).

In sum, the form-function inferences that listeners made in Experiment 1 cannot be explained solely by contextual features of music. While they are indeed predictive of listeners' function ratings, subjectively-rated musical features of each song tended to explain more unique variability in function ratings than did contextual ratings. Moreover, neither contextual nor musical features fully explained function ratings: an identifier covariate in models for all four song types explained unique variance in function ratings. Function detection in song is thus facilitated by both contextual *and* musical features of song forms — and by other features reliably present in songs that were not measured in Experiment 2.

Discussion

The present research provides evidence for the existence of recurrent, perceptible features of three domains of vocal music across 86 human societies, and the striking consistency of form-function percepts across listeners from around the globe — listeners who presumably know little or nothing about the music of indigenous peoples. Moreover, these studies suggest that song types differ from each other on the basis of both contextual and musical features, but musical features tend to be more predictive of form-function inferences than contextual features.

Why do songs that share social functions have convergent forms? If dance songs are shaped by adaptations for signaling coalition quality [15], their contextual and musical features should amplify that signal. The feature ratings in Experiment 2 support this idea: dance songs tend to have more singers, more instruments, more complex melodies, and more complex rhythms than other forms of music. If lullabies are shaped by adaptations for signaling parental attention to infants [17], their acoustic features should amplify that signal. The feature ratings in Experiment 2 support this idea: lullabies tend to be rhythmically and melodically simpler, slower, sung by one female person, and with low arousal relative to other forms of music.

This work raises two key questions about the basic facts of music. First, despite the geographic variation in listeners in Experiment 1, all participants were English-literate and had access to an expansive variety of music on the Internet. They thus share a great deal of musical experience. Do form-function inferences generalize to all listeners worldwide, even those who have no shared musical experience, or who know only the music of their own culture? A stronger test of universality would require testing the inferences of people living in isolated societies with minimal access to the music of other cultures.

Second, while we used naïve listeners' perceptions of musical forms to explore what drove form-function inferences, those perceptions are subjective, were based on brief excerpts of the songs rather than full performances, and lack rich contextual information available from ethnomusicologists and anthropologists. Are the musical and contextual features of the songs that inform function inferences universal? A stronger demonstration of universals in music would require in-depth feature analyses of a cross-culturally representative sample of music from small-scale societies, informed by expert listeners, music information retrieval, and modern approaches from data science.

Nevertheless, the present research demonstrates that cross-cultural regularities in human behavior pattern music into recurrent, recognizable forms, while maintaining its profound and beautiful variability across cultures.

STAR Methods

Contact for resource sharing

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Samuel Mehr (sam@wjh.harvard.edu).

Experimental model and subject details

Survey of academics—940 academics (390 female, 439 male, 3 other, 108 did not disclose; age 20–91 years, mean = 46.7, SD = 14.5) born in 56 countries were recruited in two fashions: first, by emailing all affiliates publicly listed in the Music and Psychology/Cognitive Science departments at the top 200 universities listed for each department in the U.S. News & World Report Best Colleges; and second, by distributing the survey anonymously to three music listservs (Society for Ethnomusicology, Society for Music Theory, and AUDITORY). No participants were excluded from analyses. Participants were given the opportunity to enter into a drawing for 50 gift cards of $25 value and could opt out of any/all questions on the survey. All participants agreed to a consent statement before the study, which was approved by Harvard University's Committee on the Use of Human Subjects. All procedures were in accordance with approved guidelines.

Experiment 1—750 participants (USA: n = 250, 115 female, age 18–65 years, mean = 35.6, SD = 10.6; India: n = 250, 60 female, age 19–65 years, mean = 30.3, SD = 6.96; World: n = 250, 80 female, age 18–65 years, mean = 29.8, SD = 7.52) were recruited through Amazon Mechanical Turk (MTurk), an online labor marketplace. The majority of MTurk workers are located in the USA and India, so we aimed to recruit cohorts of workers in the USA, in India, and in a "World" cohort of MTurk workers who were not residents of

the US, India, or 28 Western nations with high Human Development Index scores [39]; we defined "Western nations" following a classic work in international relations [40]. The full listing of countries present in the World cohort is in Figure S1. Using MTurk's interface, we made the study available to English-speaking participants who had at least a 95% successful completion rate for prior MTurk tasks. All participants were paid between $1 and $3 upon completion and agreed to a consent statement before the study, which was approved by Harvard University's Committee on the Use of Human Subjects. All procedures were in accordance with approved guidelines.

Experiment 2—1000 participants (USA: n = 500, 277 female, age 20–71 years, mean = 37.1, SD = 11.4; India: n = 500, 136 female, age 18–81 years, mean = 30.2, SD = 7.64) were recruited through MTurk. The study was available to English-speaking participants who did not participate in Experiment 1 and who had at least a 75% successful completion rate for prior MTurk tasks. All participants were paid $2 upon completion and agreed to a consent statement before the study, which was approved by Harvard University's Committee on the Use of Human Subjects. All procedures were in accordance with approved guidelines.

Method details

Survey of academics—Participants first indicated their primary and secondary fields of study, career stage, expertise in music performance, and degree of familiarity with music from small-scale societies. They then answered the two key questions described below, followed by a number of other questions about universals in music and other behaviors, human evolution, and the scientific study of music which are not relevant to the present report. The two items that participants completed are reproduced in full below:

a. Here is a thought experiment. Imagine that you are a researcher with unlimited time and resources, and have access to a fantastic time machine that can put you anywhere in the world at any time.

Imagine that you use your time machine and your unlimited time and resources to obtain a recording of every single song that has ever been sung by every person in the world (everyone from people in big cities to people in isolated hunter-gatherer groups). For each song, you also find out what the people do while listening to or while singing the song; e.g., that people dance along to it, use the song to calm down a fussy infant, etc.

Then, you run a simple experiment. You take these many recordings and play each one for many people around the world (from people in big cities to people in isolated hunter-gatherer groups).

After they listen to the recording, you ask each of these people to think about the singer, and to say what behaviors they think the singer was using the song with (e.g., "used to soothe a baby", "used for dancing", "used for healing illness", "used for expressing love to another person"). They have only heard the recording and don't know the answer: they will be guessing the behaviors on the basis of how the song sounds and nothing else.

