Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity

Elizabeth H. Margulis[a,1], Patrick C. M. Wong[b], Cara Turnbull[a], Benjamin M. Kubit[c], and J. Devin McAuley[d,1]

[a]Department of Music, Princeton University, Princeton, NJ 08544; [b]Department of Linguistics and Modern Languages, Chinese University of Hong Kong, Hong Kong SAR, China; [c]Department of Psychology, Princeton University, Princeton, NJ 08544; and [d]Department of Psychology, Michigan State University, East Lansing, MI 48824

Edited by John Hajda, Psychological and Brain Sciences, University of California, Santa Barbara, CA; received June 4, 2021; accepted December 13, 2021, by Editorial Board Member Michael S. Gazzaniga

The scientific literature sometimes considers music an abstract stimulus, devoid of explicit meaning, and at other times considers it a universal language. Here, individuals in three geographically distinct locations spanning two cultures performed a highly unconstrained task: they provided free-response descriptions of stories they imagined while listening to instrumental music. Tools from natural language processing revealed that listeners provide highly similar stories to the same musical excerpts when they share an underlying culture, but when they do not, the generated stories show limited overlap. These results paint a more complex picture of music's power: music can generate remarkably similar stories in listeners' minds, but the degree to which these imagined narratives are shared depends on the degree to which culture is shared across listeners. Thus, music is neither an abstract stimulus nor a universal language but has semantic affordances shaped by culture, requiring more sustained attention from psychology.

music | narrative | imagination | culture | semantics

Music listening rarely involves exclusive focus on the sounds themselves. Instead, listeners often experience a kind of "distributed attentional focus" (1) where the sounds are heard but attention rests more on an associated set of multisensory imaginings, including visual imagery, autobiographical memories, or kinesthetic sensations (2). One such mode of listening that has been recently well documented involves hearing instrumental music narratively—i.e., imagining a story as the music unfolds in time (3, 4).

Research on narrative listening has revealed that the degree of a piece of music's narrativity (the likelihood that the musical excerpt triggers a story in listeners' minds) as well as its narrative engagement (how vivid and clear the events and characters of the story are in listeners' minds) varies from excerpt to excerpt, with strong within-culture consistency and large between-cultural differences for responses to individual pieces (3–5). Unknown, however, is whether the content of the imagined stories varies or converges across people and cultures because the previous work from this project has used quantitative measures of narrative perceptions and not focused on or analyzed the semantic content of the narratives generated by listeners. A key question that cannot be addressed without examining the free-response stories themselves is to what extent are these imagined musical narratives shared across listeners? At stake is the degree of intersubjectivity around a fundamental aspect of musical experience and meaning construction. Are we all hearing and imagining the same thing when we listen to a piece of instrumental music, or are our experiences hopelessly subjective?

Some indirect evidence that responses are characterized by substantial intersubjectivity comes from McAuley et al. (5). Here, the authors examined individuals' ability to match narratives to their corresponding musical excerpt—where the "correct" match was a narrative generated by another individual in response to the musical excerpt and where the individual was from the same or a different culture. Results revealed that individuals had remarkable success matching story descriptions to the corresponding excerpt as long as the to-be-matched narrative was generated by an individual from the same culture. It is not clear, however, that the capacity to recognize a narrative that another listener might imagine would extend to a tendency to spontaneously generate a similar narrative when listening. Moreover, due to the binary nature of the matching task, the similarity of imagined narratives between different listeners could not be expressly assessed. The present study directly addresses these questions by examining the semantic similarity of the stories generated by listeners in different geographical locations. Of particular interest are the roles that musical style and more broadly culture may play in constraining the stories that listeners imagine in response to music.

The capacity to imagine stories constitutes one example of a broader human capacity for scene construction, which underlies episodic memory and episodic future thinking as well as the imagination of fictional scenarios (6, 7). In fact, the boundary between remembering something that happened to oneself and imagining something that might happen to someone else is thinner than might be expected—reliving a past series of events and vividly constructing a hypothetical series of events are "correlated highly enough to be considered the same empirical measure" (ref. 6, p. 2). Given this functional equivalence, story

Significance

Are we all imagining the same thing when we listen to music, or are our experiences hopelessly subjective? This research analyzes the similarity of responses from 622 participants in three locations on a highly unconstrained task: free-response descriptions of the stories they imagined while listening to instrumental music. Strikingly, participants in two separate locations that share an overarching culture imagine highly similar narratives to individual excerpts. But these similarity patterns do not extend to narratives imagined by participants in a third location with a distinct culture. This work shows that music—often considered an "abstract stimulus"—can trigger shared stories in listeners' minds but that this intersubjectivity depends on a shared underlying culture.

Author contributions: E.H.M., P.C.M.W., and J.D.M. designed research; E.H.M., P.C.M.W., C.T., and J.D.M. performed research; E.H.M., C.T., B.M.K., and J.D.M. analyzed data; and E.H.M., P.C.M.W., C.T., B.M.K., and J.D.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.H. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: margulis@princeton.edu or dmcauley@msu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2110406119/-/DCSupplemental.

Published January 21, 2022.

imagination tasks can be understood to effectively probe essential cognitive processes.

Imagination does not just share core components with memory, it also constitutes a fundamental part of perception. Perceiving a cat half-hidden behind a tree involves imagining the missing back of the cat (8). In a study from 1910, participants asked to visualize objects while staring at a white wall thought they were imagining the objects but were actually perceiving objects that had been subtly projected on the wall (9). Thus, the boundary between perception and imagination can seem as porous as the one between memory and imagination. This finding is particularly germane to the case of music, where imaginative episodes unfold dynamically as people listen. Just as participants in the Perky experiment imagined objects that followed the contours of barely visible projections, there are indications that the content of the imagined stories triggered by listening to instrumental music relate nonarbitrarily to the acoustic features of the music (10–12). To date, however, to the best of our knowledge, no rigorous large-scale study has focused specifically on the content of stories imagined in response to instrumental music in a manner that affords a consideration of the way both musical excerpts and culture shape shared narrative imaginings.

A related body of work has empirically examined questions around music's semantic dimensions but understood in a more constrained way. Brief clips of music have been shown to elicit the N400 ERP component associated with semantic priming for target words that are semantically related versus semantically unrelated to that clip—for example, the word "narrow" after a clip featuring pitches close together within a small range, or the word "basement" after a clip featuring low pitches; behaviorally, participants were able to mark which of two choices was the related target word (13). Even single chords were able to elicit N400s for semantically incongruent words (e.g., hate for a consonant chord, love for a dissonant ones) but only in participants with formal musical training (14). A functional MRI study using the same paradigm suggested that for musically trained participants, individual chords "may activate affective representations, which spread onto affectively related lexical representations" (ref. 15, p. 93). In other work, single tones with different timbres have been similarly shown to elicit N400s for incongruent words, such as "tense" or "open" (16). All of these studies reveal stable semantic associations to tones, chords, or brief clips within defined subgroups (e.g., musicians in a single geographic location), as assessed by binary-choice designs.

The present study goes beyond the work of Koelsch and colleagues in important ways. Here, we use an unconstrained task consisting of free-response descriptions to study relationships among spontaneously imagined narratives sustained while listening to longer 1-min musical excerpts (rather than brief clips). While previous empirical work on musical semantics provides tantalizing clues about lower-level mechanisms that may relate to the process of formulating full-fledged, unconstrained imaginings, nothing in that previous body of work provides evidence that shared imaginings might exist at the level of full stories to musical excerpts that extend beyond a few seconds under free-response (rather than binary choice) conditions, nor does anything in that previous body of work make use of narrative data collected at multiple geographic sites providing both within- and between-culture comparisons in order to identify the scope of any intersubjectivity.

The basis for the current study was an initial cross-cultural investigation by our team that focused on survey responses about listeners' narrative perceptions (3). In this study, we compared quantitative measures of narrativity (the likelihood that an excerpt of music triggers a story in listeners minds) and narrative engagement (how vivid and clear the events of the story are in listeners minds) for a large set of musical excerpts from Western and Chinese musical traditions for listeners in the same three distinct geographical locations as the present investigation—two suburban college towns in the US Midwest and one rural village in the Chinese province of Guizhou. Results showed that people in all three locations readily narrativize to excerpts (i.e., narrativity scores were quite high) with varying levels of narrative engagement for both Western and Chinese instrumental music; moreover, people do so with about the same degree regardless of location. Notably, however, although both excerpt narrativity and narrative engagement scores were highly correlated across the two US locations, they were not correlated (not predictive) for cross-cultural comparisons between listeners in both of the US locations and the remote rural village in Guizhou.