There are a range of possible outcomes. It might be that people can can guess what a song is used for just by hearing it, without any prior experience or knowledge about the song's cultural context. On the other hand, it might be that music around the world and over time is so variable that listeners would have trouble guessing what a song is used for just by hearing it.

What do you think the results of this imaginary experiment would be? Response options were *On average, people would be very bad at accurately guessing the behaviors; On average, people would be kind of bad at accurately guessing the behaviors; On average, people would be kind of good at accurately guessing the behaviors; On average, people would be very good at accurately guessing the behaviors*; and *I prefer not to answer*.

b. Whether or not people are good at guessing what a song is used for, people from different cultural backgrounds might interpret music in similar ways, or in different ways. In the same imaginary experiment, imagine that you measured how consistent the people's answers were with one another. What do you think the result would be? Response options were *The guesses from people all over the world would be very inconsistent with one another; The guesses from people all over the world would be kind of inconsistent with one another; The guesses from people all over the world would be kind of consistent with one another; The guesses from people all over the world would be very consistent with one another*; and *I prefer not to answer*.

For both questions, we did not analyze data from subjects who responded *I prefer not to answer*. Responses on both questions were coded as binary variables, that is, grouping together the lower two and upper two responses to both questions.

Experiment 1

Participant exclusions: To ensure the quality of the data reported, we only analyzed the responses of participants who successfully passed a series of compliance and attention checks. First, all participants were required to wear headphones: on the MTurk website we stated that this was a requirement for participation and we used a headphone screening task to ensure participants' compliance with this requirement (see Headphone screening; those participants who failed the screening task were not allowed to continue with the study and thus are not included in the summary statistics above). Second, we used geolocation to confirm the countries in which participants were located, in addition to filtering by their MTurk registration country (n.b., this method does not protect against participants who mask their true location, e.g., by using a proxy server). Third, we excluded participants who self-reported problems hearing more than 10% of the excerpts (i.e., more than 4 playback failures) to minimize variance in the number of excerpts rated across participants. Last, we excluded participants on the basis of several attention and compliance checks (see Supplemental Methods). To obtain the final N of 750, we ran 903 participants who passed the headphone check and excluded 52 for reporting more than 4 playback failures, 40 for geolocation outside of targeted countries, and 61 for failing one or more manipulation checks.

Headphone screening: This task used the method of [36]. On each of six trials, participants heard three tones and were asked to indicate which was the quietest/softest. One of the three tones on each trial was set at −6dB relative to the other two and one of the two louder tones was in antiphase between the two stereo channels. The three tones in a trial were presented in a random order. Free-field listeners (e.g., on laptop speakers) hear the antiphase tone as softer than it actually is, due to phase cancellation, and thus are likely to answer incorrectly that the antiphase tone is quietest. In contrast, listeners wearing headphones are unaffected by the antiphase manipulation and are likely to answer correctly that the −6dB tone is the quietest. The task thus distinguishes between participants who are wearing headphones and those who are not. For full details of the task, see [36]; per the task's design, participants scoring 5 or 6 correct (out of 6 trials) were included in the full study.

Experimental protocol: First, to demonstrate the structure of the study, we played a recording of the song "Happy Birthday" and asked participants to report a simple inference about the song's function: "Think of the singer(s). I think that the singers…" with response options on a 1 to 6 scale from "Definitely do not use the song to celebrate a birthday" to "Definitely use the song to celebrate a birthday". Participants who responded on the negative side of the scale were asked to replay the track and respond again. Then, the full study began. There were 36 trials, each containing an excerpt randomly drawn from the *Natural History of Song* discography (see Collection of recordings). The interface only allowed participants to play the excerpt once, did not allow participants to advance to the next page until the excerpt ended, and did not allow participants to return to the playback page after it played. Participants could report a technical issue in hearing the excerpt (i.e., answering "Yes" to "Did you have any trouble hearing that song?", in which case they advanced to the next excerpt without answering any questions). We then asked the six function questions in a random order. Each was presented in the same fashion: "Think of the singer(s). I think that the singers…" with response options of 6 radio buttons, with the left anchor labeled "Definitely do not use the song {X}" to "Definitely use the song {X}", where {X} was one of the six functional dimensions: "for dancing", "to soothe a baby", "to heal illness", "to express love for another person", "to tell a story", and "to mourn the dead". For each question, participants clicked a radio button and were immediately advanced to the next item. After completing all 36 trials, they completed a set of compliance and attention checks (see below) before returning to MTurk to receive payment.

Compliance and attention checks: We asked five questions of participants to assess their compliance with instructions and their attention to the task:

- **a.** "What color is the sky? Please answer this incorrectly, on purpose, by choosing RED instead of blue." Response options were *Green, Red, Blue,* or *Yellow*. Any participant who did not answer *Red* was excluded.
- **b.** "Did you wear headphones while listening to the sounds in this HIT? Please answer honestly. Your payment does NOT depend on your response to this question." Response options were *Yes* or *No*. Any participant who answered *No* was excluded.

c. "Turkers are working on this HIT in many different places. Please tell us about the place where you worked on this HIT. Please answer honestly. Your payment does NOT depend on your response to this question." Response options were *I worked on this HIT in a very noisy place, I worked on this HIT in a somewhat noisy place, I worked on this HIT in a somewhat quiet place,* or *I worked on this HIT in a very quiet place*. Any participant who answered *I worked on this HIT in a very noisy place* or *I worked on this HIT in a somewhat noisy place* was excluded.

d. "Turkers are working on this HIT with many different devices, browsers, and internet connections. Please tell us about whether you had difficulty loading the sounds. Please answer honestly. Your payment does NOT depend on your response to this question." Response options were *There were problems loading all of the sounds, There were problems loading most of the sounds, There were problems loading some of the sounds,* or *There were no problems loading any of the sounds*. Any participant who answered *There were problems loading all of the sounds* or *There were problems loading most of the sounds* was excluded.

e. "How carefully did you complete this survey? Please answer honestly. Your payment does NOT depend on your response to this question." Response options were *Not at all carefully, Slightly carefully, Moderately carefully, Quite carefully,* or *Very carefully*. Any participant who answered *Not at all carefully, Slightly carefully,* or *Moderately carefully* was excluded.