An obvious limitation of Margulis et al. (3) is that quantitative measures of narrativity and narrative engagement alone do not probe the content of the imagined stories. The central question raised here is whether or not listeners have shared narrative imaginings. Exploring imagined stories in an unconstrained manner—by simply prompting participants to provide free-response descriptions—is essential to ascertaining whether reliable intersubjectivity exists and affords the opportunity to begin to answer a number of provocative questions. Under conditions where a listener is free to describe the subjective contents of their musical imaginings, will their responses diverge in a way that reflects the true idiosyncrasy and interpersonally variable nature of music listening? More generally, what do the accounts of the imagined stories themselves reveal about the extent to which narrative imaginings are "guided" by the music (17, 18) or the extent to which they constitute more arbitrary episodes of mind wandering? To what extent are the stories generated in response to specific musical excerpts shared among individuals within and across cultures? Moreover, how does the capacity of musical features to shape imagined stories intersect with the set of prior experiences and dispositions that any individual listener brings to the story imagining task?

These questions lie at the heart of music's capacity to connect and divide people and groups—while listening to a particular excerpt, two people might imagine the same story or wildly divergent stories. The relationship between those imaginings can render that music an agent of shared experience or the opposite, with ramifications for music's potential role in social bonding, postulated to be a fundamental purpose of music and the explanation for its evolutionary origins (19). These questions also have clinical relevance since the Bonny Method of Guided Imagery and Music is one of the most commonly used musical therapies, applied to a variety of conditions including post-traumatic stress disorder, depression, cancer rehabilitation, and recovery after cardiothoracic surgery (20–23).

To tackle these questions head on, 622 participants in the same three geographical locations as Margulis et al. (3) provided free-response descriptions of the stories imagined when listening to excerpts of instrumental music. Two of the locations were suburban college towns in middle America (one in Arkansas and the other in Michigan). The story responses of these two groups were compared to each other (a within-culture comparison) and to the stories generated by participants in a different geographical location that was culturally distinct—a village in rural China (the Dimen group) where participants speak Dong, a tone language that is not related to Mandarin (ref. 24, p. 230), and where participants have little access to the Western media that might be expected to influence responses. All three groups of listeners (Arkansas, Michigan, Dimen) heard the same 32 musical stimuli. Half the stimuli were excerpts of instrumental Western art music without words and half were excerpts of instrumental Chinese art music without words—parallel in the sense that although Western art music did not form a part of the US participants' daily intentional listening habits, it was broadly familiar as a genre, just as Chinese art music did not form a part of the Dimen

participants' daily intentional listening habits but was broadly familiar as a genre.

If stimulus features exclusively drive the content of the imagined narratives—if stories are directly and universally implied by music—then narratives provided to individual excerpts should be highly similar across all three locations, and stories provided to excerpts from each music tradition (Western or Chinese) should be somewhat similar across locations. If differences at the level of individuals are the primary drivers of the content of the imagined stories—e.g., listeners imagine arbitrary stories unrelated to the musical excerpts themselves—then similarity among responses to individual excerpts across locations, or among responses to excerpts from a particular music tradition (Western or Chinese), should be no greater than the similarity of responses to different excerpts from different music traditions.

If differences at the level of culture are the primary drivers of the content of the imagined stories, then responses should be similar between Arkansas and Michigan participants overall but not within individual excerpts and not within music traditions; this would suggest that people who share a culture (broadly construed) tend to imagine similar stories, but these stories are not driven in consistent ways by the musical excerpts. If, however, stimulus features combine with differences at the level of culture to shape the content of imagined narratives, then responses to individual excerpts (and, to a lesser extent, excerpt types) should be highly similar between Arkansas and Michigan participants but not between Arkansas and Dimen participants or between Michigan and Dimen participants.

In order to compare stories collected at different geographic locations, we examined text similarity using collections of narratives evoked by the same music excerpt. For each location (Arkansas, Michigan, or Dimen), narratives provided in response to the same music excerpt were combined into a single narrative document or "nardoc" (*SI Appendix*, Table S1). Thus, while each document comprises a variety of individual narratives, as a whole, a document represents the narrative themes typically experienced when listening to a music excerpt. Analysis occurred at the level of nardocs for music excerpts rather than responses from individual participants, which allowed us to examine the applicability of Kassabian's (25) concept of "distributed subjectivity"—commonalities that might emerge among participants in response to particular excerpts—rather than focusing on the differences that might exist between individuals. To ascertain whether individual excerpts of music reliably drive distinct narrative imaginings, the semantic similarity of nardocs was compared within and across locations. Similarity was first assessed by visualizing relative nardoc locations in a predefined semantic space followed by a more rigorous assessment of shared semantic content using a permutation-based approach.

Results

Visualization of Semantic Similarity of Imagined Narratives. We visualized the semantic relationships between narrative documents (nardocs) by projecting text into a pretrained embedding space and reducing dimensionality prior to plotting the results. We used a word2vec model pretrained on the Google News dataset (https://code.google.com/archive/p/word2vec/) as a means to obtain feature vectors for words that reflect semantic similarities based on a large corpus of ∼3 million words and phrases rather than relying on the collection of nardocs used in the present study (*Methods*). Fig. 1 shows the average location in the embedding space for narrative documents for each of the 32 excerpts by the listeners in Arkansas, Michigan, and Dimen. The closer two nardocs are in the embedding space, the greater their semantic similarity, but the two dimensions of the embedding space do not themselves have a meaningful interpretation (*Methods*). Dimen listeners' story responses cluster together in a region of

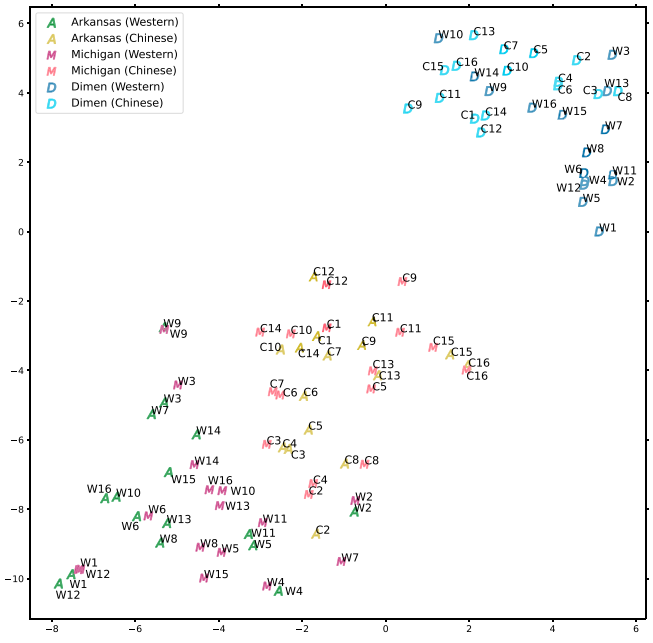

Fig. 1. Visualization of narrative documents (nardocs) in semantic space. Each symbol represents the average position of a nardoc projected into the predefined embedding space. The proximity of two nardocs in embedding space corresponds to the extent of shared semantic content (i.e., points for nardocs that are closer together represent greater semantic similarity). Individual excerpt labels are represented by W or C for music tradition (Western or Chinese) followed by the excerpt number (1 to 16 for each). Locations are labeled by a larger letter (A = Arkansas, M = Michigan, D = Dimen).

the semantic space that is distinct from that occupied by the imagined stories of the Arkansas and Michigan listeners. Furthermore, within each geographical location, responses to music from a particular music tradition (Western or Chinese) largely cluster together. These clusters by music tradition seem to overlap for Arkansas and Michigan participants but not for Arkansas and Dimen or Michigan and Dimen participants. In other words, Arkansas and Michigan stories imagined in response to Western music seem to be similar and Arkansas and Michigan stories imagined in response to Chinese music seem to be similar, but neither of these seem similar to the stories imagined by Dimen participants to these kinds of excerpts.

If imagined stories constituted arbitrary episodes of mind wandering, then such consistent patterns around music tradition and geographic location should not emerge. Moreover, if the imagined stories were arbitrary and idiosyncratic, clustering should not occur around responses to individual excerpts. Instead, responses to individual excerpts between participants at two geographically distinct, but culturally similar sites—Arkansas and Michigan—appear to be highly similar, clustering so closely together in several cases (for excerpts W2, W9, C13, and C16, for example) that the points representing their position in semantic space nearly or entirely overlap. This pattern does not emerge between Arkansas and Dimen stories imagined in response to individual excerpts or between Michigan and Dimen stories. We next examine these trends.