Note that items (b), (c), and (d) were not used in the USA cohort.

Collection of recordings: We used music from the *Natural History of Song* Discography, wherein researchers searched published collections and contacted anthropologists and ethnomusicologists to find recordings from each of 30 world regions defined by the Probability Sample Files of the Human Relations Area Files [31,32]. From the available recordings in each area, searches were limited to those that included audible singing, and were chosen as examples of each of four song types based on the below criteria by consulting their ethnographic descriptions. Preference was always given to recordings with the richest ethnographic description and to the 60 societies included in the Probability Sample Files; when more than one recording fit these criteria, the final selection was made at random. Songs were selected so as to best fit the criteria listed in Figure 1 (see also the excluded examples in Figure 1, for comparison). Judgments of each recording's goodness-of-fit to these criteria were made independently of the judgment of whether or not there was audible singing, to ensure that inclusion criteria were unbiased by the researchers' personal interpretations of the music present on the recording.

Stimuli: We randomly selected 14 sec excerpts of each track in the *Natural History of Song* Discography. If the randomly sampled period happened to contain predominantly non-sung content (e.g., an instrumental interlude) or included non-musical auditory cues that indicated the behavioral context (e.g., a baby crying during a lullaby), we rejected the excerpt and randomly selected a new one from the same recording. A similar procedure was used in the

pilot study (see below), but to ensure that pilot findings were not unique to those particular excerpts, we re-sampled all excerpts for the present study.

Pilot study: Before conducting the experiments reported here, we conducted exploratory pilot experiments in MTurk cohorts in the United States ($N = 99$) and India ($N = 95$), who listened to a variety of *Natural History of Song* recordings. In addition to a variety of questions on the content of each excerpt (e.g., number and gender of singers), we asked participants to identify the song's function in a four-alternative forced choice question. Identification accuracy was above chance for dance songs, lullabies, and healing songs, and several of the perceived features co-varied with song types. These exploratory findings led us to undertake the present work, which added a variety of controls, used more sensitive measures of song function, and sampled listeners from more countries.

Pre-registration of hypotheses and analysis plan: Based on the results of the pilot study, we designed the present research as a conceptual replication targeting the detection of song functions and pre-registered it at https://osf.io/xpbq2. The study and analyses were carried out as per the registration with two minor changes. First, we collected data from 250 participants in the "World" cohort, rather than the planned 500 participants, because we exhausted the available pool of World participants that were readily available on MTurk. However, this sample size is consistent with the rationale in our registration; that is, the World cohort is over 2.5 times the size of the MTurk cohorts in the pilot study. Second, because we found that users in the India and World cohorts reported substantially more difficulty hearing excerpts than the USA cohort, we added manipulation check questions about the environment in which they were listening and about their ability to hear the excerpts, excluding those participants who reported that they were in a noisy environment and/or who had difficulty hearing many tracks (see Participant exclusions).

Experiment 2—The headphone screening task, compliance and attention checks, collection of recordings, and stimuli were identical to those used in Experiment 1.

Experimental protocol: After successful completion of the headphone screening task, participants listened to 18 excerpts, drawn from the same set of *Natural History of Song* discography excerpts in Experiment 1 (see Collection of recordings). After listening to each question, they answered five questions probing their perceptions of song features drawn at random from the full set of 10 items (three contextual and seven musical; see Main text). The full text of each item is reproduced below:

a. "How many singers do you hear?" Response options were *1, 2, 3, 4, 5,* or *More than 5*.

b. "What is the gender of the singer or singers? If you're not sure, please make a guess." Response options were *Male, Female* or *Both*.

c. "How many musical instruments did you hear? Please do not count the singer as a musical instrument (for example, if you heard a singer and a guitar, you would answer "1 instrument"; but if you only heard a solo singer, you would answer

"No instruments")." Response options were *No instruments, 1 instrument, 2 instruments, 3 instruments, 4 instruments,* or *5 or more instruments*.

d. "Think about the melody of this song. By "melody", we mean the pattern of notes, pitches, or tones, that make up the song. You could also call the melody the "tune". How complex is the melody? You may include in your answer a consideration of the melodies played in accompanying instruments, if any were present." Response options were six radio buttons, with the first labeled *Very simple* and the last labeled *Very complex*.

e. "Think about the rhythms of this song. By "rhythms", we mean the timing of the singing and instruments, the pattern of beats in one or more voices or instruments, the regularity or irregularity of the pulses, etc. How complex are the rhythms? You may include in your answer a consideration of the rhythms played in accompanying instruments, if any were present." Response options were six radio buttons, with the first labeled *Very simple* and the last labeled *Very complex*.

f. "How fast is this song?" Response options were six radio buttons, with the first labeled *Very slow* and the last labeled *Very fast*.

g. "How steady is the beat in this song?" Response options were six radio buttons, with the first labeled *Very unsteady beat* and the last labeled *Very steady beat*.

h. "How exciting is this song?" Response options were six radio buttons, with the first labeled *Not exciting at all* and the last labeled *Very exciting*.

i. "How happy is this song?" Response options were six radio buttons, with the first labeled *Very sad* and the last labeled *Very happy*.

j. "How pleasant is this song?" Response options were six radio buttons, with the first labeled *Very unpleasant* and the last labeled *Very pleasant*.

As in Experiment 1, the interface permitted participants to play each excerpt only once, prevented them from advancing until the excerpt ended, prevented listeners from returning to the playback page, and gave them the option to report difficulties hearing the excerpt (in which case they were advanced to the next excerpt without answering any questions). After completing the 18 trials, participants completed the requisite compliance and attention checks before returning to MTurk for their payment.

Participant exclusions: We used the same exclusion criteria as Experiment 1, with one exception: we excluded participants who reported technical difficulties with at least half of the excerpts. To obtain the final N of 1000, we ran 1136 participants who passed the headphone check and excluded 6 for reporting more than 9 playback failures, 44 for geolocation outside of targeted countries, and 86 for failing one or more attention checks.