Comparison of Narrative Similarity within and across Music Tradition for Each Geographic Location. For all quantitative comparisons, narrative documents were preprocessed and transformed into a feature vector. Each feature corresponded to the term frequency-inverse document frequency (TF-IDF) score of a word (26). This TF-IDF score measures a word's importance within a particular text document by multiplying its frequency in that document

by the inverse frequency of the word across all documents. Thus, a high TF-IDF score indicates that a word is uniquely relevant to a particular nardoc and a vector of TF-IDF scores reflects the unique composition of words in a nardoc.

If overarching music style characteristics constrain the narratives that listeners hear in the music, then for each group of listeners (Arkansas, Michigan, and Dimen), stories generated in response to musical excerpts from a particular music tradition should be more similar to each other than to stories generated in response to excerpts from a different music tradition. To test this, we measured the cosine similarities between nardoc feature vectors generated in response to excerpts from within the same music tradition (e.g., Western tradition or Chinese tradition) and compared these to the cosine similarities between nardocs feature vectors generated in response to excerpts across music traditions. Cosine similarity expresses the cosine of the angle between the two to-be-compared feature vectors, which varies between 0 and 1. A value of 1 indicates an angle of 0° between the two vectors (maximum similarity) while a value of 0 indicates that the two vectors are at 90° (minimum similarity).

For each geographic location, we used Welch's t test to compare the different-excerpt cosine similarity distributions within and across music traditions, deriving one-tailed difference thresholds (>95th percentile of null) for each comparison by permuting the excerpt labels of participants' narratives at each location prior to combining them into nardocs (*Different-Excerpt Comparisons*). The results, summarized in Fig. 2 and reported as follows, show that the overarching music style characteristics particular to a music tradition (Western vs. Chinese) constrain the narratives that listeners hear in the music in all three geographical locations (Arkansas, Michigan, and Dimen); supporting descriptive statistics are provided in *SI Appendix*, Table S2.

Arkansas. The leftmost three bars of Fig. 2 show that the similarity between nardocs generated by the Arkansas listeners in response to excerpts within the Western tradition and within the Chinese tradition was greater than the similarity between nardocs generated in response to excerpts across traditions (Western-Western vs. Chinese-Western: $t = 8.70$, difference threshold, $t = 3.68$; Chinese-Chinese vs. Chinese-Western: $t = 9.92$. difference threshold, $t = 2.61$). Moreover, the similarity between Western tradition nardocs was not significantly different from the similarity between Chinese tradition nardocs ($t = 0.37$; difference threshold, $t = 6.12$).

Michigan. The middle three bars of Fig. 2 show that the similarity between nardocs generated by the Michigan listeners in response to excerpts within the Western tradition and within the Chinese tradition was greater than it was between nardocs generated in response to excerpts across traditions (Western-Western vs. Chinese-Western: $t = 11.65$, difference threshold, $t = 6.71$; Chinese-Chinese vs. Chinese-Western: $t = 7.78$, difference threshold, $t = 0.01$). Moreover, as found with the Arkansas listeners, the similarity between Western tradition nardocs was not significantly different from the similarity between Chinese tradition nardocs ($t = 3.70$; difference threshold, $t = 10.50$).

Dimen. The rightmost three bars of Fig. 2 show that the similarity between nardocs generated by the Dimen listeners in response to excerpts within the Western tradition and within the Chinese tradition was greater than the similarity between nardocs generated in response to excerpts across traditions (Western-Western vs. Chinese-Western: $t = 2.88$, difference threshold, $t = 1.75$; Chinese-Chinese vs. Chinese-Western: $t = 7.73$, difference threshold, $t = 5.24$). Moreover, as found with the two US samples, the similarity between Western tradition nardocs was not significantly different from the similarity between Chinese tradition nardocs ($t = -3.58$; difference threshold, $t = -8.27$).

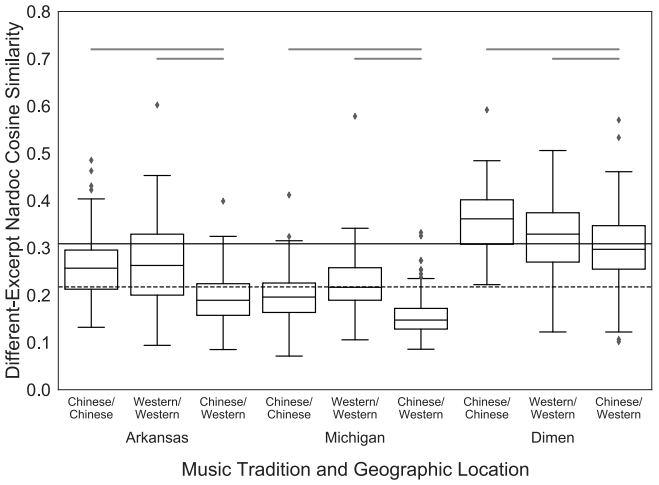

Fig. 2. Comparison of stories prompted by different music excerpts at the same geographic location. For each of the 32 excerpts, we calculated the pairwise cosine similarities between the TF-IDF vectors for narrative documents collected from the same geographic location. We excluded same-excerpt comparisons and instead examined different-excerpt similarity based on the music tradition an excerpt belongs to. The nine box-and-whisker plots depict the median value and quantiles of the distribution of different-excerpt similarity values in each comparison within geographic locations. For each geographic location, we used Welch's t test to compare the different-excerpt similarity distributions within and between music traditions. Individual data points (diamonds) correspond to document similarity values that exceed 1.5× the IQR. Black lines spanning two distributions at the top of the figure represent significant t tests relative to the permuted difference thresholds. The long solid and dotted lines depict the 95th percentile and median value of the control narrative distributions and represent estimates of the maximum and average similarity expected between unprompted stories by US college undergraduates, respectively. The values serve as an additional reference point and not as a threshold for significance.

Comparison of Narrative Similarity within/between Music Traditions across Geographic Locations. Given that music style characteristics constrain narratives within a location (i.e., a distinct set of stories emerges in relation to excerpts from the Western music tradition and the Chinese music tradition, respectively, regardless of geographic location), the next natural question is whether stories generated in response to a particular music tradition share characteristics across locations. To address this question, similarity was compared between nardocs prompted by excerpts within the same music tradition (Western or Chinese) across pairs of locations (Arkansas-Michigan, Arkansas-Dimen, and Michigan-Dimen). For example, for the Arkansas-Michigan comparison of the stories prompted by music excerpts from the Western tradition, we calculated the pairwise similarity between all Western tradition nardocs generated by Arkansas listeners and all Western tradition nardocs generated by Michigan listeners, excluding the 32 same-excerpt comparisons. We then compared the cross-location Western-Western and Chinese-Chinese distributions to the distribution of pairwise similarities between all nardocs generated in response to the two different music traditions at the two locations.

For each cross-location comparison (Arkansas-Michigan, Arkansas-Dimen, Michigan-Dimen), we used Welch's t test to compare the different-excerpt cosine similarity distributions within and between music traditions, deriving difference thresholds (>95th percentile of the null distribution) for each comparison by permuting the excerpt labels of participants' narratives at each location prior to combining them into nardocs (*Different-Excerpt Comparisons*).

The results, summarized in Fig. 3 and reported as follows, show that narratives generated by Arkansas and Michigan listeners were similarly influenced by the overarching music style

characteristics particular to a music tradition. However, the results for cross-culture comparisons at different locations (Arkansas-Dimen and Michigan-Dimen) revealed little commonality across cultures in the types of stories that listeners generated in response to excerpts from either the Western or Chinese music tradition; supporting descriptive statistics are provided in *SI Appendix*, Table S3.

Arkansas-Michigan comparison. The leftmost three bars of Fig. 3 show that similarity was significantly greater between Arkansas and Michigan nardocs of the same music tradition than between Arkansas and Michigan nardocs of different music traditions (e.g., Western/Chinese or Chinese/Western; Western/Western, $t = 14.71$, difference threshold, $t = 5.88$; Chinese/Chinese, $t = 12.89$, difference threshold, $t = 0.32$). The comparison across locations also revealed no overall difference in similarity for nardocs generated in response to excerpts within the Chinese tradition and excerpts within the Western tradition ($t = 3.08$; difference threshold, $t = 9.84$), confirming that shared stories can emerge in response to music from both more and less familiar styles.

Arkansas-Dimen and Michigan-Dimen comparisons. Overall, the pattern of results between geographically distinct locations was different when the two locations did not share an overarching culture. Both between-culture comparisons, shown in the middle three and rightmost three bars of Fig. 3, respectively, revealed no overall difference in similarity for nardocs generated in response to excerpts within the Chinese tradition and excerpts within the Western tradition (Arkansas-Dimen: $t = 3.37$, difference threshold, $t = 4.10$; Michigan-Dimen: $t = 3.26$, difference threshold, $t = 6.61$). However, when music tradition is considered separately, the similarity of nardocs generated in response to excerpts from the Western tradition, but not the Chinese tradition, was greater than the similarity of nardocs generated in response to excerpts from across music traditions for both the Arkansas-Dimen comparison (Western: $t = 4.95$, difference threshold, $t = 2.37$; Chinese: $t = 1.63$, difference threshold, $t = 4.23$) and the Michigan-Dimen comparison (Western: $t = 5.52$; difference threshold, $t = 3.96$; Chinese: $t = 1.76$; difference threshold, $t = 2.47$). Taken together, these results suggest that Chinese music prompts more different associations across cultures than the Western music, likely because of the greater prevalence of Western music in globalized contexts such as mass media.

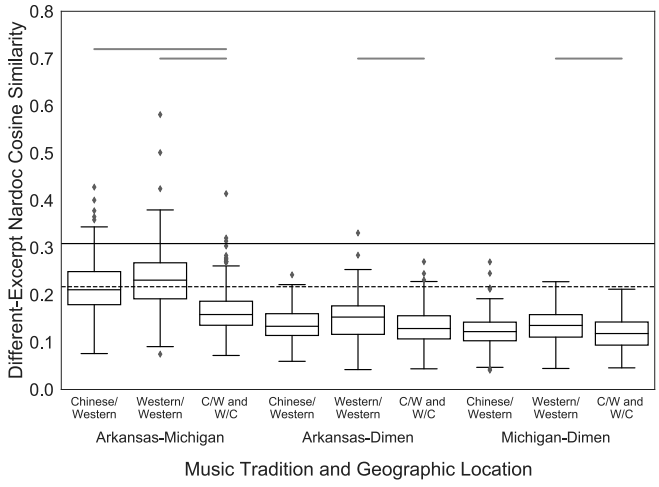

Fig. 3. Comparison of stories prompted by different music excerpts at different geographic locations. For each of the 32 excerpts, we calculated the pairwise cosine similarities between the TF-IDF vectors for narrative documents collected at each geographic location. We excluded same-excerpt comparisons and instead examined different-excerpt similarity based on the music tradition an excerpt belongs to. The nine box-and-whisker plots depict the median value and quantiles of the distribution of different-excerpt similarity values in each comparison between geographic locations. Individual data points (diamonds) correspond to document similarity values that exceed 1.5× the IQR. For each location comparison, we used Welch's *t* test to compare the different-excerpt similarity distributions within and between music traditions. Black lines spanning two distributions at the top of the figure represent significant *t* tests relative to the permuted difference thresholds. The long solid and dotted lines depict the 95th percentile and median value of the control narrative distributions and represent estimates of the maximum and average similarity expected between unprompted stories by US college undergraduates, respectively. The values serve as an additional reference point and not as a threshold for significance.

Comparison of Narrative Similarity for the Same Excerpts across Geographic Locations. The preceding two sections show that the broad category of music tradition constrains patterns of narrative response within and across geographic locations, but how closely do stimulus features drive imagined narratives? If the stimulus features of individual musical excerpts shape narratives across listeners in different geographic locations more directly, then even narratives prompted by the same excerpt should be similar across geographic locations, especially for locations with a shared culture. To address this hypothesis, we examined the percentage of same-excerpt nardoc comparisons that demonstrated significant cosine similarity values for each location pair (Arkansas-Michigan, Arkansas-Dimen, Michigan-Dimen). We derived two separate thresholds for significance: the maximum similarity expected by chance between stories prompted by the 32 different music excerpts used in the study (same-sample threshold; *Excerpt Label Permutations*) and the maximum similarity expected between stories by US college undergraduates promoted by a generic task cue (control-sample threshold; *Control Narrative Permutations*). The results are shown in Fig. 4 and reported as follows.

Arkansas-Michigan comparison. Overall, the majority of excerpts prompted similar narratives in participants from Arkansas and Michigan. The median cosine similarity for nardocs produced in response to the same excerpts across locations was 0.357 (interquartile range [IQR] = 0.325 to 0.417). A total of 65.6% of excerpts (21 out of 32) prompted nardocs from Arkansas listeners that were more similar to the nardocs generated for that excerpt by Michigan listeners than expected by chance based on the same-sample threshold (cosine similarity = 0.346). A total of 84.4% of excerpts (27 out of 32) prompted nardocs from Arkansas listeners that were more similar to nardocs generated for that excerpt by Michigan listeners than expected by chance based on the control-sample threshold (cosine similarity = 0.309). The finding suggests that same-culture listeners are likely to experience similar narratives when listening to the same piece of music.

Arkansas-Dimen and Michigan-Dimen comparisons. Overall, the majority of excerpts prompted much less similar narratives in participants from Arkansas and Dimen and from Michigan and Dimen. The median cosine similarity for nardocs produced in response to the same excerpts across locations was only 0.169 (IQR = 0.149 to 0.191) for the Arkansas-Dimen comparison and 0.153 (IQR = 0.130 to 0.169) for the Michigan-Dimen comparison. Moreover, only a single excerpt (1 out of 32) prompted narratives in either Arkansas or Michigan listeners more similar to Dimen listeners than expected by chance based on the control-sample threshold (cosine similarity = 0.309), and only a single excerpt (1 out of 32) for the Arkansas sample and only 6.3% of excerpts (2 out of 32) for the Michigan sample prompted narratives more similar than expected to Dimen listeners based on the same-sample threshold (cosine similarity = 0.244); potential effects of age across locations are considered in *SI Appendix*, *SI Results* and Table S4. Overall, these results suggest music can evoke similar narratives in different individuals, but only if those

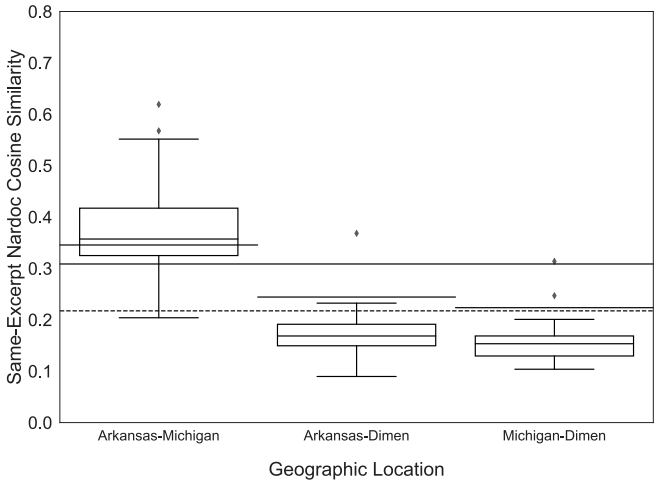

Fig. 4. Comparison of stories prompted by the same music excerpt at different geographic locations. For each of the 32 excerpts, we calculated the cosine similarities between the corresponding TF-IDF vectors for narrative documents collected at each geographic location. The three box-and-whisker plots depict the median value and quantiles of the distribution of same-excerpt similarity values in each comparison between geographic locations. Individual data points (diamonds) correspond to document similarity values that exceed 1.5× the IQR. The values of the solid lines represent particularly conservative estimates of document similarities corrected for multiple comparisons (short lines = same-sample thresholds; long line = control-sample threshold). The long dotted line corresponds to the median value of the Princeton-Michigan permuted distribution and represents an estimate of the average similarity expected between unprompted stories by US college undergraduates. The long dotted line serves as an additional reference point and not as a threshold for significance.

individuals have overlapping life experiences, stemming from (for example) a shared culture.

Discussion

Results show that people spontaneously generate similar stories to instrumental excerpts across geographic locations that share a culture (Arkansas-Michigan); these shared imaginings do not convey across cultural boundaries (Arkansas-Dimen and Michigan-Dimen). Narrative imaginings, which are usually silent and invisible, also appear to be phenomenologically distinct and reportable, revealing shared experiences of music among listeners with shared cultural backgrounds. Given that these imaginings can feel highly subjective and personal, the realization that aspects of them are shared underscores music's role in social bonding, hypothesized as a key driver of natural selection for musical capacities (19). The clear within-excerpt similarity of free responses from listeners at two geographically distinct, culturally similar research sites in the United States reveals that the narrative imaginings people sustain while listening to music are reliably driven by characteristics of the music itself. Yet the fact that this relationship between excerpt and narrative response only holds within but not across cultures reveals that the broadly shared pool of experiences that constitute a culture also plays a formative role in determining the content of music-evoked stories. Thus, any individual story imagined in response to music bears the imprint of both the sounds that triggered it and the prior experiences the listener brings to the encounter. Imagined stories can connect listeners with similar cultural experiences as much as they can divide listeners with different ones, an important corrective to the assumption that music necessarily aids intercultural understanding.