Item reliability: Because of the nested random assignment of excerpts and items, standard reliability metrics (e.g., alpha) are not appropriate. Instead, we computed split-half reliability for each of the 10 features. For each song, we split the available ratings into two sets, took their song-wise means, and computed a Pearson correlation (n = 118) for the means. Split-

half reliability was acceptable for all items (number of singers: r = .99; gender of singer(s): r = .99; number of instruments: r = .98; melodic complexity: r = .82; rhythmic complexity: r = .82; tempo: r = .95; steady beat: r = .83; arousal: r = .91; valence: r = .93; pleasantness: r = .87).

Quantification and statistical analysis

The pre-registration (see https://osf.io/xpbq2) details many of the methods and analyses reported here and was finalized before the experiment or analyses were conducted. Statistical analyses were conducted in Stata and visualizations were created in R. All statistical details of the experiments, including the statistical tests used, exact values of n, what n represents, definition of center, and dispersion and precision measures can be found in the main text. Significance was defined before the analyses were conducted as an alpha level of .05. We report exact p-values in the main text and in the tables. Details of the sample size estimation and subject exclusion are in Participant exclusions. Standard regression assumptions were checked by visual inspection of the data; no assumptions were violated in any analysis.

Data and software availability

All data are available at https://osf.io/xpbq2.

Additional resources

Song excerpts and interactive versions of the 3D scatterplots in Figures 2 and 3 are available at https://osf.io/xpbq2. A demonstration version of Experiment 1 is also available and can be viewed at https://harvard.az1.qualtrics.com/jfe/form/SV_e8M5XpwzWS7A0Nn.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Harvard University Department of Psychology (grant to M.M.K.), the ANR Labex at the Institute for Advanced Studies Toulouse (fellowship to L.G.), the Harvard Data Science Initiative (fellowship to S.A.M.), and the National Institutes of Health (NIH Director's Early Independence Award DP5OD024566 to S.A.M.). We thank the participants; J. McDermott and K. Woods for sharing their headphone screening task and assisting us with it; R. Howard and L. Lopez for research assistance; G. Bryant, D. Locke, A. Lomax Wood, A. Martin, J. McDermott, J. Nemirow, T. O'Donnell, K. Panchanathan, J. Rekedal, and E. Spelke for comments on the manuscript; G. North and four anonymous reviewers for their constructive feedback; and the members of the Evolutionary Psychology Laboratory at Harvard University for many productive discussions that led to this work.

References

1. Morton ES. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. Am Nat. 1977; 111:855–869.
2. Owren MJ, Rendall D. Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. Evol Anthropol. 2001; 10:58–71.
3. Endler JA. Some general comments on the evolution and design of animal communication systems. Philos Trans R Soc B Biol Sci. 1993; 340:215–225.

4. Fitch WT, Neubauer J, Herzel H. Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. Anim Behav. 2002; 63:407–418.
5. Wagner WE. Fighting, assessment, and frequency alteration in Blanchard's cricket frog. Behav Ecol Sociobiol. 1989; 25:429–436.
6. Ladich F. Sound production by the river bullhead, Cottus gobio L. (Cottidae, Teleostei). J Fish Biol. 1989; 35:531–538.
7. Mueller HC. Displays and vocalizations of the sparrow hawk. Wilson Bull. 1971; 83:249–254.
8. Clutton-Brock TH, Albon SD. The roaring of red deer and the evolution of honest advertisement. Behaviour. 1979; 69:145–170.
9. Filippi P, Congdon JV, Hoang J, Bowling DL, Reber SA, Pašukonis A, Hoeschele M, Ocklenburg S, de Boer B, Sturdy CB, et al. Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. Proc R Soc B. 2017; 284:20170990.
10. Sell A, Bryant GA, Cosmides L, Tooby J, Sznycer D, von Rueden C, Krauss A, Gurven M. Adaptations in humans for assessing physical strength from the voice. Proc R Soc Lond B Biol Sci. 2010; 277:3509–3518.
11. Puts DA, Apicella CL, Cárdenas RA. Masculine voices signal men's threat potential in forager and industrial societies. Proc R Soc Lond B Biol Sci. 2011:rspb20110829.
12. Bryant GA, Fessler DMT, Fusaroli R, Clint E, Aarøe L, Apicella CL, Petersen MB, Bickham ST, Bolyanatz A, Chavez B, et al. Detecting affiliation in colaughter across 24 societies. Proc Natl Acad Sci. 2016; 113:4682–4687. [PubMed: 27071114]
13. Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH. Sound–meaning association biases evidenced across thousands of languages. Proc Natl Acad Sci. 2016; 113:10818–10823. [PubMed: 27621455]
14. Bryant GA, Barrett HC. Recognizing intentions in infant-directed speech: Evidence for universals. Psychol Sci. 2007; 18:746–751. [PubMed: 17680948]
15. Hagen EH, Bryant GA. Music and dance as a coalition signaling system. Hum Nat. 2003; 14:21–51. [PubMed: 26189987]
16. Bryant GA. Animal signals and emotion in music: Coordinating affect across groups. Front Psychol. 2013; 4:990. [PubMed: 24427146]
17. Mehr SA, Krasnow MM. Parent-offspring conflict and the evolution of infant-directed song. Evol Hum Behav. 2017; 38:674–684.
18. Singh M. The cultural evolution of shamanism. Behav Brain Sci. in press.
19. Balkwill LL, Thompson WF. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music Percept. 1999; 17:43–64.
20. Balkwill LL, Thompson WF, Matsunaga R. Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners1. Jpn Psychol Res. 2004; 46:337–349.
21. Meyer RK, Palmer C, Mazo M. Affective and coherence responses to Russian laments. Music Percept. 1998; 16:135–150.
22. Fritz T, Jentschke S, Gosselin N, Sammler D, Peretz I, Turner R, Friederici AD, Koelsch S. Universal recognition of three basic emotions in music. Curr Biol. 2009; 19:573–576. [PubMed: 19303300]
23. Eerola T, Vuoskoski JK. A comparison of the discrete and dimensional models of emotion in music. Psychol Music. 2011; 39:18–49.
24. Trehub SE, Unyk AM, Trainor LJ. Adults identify infant-directed music across cultures. Infant Behav Dev. 1993; 16:193–211.
25. Unyk AM, Trehub SE, Trainor LJ, Schellenberg EG. Lullabies and simplicity: A cross-cultural perspective. Psychol Music. 1992; 20:15–28.
26. Savage PE, Brown S, Sakai E, Currie TE. Statistical universals reveal the structures and functions of human music. Proc Natl Acad Sci. 2015; 112:8987–8992. [PubMed: 26124105]
27. Brown, DE. Human universals. Philadelphia: Temple University Press; 1991.
28. Brown S, Jordania J. Universals in the world's musics. Psychol Music. 2013; 41:229–248.
29. Lomax A. Universals in song. World Music. 1977; 19:117–129.
30. Pinker, S. The blank slate: the modern denial of human nature. New York: Viking; 2002.