For example, Fig. 5 shows the words most representative of the narratives generated in response to a sample Chinese excerpt (C16) and a sample Western excerpt (W9). While listening to excerpt W9, participants in Arkansas and Michigan tended to imagine a sunrise over a forest, with woodland creatures awaking and birds chirping; participants in Dimen tended to imagine a man blowing a leaf on the top of a mountain, romantically wooing his beloved. While listening to C16, participants in Arkansas and Michigan imagined an old cowboy sitting alone in the scorching desert sun looking out over an empty town; participants in Dimen imagined a man in ancient times sorrowfully contemplating the loss of his beloved. It is striking that individual excerpts prompt such consistently similar imagined stories from participants in Arkansas and Michigan yet equally clear that this similarity does not extend to stories imagined by Dimen participants.

The Western excerpts drew from a musical tradition that features heavily in Western mass media (25). The majority of Dimen participants (63%) reported no exposure to Western media, but some (37%) did report exposure—albeit quite minimal. Although the extent of shared exposure to Western mass media was not sufficient to bring imagined stories in response to individual music excerpts into alignment between the Dimen participants and the Arkansas and Michigan participants, it likely drove the significant difference between the similarity of imagined stories to any one of the Western excerpts to any of the other Western excerpts across all geographic location pairings (not only Arkansas-Michigan but also Arkansas-Dimen and Michigan-Dimen) compared to the similarity of imagined stories to any one of the Western excerpts compared to any one of the Chinese excerpts across these same geographic location pairings. That is, even across cultures, the general kind of story that Western excerpts were understood to imply seemed to be at least somewhat shared, likely due to experience (however slight) with Western mass media (this issue is given additional consideration in *SI Appendix*, *SI Results*).

Furthermore, within each geographic location, stories imagined in response to two excerpts from any one musical tradition (Western or Chinese) tended to be more similar than stories imagined to two excerpts from different musical traditions. This increased similarity in imagined narratives for excerpts from particular musical traditions constitutes additional evidence that stories imagined while listening to music are significantly driven at least in part by characteristics of the sound itself (in this case, whether the excerpt stemmed from a Western or Chinese art music tradition). This interpretation is bolstered by the finding that participants could identify which of two stories was the one typically provided by other participants who shared a culture but not which of two stories was the one typically provided by participants who did not share a culture (5).

There are a number of outstanding questions related to the relative roles of exposure to Western and Chinese mass media in the different geographic locations. Whether the robust within-culture shared associations were gleaned from media or other kinds of cultural experiences, it is still a striking phenomenon. Since we have good reason to believe that the specific excerpts in the study were not used in widely seen media contexts, the convergence in imagined stories—if influenced by media—would have to reflect abstraction away from the individual excerpts to their underlying patterns or attributes, which would then be matched with stored templates that have been abstracted from experience with specific media contexts. This would amount to an internalized musical taxonomy for listening of great potential interest to musicology, cognitive science, and media studies. Arkansas and Michigan participant responses to Chinese excerpts are particularly illustrative of this process. Although it is highly unlikely that Arkansas and Michigan participants encountered the Chinese excerpts in mass market movies, they nevertheless experienced broadly similar imaginings, hearing them in terms of other more familiar sound-sequence pairings—for example, a lonely cowboy scene. Despite the fact that

Fig. 5. Word clouds of nardocs generated in response to a Chinese excerpt (C16) and a Western excerpt (W9). Each cloud depicts the words with the top 30 TF-IDF scores in a nardoc. The size of a word corresponds to the magnitude of its score.

they had almost certainly never heard an excerpt of Chinese art music accompany a cowboy movie, one potential mechanism subserving the shared narrative imagery for this excerpt involves cueing into similarities between the excerpt's acoustic features and music that Arkansas and Michigan participants had heard in movies about the "Old West." An important topic for future research is studying how these mappings develop and function.

Overall, this study's findings suggest that music-evoked narratives are both nonarbitrary and nonuniversal. The discipline of ethnomusicology has long argued for the cultural contingency of most aspects of music making and understanding, but that message is frequently bypassed by scientific research (for thorough critiques of scientific accounts of music and universality, see refs. 27 and 28). By demonstrating the malleability of perceived musical meaning using empirical tools, this study provides a path forward for complex cross-cultural scientific work about music and its communicative potential. The stories we imagine while listening might reveal more about our shared or divergent previous experiences than we would expect. The stories people imagine while listening to instrumental music are fundamental to its aesthetic valuation (29) and to its emotional power (30). Imagination aids transportation, the process whereby an aesthetic stimulus grips the listener, carrying them out of their everyday world and giving them a sense that they have entered the universe of the piece (31), a process that is critical to peak experiences of music (32).

In a broader context, these results provide what is akin to a contemporary version of the Perky experiment from 1910. People imagine stories while listening to music that may seem highly subjective and personal but which in fact follow the contours of the music's acoustic features and of the shared cultural resonances that have accrued around them. This speaks to a broader blend between the more idiosyncratic aspects of any individual's imagination and the shared group-level determiners of any individual's private imagining. A key next question would be if people experience shared narrative imagining to musical prompts, would they experience similar levels of commonality among imaginings produced by other types of prompts, or is there something special about music? This debate recalls arguments around whether music triggers autobiographical memories more frequently and vividly than other prompts (33, 34). Given the empirical similarity between imagined third-person narratives and recalled first-person episodic memories, these two lines of research considered in parallel can shed new light on how music shapes the imagination (see ref. 2). One way that music might be uniquely capable of engendering particularly vivid and dynamic imagined scenes—whether episodes of autobiographical memory or fictional stories—is that it features a series of time-fluctuating characteristics such that changes can dynamically guide and support the contents of the imagination.

Another important topic for future research concerns the relationship between imagined stories and emotional responses to music (see ref. 21). It has been suggested that some aspects of emotional response to music are shared across cultures (35–37), yet we find here that similarities in the semantic content of narrative imaginings do not seem to extend across cultures. To consider whether, in the present study, similarity in the general sentiment of the narratives might extend across cultures, even if the concrete imaginings do not, we had a set of independent judges rate the sentiment (negative to positive) of each of the produced narratives by each participant in all three locations and then examined the correlation of sentiment across pairs of locations (details in *SI Appendix*, *SI Results*). We find that, similar to the semantic similarity results, the Western music excerpts trigger shared sentiment for the within-culture comparison but not for the two between-culture comparisons (Arkansas-Dimen, Michigan-Dimen). The Chinese excerpts also lead to shared sentiment for the within-culture comparison but also lead to shared sentiment for the two between-culture comparisons, suggesting that commonalties in valence can arise independently of commonalities in concrete semantics. It is difficult to make robust inferences from these supplemental sentiment results since most of the stories were neutral, neither positively nor negatively valenced. Stories tended to feature objects, actions, and settings that themselves may have had inspired varied emotional response, depending on the background and perspective of the participant, but the story descriptions simply recounted the story, not the participant's reaction to the story. Future research should expressly assess both narrative imaginings and emotional response to understand the extent to which there are cross-cultural similarities in affective experience when the corresponding concrete semantic imaginings diverge.

Another outstanding question for future work is the degree to which listeners' imagined stories constitute an available response

to instrumental music that participants can call upon when asked versus a pervasive response that participants employ in everyday listening contexts. Earlier work from this project shows that listeners readily imagine narratives in response to instrumental music, do so with relative ease, and tend to enjoy excerpts more when they are more narratively engaged (3). While this suggests that imagining narratives is a broadly accessible and enjoyable form of relating to music that may constitute a pervasive response, there are not yet direct data that address how frequently narrative listening occurs under naturalistic conditions. Thus, this study demonstrates that intersubjective imaginings are an available response—shared semantic resonances are at hand should a listener choose to attend to them—but the extent to which listeners do or do not employ this response in everyday situations requires additional investigation.

One potential approach to this issue would be to consider implicit tasks to measure narrative engagement without cueing participants about the possibility of hearing a story. If people regularly use narrative frameworks to understand instrumental music, then these imagined narratives may play an important mediating role in emotional experiences of music (21). This could potentially explain why Cowen et al. (38) found that highly specific concrete feelings in response to musical excerpts, such as "triumphant" or "dreamy," were preponderant and better preserved across individuals then levels of valence and arousal more broadly—if feeling labels had been mediated by specific imagined dramatic situations, it might have been easier to overlay specific labels on them rather than more abstract dimensions like valence or arousal. Indeed, the presence of intersubjectivity around concrete stories imagined in response to music within, but not across, cultures provides a potential scaffold for the kind of complex, differentiated affective experiences beyond mere positive or negative valence that make music such a rich communicative domain.