31. Naroll R. The proposed HRAF probability sample. Cross-Cult Res. 1967; 2:70–80.
32. Murdock, GP., Ford, CS., Hudson, AE., Kennedy, R., Simmons, LW., Whiting, JWM. Outline of cultural materials. New Haven, CT: Human Relations Area Files, Inc; 2008.
33. Lomax, A. Folk song style and culture. Washington, DC: American Association for the Advancement of Science; 1968.
34. Nettl, B. The study of ethnomusicology: Thirty-three discussions. Urbana, IL: University of Illinois Press; 2015.
35. Miller, GF. The mating mind: How sexual choice shaped the evolution of human nature. New York: Doubleday; 2000.
36. Woods KJP, Siegel MH, Traer J, McDermott JH. Headphone screening to facilitate web-based auditory experiments. Atten Percept Psychophys. 2017:1–9.
37. Gray JP, Wolfe LD. Height and sexual dimorphism of stature among human societies. Am J Phys Anthropol. 1980; 53:441–456. [PubMed: 7468783]
38. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349:aac4716. [PubMed: 26315443]
39. UNDP. Human development for everyone. New York, NY: United Nations Development Programme; 2016.
40. Huntington, SP. The clash of civilizations and the remaking of world order. New York: Simon & Schuster; 1997.

Highlights

1. People in 60 countries listened to songs from 86 mostly small-scale societies
2. They successfully inferred song functions on the basis of song form alone
3. Listener ratings were guided by both contextual and musical features of the songs
4. Human song therefore exhibits widespread form-function associations

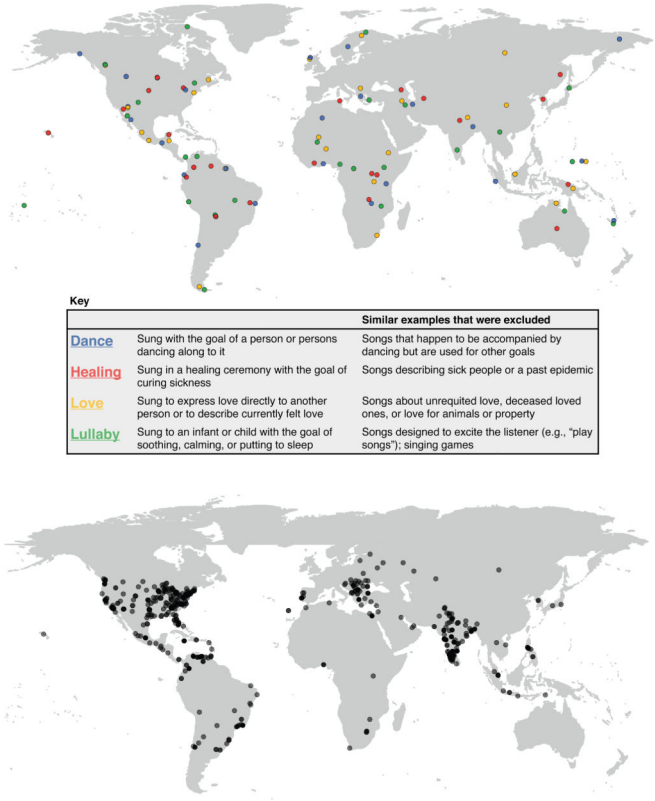

Figure 1. Locations of song recordings and listeners

The 118 recordings were made in 86 societies, the locations of which are plotted in **A,** and sorted by song function (see Legend). Details on the societies and recordings are in Table 1 and in Method details. **B,** Locations of the listeners in Experiment 1 (n = 750), plotted with geolocation data gathered from IP addresses. Each gray dot represents a single listener; darker dots represent multiple listeners in the same region. Details on listener demographics and countries represented are in STAR Methods and Figure S1. See also Figure S1.

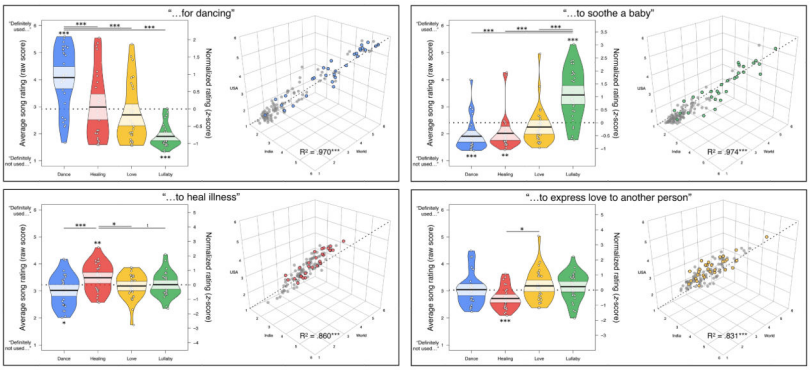

Figure 2. Accuracy and international consistency of form-function ratings

Participants, who were unaware of the functions of songs from which excerpts were drawn, were asked to judge the function of each excerpt on each dimension on a scale from 1 ("Definitely not for…") to 6 ("Definitely for…"). Results are grouped by question, one per box, with the text of each question at the top of each box. The left side of each box presents listeners' *perceived* function of each song plotted as a function of the songs' *actual* functions, in violin plots. The right side of each box presents the degree of agreement in ratings across the three cohorts of listeners. In all plots, each point represents a song's average rating. In the violin plots (left side), song-wise averages are reported both as raw ratings (left y-axis) and as z-scores (right y-axis); the latter included for reference to effect sizes relative to a Normal distribution. The violin plots are kernel density estimations, the black lines are means, and the shaded white areas are the 95% confidence intervals of the means. Dotted lines denote the grand mean on each question, which varies in units of raw ratings, but due to normalization, is always 0 in z-scores. In the 3D scatterplots (right side), the dotted line is the equation $z = y = x$; that is, perfect consistency across cohorts. Please visit https://osf.io/xpbq2 to explore the 3D plots directly; these online versions can be rotated and zoomed interactively. **A,** Listeners rate dance songs higher on "for dancing" than all other song types and higher than the average song. **B,** Across all songs, ratings on "for dancing" are highly consistent across listeners in the USA, India and 58 other countries in the "World" cohort, and dance songs (in blue) form a strikingly distinct group from other songs. **C** and **D,** Comparable patterns are found with lullabies: listeners rate them higher than any other song type as used "to soothe a baby", higher than average, and consistently across cohorts. **E,** Similarly, function inferences for healing songs are distinct from other song types, but with smaller effect sizes, and **F,** listener ratings are consistent across cohorts but with more variation than in **B** and **D**. In **G,** listeners do not accurately rate love songs as used "to express love for another person"; but nevertheless, **H,** ratings on this dimension are consistent across cohorts. Asterisks denote p-values from general linear hypothesis tests (left panels) or multiple regression omnibus tests (right panels). ***p < .001, **p < .01, *p < .05, ^{t}p < .1. See also Tables S4–S7.

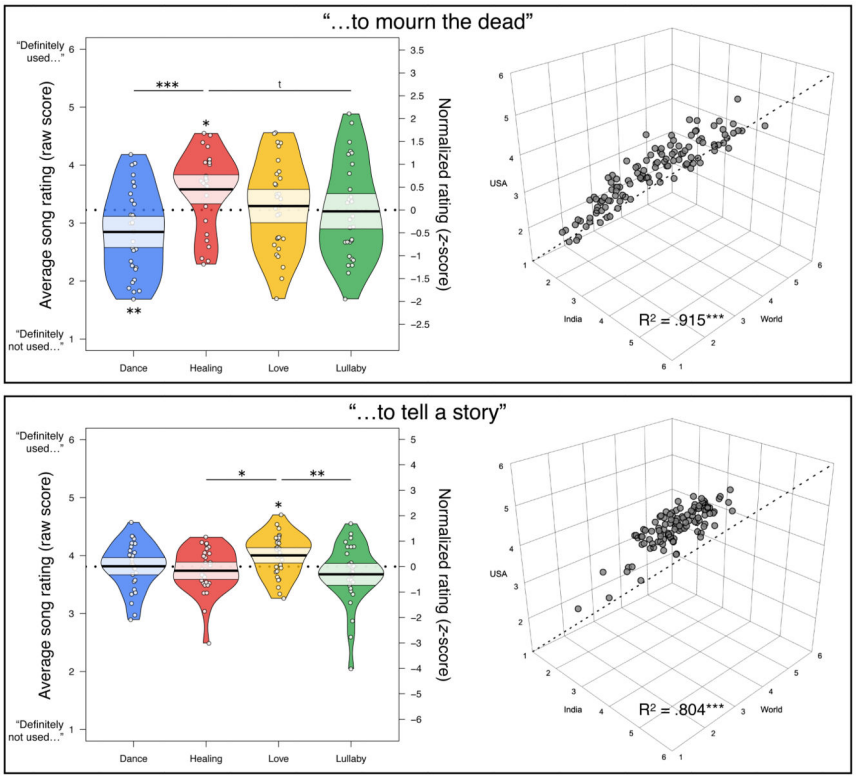

Figure 3. Exploratory findings from foil dimensions

To mask the number of known song functions presented in the study, participants also rated the songs on two dimensions that were not explicitly represented by the songs in corpus. Thus, we had no predictions for responses on these dimensions. However, listener responses demonstrated modest, but consistent differences across song types. **A,** Healing songs are rated higher than average on "to mourn the dead", and higher than both dance songs and lullabies, with **B,** consistent responses across cohorts. **C,** Despite the fact that listeners were unable to detect love songs on the dimension "to express love to another person" (see Figure 2), they *did* rate love songs higher than average on "to tell a story", higher than healing songs and lullabies, and **D,** ratings across cohorts were highly consistent with one another. To explore the 3D plots directly, including rotation and zoom, please visit https://osf.io/xpbq2. Asterisks denote p-values from general linear hypothesis tests (left panels) or multiple regression omnibus tests (right panels). $p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{t}p < .1$.