Finally, at the broadest level, the present research underscores the fact that music is anything but an "abstract stimulus," as it is often called in the research literature (39–41). Rather, it conveys concrete imaginings that can be broadly shared among listeners but fail to transmit across cultural boundaries, making music a potential driver both of mutual understanding and social bonding but also of division and distancing. Given the possible therapeutic uses of musical imaginings (22), the discovery that such imaginings are tied closely to acoustic features, but with semantic content that is shared within, but not across, cultures, could allow for the development of more precisely calibrated therapeutic tools.

Methods

Participants.

Arkansas group. Three-hundred eighteen individuals (n = 198, female), ages 17 to 30 y (M = 19.0, SD = 1.6) enrolled in a psychology class participated in the experiment individually at the Music Cognition Laboratory at the University of Arkansas in Fayetteville, AR, in exchange for partial course credit. A total of 58% of participants reported no formal music training, defined as explicit instruction in music of any sort; 42% reported 1 to 14 y of music training (M = 5.1, SD = 3.0). Over 98% of participants reported watching English-language media (currently: M = 11.8 h/wk, SD = 12.9; as a child: M = 15.9 h/wk, SD = 14.0). In contrast, only 15% of participants reported ever having been exposed to Chinese-language media. For those 15%, current exposure to Chinese-language media was estimated to be M = 1.4 h/wk, SD = 2.5, while exposure as a child was estimated to be M = 3.4 h/wk, SD = 4.4.

Dimen group. One-hundred forty-seven individuals (n = 126, female), ages 19 to 80 y (M = 46.6, SD = 14.5) from Dimen, China, participated in the experiment individually at the Dimen Dong Community Cultural Research Center in Dimen, Guizhou Province, China. A total of 70% of participants reported no formal music training; 30% reported between 1 and 60 y of music training (M = 14.5, SD = 14.4). Over 97% of participants reported watching Chinese-language media (currently: M = 11.6 h/wk, SD = 8.9; as a child: M = 5.4 h/ wk, SD = 7.9). In contrast, only 37% of participants reported ever having been exposed to English-language media. For those 37%, current exposure to English-language media was estimated to be M = 2.1 h/wk, SD = 3.5, while exposure as a child was estimated to be M = 0.8 h/wk, SD = 3.2. Dimen listeners responded in Dong and narratives were manually translated to English using the following procedure. To minimize the possibility of systematic bias introduced by the translation process, these responses were first translated into Mandarin by translators masked to the hypotheses of the study and masked to the identity of the stimulus prompting each response. The responses were then translated from Mandarin into English by a translator masked to the hypotheses of the study and masked to the identity of the stimulus prompting each response. At each stage, back translation was conducted to confirm accuracy.

Michigan group. The within-culture comparison group to the Arkansas sample consisted of 157 individuals (n = 120, female), ages 18 to 31 y (M = 19.2, SD = 1.8) enrolled in a psychology class who participated in the experiment individually at the Timing, Attention, and Perception Laboratory at Michigan State University in East Lansing, MI, in exchange for partial course credit. A total of 32% of participants reported no formal music training, while 68% reported between 1 and 18 y of music training (M = 5.9, SD = 3.7).

The study was approved by the Survey and Behavioral Research Ethics Committee (SBREC) at the Chinese University of Hong Kong and by the Institutional Review Board (IRB) at the University of Arkansas and Michigan State University. At the Dimen site, informed consent was obtained as approved by the Chinese University of Hong Kong SBREC. The contents of the consent form were explained to participants in Mandarin. For participants who did not speak Mandarin, the consent form was explained in Dong. Both the person explaining the form in Mandarin and the person explaining it in Dong signed the consent form as well as the participant. The participants who required translation also usually did not know how to write. These participants wrote a symbol (标记) on the consent form as their signature. All participants at the Arkansas site signed informed consent, approved by the University of Arkansas IRB. All participants at the Michigan site signed informed consent, approved by the Michigan State University IRB.

Materials.

Stimuli. Stimuli, listed in *SI Appendix*, Table S5, were thirty-two 60-s excerpts drawn from commercial recordings of instrumental music with no lyrics or vocal part. Half of the 32 excerpts (n = 16) were drawn from recordings of Western art music and half (n = 16) from recordings of Chinese art music. The selected excerpts were from a larger set of 128 excerpts for which we had normative data from all three geographical locations that enabled us to select excerpts that were matched for enjoyment and low familiarity (3). Moreover, preliminary pilot work determined that although participants in the Dimen group were broadly familiar with the style of Chinese music presented in the experiment and participants in the Arkansas and Michigan groups were broadly familiar with the style of Western music presented in the experiment, these styles of music were not the ones to which participants tended to listen most frequently, and these specific excerpts were unlikely to be ones that participants had heard prior to the experimental session (i.e., they were relatively novel) nor was it likely that listeners had a priori explicit associations between the selected excerpts and specific films and/or TV shows (see *SI Appendix, SI Methods* for details of a media search for the selected excerpts).

Procedure. Once seated for the experiment, participants were instructed, "You'll be asked to report aspects of your experience listening to musical excerpts, including whether or not you imagined a story while listening. Please do NOT specifically ATTEMPT to imagine a story. Simply listen to the music as you ordinarily would. If you imagine a story, that's fine, and if you don't imagine a story, that's fine too." Following these instructions, each participant heard 1 of 4 subsets of 8 musical excerpts from the full set of 32—4 Western and 4 Chinese. Subsets were rotated across participants so that responses were obtained for all 32 excerpts, with approximately an equal number of participants listening to each subset. Prior to the presentation of each excerpt, participants were told they should try to listen attentively as if they were intending to enjoy the piece. After listening to each excerpt, participants indicated whether they imagined a story or elements of a story while listening to the music (yes/no): the Story Response Question (SRQ). Next, they answered a series of questions about their narrative engagement with the excerpt, reported in McAuley et al. (4). After completing these items, participants responded to one of two free-response questions, depending on their response to the SRQ. If they answered yes to the SRQ, they described the story they imagined in as much detail as they were able. If they answered no to the SRQ, they were asked to speculate about why they did not imagine a story. Requesting free responses in both cases ensured that participants were not

incentivized to select one option or the other simply on the basis of extent of subsequent task demands.

Participants listened to the excerpts over high quality headphones with the presentation order of the eight excerpts within the subset randomized. Different listeners heard different subsets of 8 excerpts so that narrative responses were obtained for all 32 excerpts across all participants. All participants took part in the experiment via individual sessions. The entire experiment took ∼50 min.

Statistical Analyses.

Data cleaning and preprocessing. All preprocessing and analyses were done in Python version 3.6.2 (Python Software Foundation, https://www.python.org/). First, we identified and manually corrected misspelled words. Narrative texts were then stripped of all punctuation and capitalization, and English stop-words from Natural Language Toolkit version 3.5 (42) were removed. In order to focus our analysis on narrative content provided by participants, we also removed words that referenced acoustic features of music excerpts (instrument names from a list of instruments in a catalog of the world's musical instruments and music traditions, e.g., "Chinese" and "Western") and words that stemmed from task instructions (e.g., "imagine," "music," and "story"); see *SI Appendix*, Table S6 for the full list. Next, we used lemmatize() from the Gensim package version 3.8.3 (43) to convert words to their dictionary forms so that inflected forms of words were treated as the same word. Finally, we removed words that were only mentioned a single time across all nardocs, as these words could not be used to assess the similarity of imagined content.

Narrative documents. For each geographic location, narratives provided in response to the same music excerpt were combined into a single narrative document (nardoc). Narrative documents were created for each of the 32 music excerpts heard at each of the three geographic locations (Arkansas, Michigan, Dimen). Participants contributed no more than one narrative per nardoc. Thus, while each document comprises a variety of individual narratives, as a whole, a document represents the narrative themes typically experienced by individuals at a particular geographic location when listening to a particular music excerpt.