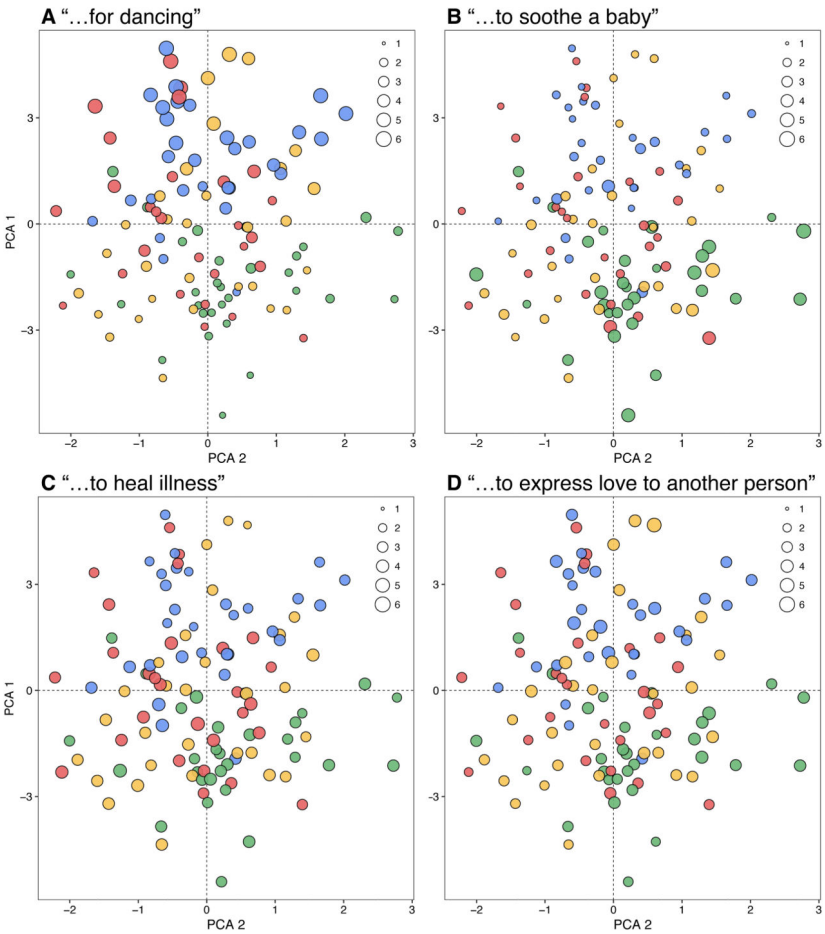

Figure 4. Relations between strength of form-function inferences and musical forms

In the scatterplots (**A–D**), each point shows the location of a song in principal-components space, along with the strength of its form-function inference (i.e., in **A**, the larger the point, the higher the song's rating on "for dancing"). Bubble sizes are unstandardized across plots. As in the previous figures, dance songs are depicted in blue, healing songs in red, love songs in yellow, and lullabies in green. See also Figure S2 and Tables S1–S7.

Table 1
Listing of societies and locations from which recordings were gathered

All data are used with permission from the *Natural History of Song* project and are subject to correction. When multiple song types are indicated for the same society, they correspond to multiple recordings (i.e., not multiple types for the same recording). See also Figure 1.

Society	Subsistence type	Region	Sub-region	Song type(s) used
Ainu	Primarily hunter-gatherers	Asia	East Asia	Dance, Lullaby
Aka	Hunter-gatherers	Africa	Central Africa	Dance, Lullaby
Akan	Horticulturalists	Africa	Western Africa	Healing
Alacaluf	Hunter-gatherers	South America	Southern South America	Love
Amhara	Intensive agriculturalists	Africa	Eastern Africa	Love
Anggor	Horticulturalists	Oceania	Melanesia	Healing
Aymara	Horticulturalists	South America	Central Andes	Dance
Bahia Brazilians	Intensive agriculturalists	South America	Eastern South America	Dance, Healing
Bai	Intensive agriculturalists	Asia	East Asia	Love
Blackfoot	Hunter-gatherers	North America	Plains and Plateau	Dance, Lullaby
Chachi	Horticulturalists	South America	Northwestern South America	Dance
Chewa	Horticulturalists	Africa	Southern Africa	Lullaby
Chukchee	Pastoralists	Asia	North Asia	Dance, Lullaby
Chuuk	Other subsistence combinations	Oceania	Micronesia	Dance, Love
Emberá	Horticulturalists	Middle America and the Caribbean	Central America	Dance
Ewe	Horticulturalists	Africa	Western Africa	Dance
Fulani	Pastoralists	Africa	Western Africa	Love
Fut	Horticulturalists	Africa	Western Africa	Lullaby
Ganda	Intensive agriculturalists	Africa	Eastern Africa	Healing
Garifuna	Horticulturalists	Middle America and the Caribbean	Central America	Love
Garo	Horticulturalists	Asia	South Asia	Dance
Georgia	Intensive agriculturalists	Europe	Southeastern Europe	Healing
Goajiro	Pastoralists	South America	Northwestern South America	Lullaby
Gourara	Agro-pastoralists	Africa	Northern Africa	Dance
Greeks	Intensive agriculturalists	Europe	Southeastern Europe	Dance, Lullaby
Guarani	Other subsistence combinations	South America	Eastern South America	Love, Lullaby
Haida	Hunter-gatherers	North America	Northwest Coast and California	Lullaby
Hawaiians	Intensive agriculturalists	Oceania	Polynesia	Dance, Healing, Love
Highland Scots	Other subsistence combinations	Europe	British Isles	Dance, Love, Lullaby
Hopi	Intensive agriculturalists	North America	Southwest and Basin	Dance, Lullaby
Huichol	Horticulturalists	Middle America and the Caribbean	Northern Mexico	Love
Iglulik Inuit	Hunter-gatherers	North America	Arctic and Subarctic	Lullaby

Society	Subsistence type	Region	Sub-region	Song type(s) used
Iroquois	Horticulturalists	North America	Eastern Woodlands	Dance, Healing, Lullaby
Iwaidja	Hunter-gatherers	Oceania	Australia	Love
Javaé	Horticulturalists	South America	Amazon and Orinoco	Lullaby
Kanaks	Horticulturalists	Oceania	Melanesia	Dance, Lullaby
Kelabit	Horticulturalists	Asia	Southeast Asia	Love
Kogi	Horticulturalists	South America	Northwestern South America	Healing, Love
Korea	Intensive agriculturalists	Asia	East Asia	Healing
Kuna	Horticulturalists	Middle America and the Caribbean	Central America	Healing, Lullaby
Kurds	Pastoralists	Middle East	Middle East	Dance, Love, Lullaby
Kwakwaka'wakw	Hunter-gatherers	North America	Northwest Coast and California	Healing, Love
Lardil	Hunter-gatherers	Oceania	Australia	Lullaby
Lozi	Other subsistence combinations	Africa	Southern Africa	Dance
Lunda	Horticulturalists	Africa	Southern Africa	Healing
Maasai	Pastoralists	Africa	Eastern Africa	Dance
Marathi	Intensive agriculturalists	Asia	South Asia	Lullaby
Mataco	Primarily hunter-gatherers	South America	Southern South America	Dance, Healing
Maya (Yucatan Peninsula)	Horticulturalists	Middle America and the Caribbean	Maya Area	Healing
Mbuti	Hunter-gatherers	Africa	Central Africa	Healing
Melpa	Horticulturalists	Oceania	Melanesia	Love
Mentawaians	Horticulturalists	Asia	Southeast Asia	Dance
Meratus	Horticulturalists	Asia	Southeast Asia	Healing
Mi'kmaq	Hunter-gatherers	North America	Eastern Woodlands	Love
Nahua	Other subsistence combinations	Middle America and the Caribbean	Maya Area	Love, Lullaby
Nanai	Primarily hunter-gatherers	Asia	North Asia	Healing
Navajo	Intensive agriculturalists	North America	Southwest and Basin	Love
Nenets	Pastoralists	Asia	North Asia	Love
Nyangatom	Pastoralists	Africa	Eastern Africa	Lullaby
Ojibwa	Hunter-gatherers	North America	Arctic and Subarctic	Dance, Healing, Love
Ona	Hunter-gatherers	South America	Southern South America	Lullaby
Otavalo Quichua	Horticulturalists	South America	Central Andes	Healing
Pawnee	Primarily hunter-gatherers	North America	Plains and Plateau	Healing, Love
Phunoi	Horticulturalists	Asia	Southeast Asia	Lullaby
Q'ero Quichua	Agro-pastoralists	South America	Central Andes	Love, Lullaby
Quechan	Intensive agriculturalists	North America	Southwest and Basin	Healing
Rwandans	Intensive agriculturalists	Africa	Central Africa	Love
Saami	Pastoralists	Europe	Scandinavia	Love, Lullaby
Samoans	Horticulturalists	Oceania	Polynesia	Lullaby
Saramaka	Other subsistence combinations	South America	Amazon and Orinoco	Dance, Love

Society	Subsistence type	Region	Sub-region	Song type(s) used
Serbs	Intensive agriculturalists	Europe	Southeastern Europe	Love
Seri	Hunter-gatherers	Middle America and the Caribbean	Northern Mexico	Healing, Lullaby
Sweden	Intensive agriculturalists	Europe	Scandinavia	Dance
Thakali	Agro-pastoralists	Asia	South Asia	Love
Tlingit	Hunter-gatherers	North America	Northwest Coast and California	Dance
Tuareg	Agro-pastoralists	Africa	Northern Africa	Love, Lullaby
Tunisians	Intensive agriculturalists	Africa	Northern Africa	Healing
Turkmen	Intensive agriculturalists	Middle East	Middle East	Healing
Tzeltal	Horticulturalists	Middle America and the Caribbean	Maya Area	Dance
Uttar Pradesh	Intensive agriculturalists	Asia	South Asia	Healing
Walbiri	Hunter-gatherers	Oceania	Australia	Healing
Yapese	Horticulturalists	Oceania	Micronesia	Healing, Lullaby
Yaqui	Intensive agriculturalists	Middle America and the Caribbean	Northern Mexico	Dance
Ye’kuana	Horticulturalists	South America	Amazon and Orinoco	Healing
Yolngu	Hunter-gatherers	Oceania	Australia	Dance
Zulu	Horticulturalists	Africa	Southern Africa	Love

Table 2

Main effects

Each section of the table reports general linear hypothesis tests comparing the four main function ratings corresponding to the target song type to the function ratings for the other three song types (e.g., Are dance songs rated higher on the function “for dancing” than lullabies, love songs, or healing songs?). Comparisons for each item are listed in descending order of effect size. Bolded results are statistically significant at alpha = .05. See also Figure 2.

	M_{diff}	95% CI	F(1,114)	p	z-score
Dance songs as used “for dancing”					
vs. lullabies	**2.18**	**[1.66, 2.70]**	**68.5**	**2.74×10^{-13}**	**1.70**
vs. love songs	**1.38**	**[0.86, 1.90]**	**27.6**	**7.11×10^{-7}**	**1.08**
vs. healing songs	**1.09**	**[0.56, 1.62]**	**16.6**	**8.68×10^{-5}**	**0.85**
Lullabies as used “to soothe a baby”					
vs. dance songs	**1.53**	**[1.15, 1.91]**	**63.3**	**1.44×10^{-12}**	**1.60**
vs. healing songs	**1.42**	**[1.03, 1.80]**	**52.4**	**5.59×10^{-11}**	**1.48**
vs. love songs	**1.19**	**[0.81, 1.57]**	**38.0**	**1.08×10^{-8}**	**1.24**
Healing songs as used “to heal illness”					
vs. dance songs	**0.47**	**[0.20, 0.73]**	**11.8**	**.000826**	**0.87**
vs. love songs	**0.31**	**[0.04, 0.58]**	**5.14**	**.0253**	**0.57**
vs. lullabies	0.26	[−0.01, 0.52]	3.58	.0611	0.48
Love songs as used “to express love to another person”					
vs. healing songs	**0.46**	**[0.19, 0.74]**	**11.0**	**.00122**	**0.83**
vs. dance songs	0.14	[−0.13, 0.41]	1.00	.319	0.25
vs. lullabies	0.03	[−0.24, 0.30]	0.04	.839	0.05

Table 3

Exploratory effects

Each section of the table reports general linear hypothesis tests of ratings on the two foil dimensions for two target song types. Comparisons are between a target song type (e.g., healing songs for “to mourn the dead”) and the other three song types and are listed in descending order of effect size. Bolded results are statistically significant at alpha = .05. The statistics in this table correspond to the visualizations in the left-hand panels of Figure 3.

	M_{diff}	95% CI	F(1,114)	p	z-score
Healing songs as used “to mourn the dead”					
vs. dance songs	**0.73**	**[0.34, 1.13]**	**13.8**	**.000320**	**0.93**
vs. lullabies	0.38	[−0.01, 0.77]	3.68	.0576	0.48
vs. love songs	0.29	[−0.10, 0.68]	2.11	.149	0.36
Love songs as used “to tell a story”					
vs. lullabies	**0.33**	**[0.11, 0.54]**	**8.79**	**.00368**	**0.74**
vs. healing songs	**0.26**	**[0.04, 0.49]**	**5.57**	**.0199**	**0.60**
vs. dance songs	0.19	[−0.03, 0.41]	2.91	.0910	0.43