Feature vectors. Prior to calculating similarity, we used TfidfVectorizer() from the Scikit-learn package version 0.21.3 (44) with the default parameters to transform each narrative document into a feature vector. All feature vectors have the same length, which is determined by the size of the overall vocabulary (i.e., the number of unique words across all documents), not the number words in a nardoc. Each feature corresponded to the TF-IDF score of a word (26) (details in *SI Appendix, SI Methods*). TF-IDF measures the importance of a word in a given narrative document by multiplying the frequency of a word in the document by the inverse frequency of the word across all documents. Thus, a high TF-IDF score suggests a word is uniquely relevant to a particular document, and a vector of TF-IDF scores reflects the unique composition of words in a document. The TF-IDF vectors were normalized using the Euclidean norm to account for differences in document lengths, which might arise due to differences in sample size; additional analyses that further address sample size considerations are provided in *SI Appendix, SI Results*.

Cosine similarity. We used cosine_similarity() from the Scikit-learn to measure the cosine of the angle between two nardoc feature vectors. Cosine similarity is bounded between 0 and 1 where a value of 1 corresponds to an angle of 0° between the two feature vectors (maximum similarity), while a value of 0 corresponds to an angle of 90° (minimum similarity). The greater the cosine similarity between two feature vectors, the more similar the composition of the corresponding narrative documents. The metric is commonly used alongside TF-IDF in automated text analysis to find related texts in large corpora, e.g., research articles (45), patient records (46), and Wikipedia articles (47). Note, in this case, that measuring the cosine similarity between the Euclidean-normed TF-IDF vectors produces the same result as the inner product of the vectors.

Same-excerpt comparisons. In order to test our hypothesis that stimulus features shape music-evoked imagery, we compared narrative documents prompted by the same music excerpts in geographically and or culturally distinct samples of participants. For each of the 32 excerpts, we calculated the cosine similarities between the corresponding TF-IDF vectors for nardocs collected at each geographic location. We grouped the similarities into separate distributions for each of the three possible between geographic location comparisons (Michigan-Arkansas, Michigan-Dimen, and Arkansas-Dimen). Each distribution comprised the 32 same-excerpt comparisons between two geographic locations. We used two separate permutation-based approaches to assess the significance of same-excerpt similarity values against chance. One method was based on excerpt label permutation and the second method involved permutations of a set of control narratives generated by two control groups (control group details in *SI Appendix, SI Methods*). Comparisons were directional (one-tailed) because of interest was whether the observed value was greater than 95% of the permuted values ("chance").

Excerpt label permutations. The first method to derive null ("chance") similarity distributions for the same-excerpt comparisons used participants' narratives. Excerpt labels were permuted within each geographic location at the level of individual participant stories prior to combining them into narrative documents and comparing same-excerpt similarities as previously described. Labels were randomized such that each participant contributed no more than one narrative per nardoc. On each permutation, we extracted the maximum cosine similarity value from the 32 same-excerpt comparisons between two geographic locations. This process was repeated 2,000 times to create a null distribution for each of the three between geographic location comparisons. We used null distributions of the maximum statistic to control the family-wise error rate and correct for multiple comparisons (48). The 95th percentile of each null distribution represents an estimate of the maximum similarity expected by chance between nardocs prompted by the 32 different music excerpts used in the study (same-sample threshold). We consider same-excerpt similarity values greater than 95% of a permuted null distribution as significant.

Control narrative permutations. The second method to derive null (chance) similarity distributions used control narratives provided by separate samples of participants from Princeton and the Michigan geographic locations (*SI Appendix, SI Methods*). Comparing control narratives provided an additional threshold to assess the significance of same-excerpt similarities. Unlike the narratives generated in response to music excerpts, the control narratives were promoted by a generic task cue (participants wrote short story descriptions during a 60-s silent period following a brief visual cue) and could not be grouped into narrative documents based on a priori labels. As a result, we randomly labeled each narrative provided by a participant as belonging to one of eight possible control documents. Dividing the control narratives into eight documents allowed us to best match the number of narratives in each control document with the number of narratives in each nardoc (*SI Appendix*, Table S1). Participants contributed no more than one narrative per document.

For each of the eight control documents, labels were randomly assigned to the unpromoted control stories prior to combining them into documents. We then measured the pairwise cosine similarities between the TF-IDF vectors for control documents collected at Princeton and Michigan. We did this 1,000 times and created two null similarity distributions by extracting the maximum and median cosine similarity values from all Princeton-Michigan comparisons on each permutation. The 95th percentile of the maximum-based null distribution represents an estimate of the maximum similarity expected between unprompted stories by US college undergraduates (control-sample threshold). We consider same-excerpt similarity values greater than 95% of the permuted null distribution as significant. Of additional interest was the median value of the null distribution, which served as an estimate of the similarity between random, unprompted stories typically generated by US college undergraduates.

Different-excerpt comparisons. In order to test our hypothesis that stimulus features drive music-evoked imagery, we also examined how the music tradition of an excerpt influenced participants' narratives. To do so, we compared distributions of pairwise similarities between nardocs for different excerpts rather than focusing on the similarity of same-excerpt nardoc pairs as described previously. For each of the 32 excerpts, we calculated the pairwise cosine similarities between the corresponding TF-IDF vectors for nardocs collected at each geographic location. We excluded same-excerpt comparisons and instead examined different-excerpt similarity based on the music tradition of excerpts. For each geographic location comparison, we grouped similarities into three separate distributions: narrative documents for Chinese music tradition excerpts, narrative documents for Western music tradition excerpts, and narrative documents for excerpts of different music traditions (Chinese/Western and Western/Chinese). Welch's *t* test was used to compare the different-excerpt cosine similarity distributions within and between music traditions for each geographic location comparison. We used a permutation-based approach to assess the significance of the *t* tests.

We also ran the same analyses for the pairwise cosine similarities between the corresponding TF-IDF vectors for nardocs collected at the same geographic location. For each geographic location, we grouped similarities into three separate distributions: nardocs for Chinese music tradition excerpts, nardocs for Western music tradition excerpts, and nardocs for excerpts of different music traditions (Chinese-Western and Western-Chinese). Welch's *t* test was used to compare the different-excerpt cosine similarity distributions within and between music traditions for each geographic location. We used a permutation-based approach to assess the significance of the *t* tests.

Excerpt label permutations. We used participants' narratives to derive null *t*-statistic distributions. Excerpt labels were permuted within each geographic location at the level of individual subject stories prior to combining them into nardocs and comparing different-excerpt similarities as described previously. Labels were randomized such that each participant contributed no more than one narrative per nardoc. On each permutation, we calculated Welch's *t* test between different-excerpt comparisons between two geographic locations. We did this 2,000 times to create a null distribution for each of the nine between geographic location comparisons. The 95th percentile of each null distribution represents the maximum difference expected by chance (difference threshold). We considered comparisons between different-excerpt similarity distributions as significantly different if the *t*-statistic was greater than 95% of the values from the corresponding permuted null distribution (a one-tailed test).

We conducted similar permutation tests to derive null *t*-statistic distributions for within geographic location comparisons. On each permutation, we calculated Welch's *t* test between different-excerpt comparisons for each geographic location. We did this 2,000 times to create a null distribution for each of the nine within geographic location comparisons. The 95th percentile of each null distribution represents the maximum difference expected by chance. We considered comparisons between different-excerpt similarity distributions as significantly different if the *t*-statistic was greater than 95% of the values from the corresponding permuted null distribution (a one-tailed test).

Visualization. We visualized the semantic relationships between narrative documents by projecting text into a pretrained embedding space and reducing dimensionality prior to plotting the results. We used the Python package Gensim to load a word2vec model pretrained on the Google News dataset (https://code.google.com/archive/p/word2vec/). The model provided a means to obtain feature vectors for words that reflect semantic similarities based on a large corpus of ∼3 million words and phrases rather than relying on the collection of nardocs used in the present study. The closer two words are in embedding space, the greater the semantic similarity. Each word within a narrative was projected into the 300-dimensional embedding space. Next, in order to equally weight each person's narrative within a nardoc, we averaged across word embeddings in each narrative. We then averaged across narratives to find the average position in embedding space for each nardoc. We used the Scikit-learn implementation of principal components analysis [PCA()] to reduce the dimensionality of the average nardoc embeddings to 50 before further reducing to two-dimensional space using the Scikit-learn implementation of *t*-distributed stochastic neighboring entities [T-SNE()] and, finally, plotting the two-dimensional nardoc vectors. We chose to first reduce the 300-dimension feature vectors using PCA based on the Scikit-learn documentation's recommendation for reducing noise in the T-SNE computations. We used the default TSNE() parameters with perplexity set to 30 and random state set to 5. We ran the computations using a 2.60 GHz Intel Core i5 processor with 16 GB of RAM running Windows 10 Home (version 21H1) on a 64-bit operating system. While the relationships between vectors in this space are interpretable, the reduced dimensions for the space are, at best, only coincidently interpretable. This is because the word2vec projection prior to the dimension reduction is nonlinear and probability weighted (there is not a unique solution when training the model). Thus, the reduced number of dimensions in Fig. 1 is not interpretable in the way that one would expect a multidimensional scaling solution to have interpretable dimensions. See ref. 49 for a formal discussion.

Data Availability. All quantitative data have been deposited in OSF (https://osf.io/43nqy/) (50).

ACKNOWLEDGMENTS. Xin Kang, Jieqiong Che, Xiyu Wang, Xueying Xu, Zhentin Liu, Chunzi Li, Xiaotong Ge, and Shengnan Zhao helped with data collection and translation for Dimen participants. We thank Mr. Lee Wai Kit and the staff at the Dimen Dong Eco-Museum for making data collection possible and we also thank the people in Dimen who participated in this research. Jewelian Fairchild, Gabby Kindig, and Anusha Mamidipaka helped with the collection of the control data. We also thank Natalie Phillips and the members of the Digital Humanities and Literary Cognition and Timing, Attention and Perception laboratories at Michigan State University for their many helpful and insightful comments at various stages of this project. Members of the University of Arkansas Music Cognition Laboratory and the Princeton Music Cognition Laboratory helped collect the data and contributed important insights to the project. This research was supported by the Division of Behavioral and Cognitive Sciences of the NSF, Award Numbers 1734063 (Principal Investigator: J.D.M.) and 1734025 (Principal Investigator: E.H.M.).

1. R. Herbert, *Everyday Music Listening: Absorption, Dissociation and Trancing* (Ashgate, 2011).
2. K. Jakubowski, "Musical imagery" in *The Cambridge Handbook of the Imagination*, A. Abraham, Ed. (Cambridge, 2020), pp. 187–206.
3. E. H. Margulis, P. C. M. Wong, R. Simchy-Gross, J. D. McAuley, What the music said: Narrative listening across cultures. *Palgrave Commun.* **5**, 146 (2019).
4. J. D. McAuley, P. C. M. Wong, L. Bellaiche, E. H. Margulis, What drives narrative engagement with music? *Music Percept.* **38**, 509–521 (2021).
5. J. D. McAuley, P. C. M. Wong, A. Mamidipaka, N. Phillips, E. H. Margulis, Do you hear what I hear? Perceived narrative constitutes a semantic dimension for music. *Cognition* **212**, 104712 (2021).
6. D. C. Rubin, S. Umanath, Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychol. Rev.* **122**, 1–23 (2015).
7. D. Hassabis, D. Kumaran, S. D. Vann, E. A. Maguire, Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1726–1731 (2007).
8. B. Nanay, The role of imagination in decision-making. *Mind Lang.* **31**, 127–143 (2016).
9. C. W. Perky, An experimental study of imagination. *Am. J. Psychol.* **21**, 422–452 (1910).
10. P. Tagg, R. Clarida, *Ten Little Title Tunes: Towards a Musicology of the Mass Media* (Mass Media Music Scholars' Press, 2003).
11. E. Huovinen, A.-K. Kaila, The semantics of musical topoi: An empirical approach. *Music Percept.* **33**, 217–243 (2015).
12. E. H. Margulis, An exploratory study of narrative experiences of music. *Music Percept.* **35**, 235–248 (2017).
13. S. Koelsch *et al.*, Music, language and meaning: Brain signatures of semantic processing. *Nat. Neurosci.* **7**, 302–307 (2004).
14. N. Steinbeis, S. Koelsch, Comparing the processing of music and language meaning using EEG and fMRI provides evidence for similar and distinct neural representations. *PLoS One* **3**, e2226 (2008).
15. S. Koelsch, Towards a neural basis of processing musical semantics. *Phys. Life Rev.* **8**, 89–105 (2011).
16. J. G. Painter, S. Koelsch, Can out-of-context musical sounds convey meaning? An ERP study on the processing of meaning in music. *Psychophysiology* **48**, 645–655 (2011).
17. O. Vartanian, "Imagination in aesthetic experience" in *The Cambridge Handbook of the Imagination*, A. Abraham, Ed. (Cambridge, 2020), pp. 578–592.
18. A. Sheppard, The role of imagination in aesthetic experience. *J. Aesthet. Educ.* **25**, 35–42 (1991).
19. P. E. Savage *et al.*, Music as a coevolved system for social bonding. *Behav. Brain Sci.* **44**, e59 (2020).
20. D. Grocke, The role of the therapist in the Bonny Method of Guided Imagery and Music (BMGIM). *Music Ther. Perspect.* **23**, 45–52 (2005).
21. M.-F. Lin *et al.*, Pivotal moments and changes in the Bonny Method of Guided Imagery and Music for patients with depression. *J. Clin. Nurs.* **19**, 1139–1148 (2010).
22. R. L. Blake, S. R. Bishop, The Bonny Method of Guided Imagery and Music (GIM) in the treatment of post-traumatic stress disorder (PTSD) with adults in the psychiatric setting. *Music Ther. Perspect.* **12**, 125–129 (1994).
23. D. S. Burns, The effect of the Bonny Method of Guided Imagery and Music on the mood and life quality of cancer patients. *J. Music Ther.* **38**, 51–65 (2001).
24. S. R. Ramsey, *The Languages of China* (Princeton University Press, 1999).
25. A. Kassabian, *Hearing Film: Tracking Identifications in Contemporary Hollywood Film Music* (Routledge, 2001).
26. S. E. Robertson, K. Sparck Jones, Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **27**, 129–146 (1976).
27. D. Loughridge, 'Always already technological': New views of music and the human in musicology and the cognitive sciences. *Music Res. Annu.* **2**, 1–22.
28. N. Jacoby *et al.*, cross-cultural work in music cognition: Challenges, insights and recommendations. *Music Percept.* **37**, 185–195 (2020).
29. N. Wiltsher, A. Meskin, "Art and imagination" in *The Routledge Handbook of the Philosophy of Imagination*, A. Kind, Ed. (Routledge, 2016), pp. 179–191.
30. P. N. Juslin, D. Västfjäll, Emotional responses to music: The need to consider underlying mechanisms. *Behav. Brain Sci.* **31**, 559–575, discussion 575–621 (2008).
31. M. Strick, H. L. de Bruin, L. C. de Ruiter, W. Jonkers, Striking the right chord: Moving music increases psychological transportation and behavioral intentions. *J. Exp. Psychol. Appl.* **21**, 57–72 (2015).
32. A. Gabrielsson, *Strong Experiences with Music: Music Is Much More Than Just Music* (Oxford University Press, 2011).
33. A. M. Belfi, B. Karlan, D. Tranel, Music evokes vivid autobiographical memories. *Memory* **24**, 979–989 (2016).
34. B. M. Kubit, P. Janata, Spontaneous mental replay of music improves memory for incidentally associated event knowledge. *J. Exp. Psychol. Gen.*, 10.1037/xge0001050 (2021).
35. L.-L. Balkwill, W. F. Thompson, A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Percept.* **17**, 43–64 (1999).

36. L.-L. Balkwill, W. F. Thompson, R. Matsunaga, Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Jpn. Psychol. Res.* **46**, 337–349 (2004).
37. P. Laukka, T. Eerola, N. S. Thingujam, T. Yamasaki, G. Beller, Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion* **13**, 434–449 (2013).
38. A. S. Cowen, X. Fang, D. Sauter, D. Keltner, What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1924–1934 (2020).
39. M. Tervaniemi, S. Maury, R. Näätänen, Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. *Neuroreport* **5**, 844–846 (1994).
40. V. N. Salimpoor, M. Benovoy, K. Larcher, A. Dagher, R. J. Zatorre, Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nat. Neurosci.* **14**, 257–262 (2011).
41. P. D. Fletcher, C. N. Clark, J. D. Warren, Music, reward and frontotemporal dementia. *Brain* **137**, e300 (2014).
42. S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python* (O'Reilly Media, 2009).
43. R. Řehůřekm, P. Sojka, "Software framework for topic modelling with large corpora" in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, 2010), pp. 45–50.
44. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. G. Salton, C. Buckley, Term-weighted approaches in automatic text retrieval. *Inf. Process. Manage.* **24**, 513–523 (1988).
46. F. S. Roque *et al.*, Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).
47. D. M. Lydon-Staley, D. Zhou, A. S. Blevins, P. Zurn, D. S. Bassett, Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nat. Hum. Behav.* **5**, 327–336 (2021).
48. P. H. Westfall, S. S. Young, S. Paul Wright, On adjusting P-values for multiplicity. *Biometrics* **49**, 941–945 (1993).
49. C. Szegedy *et al.*, Intriguing properties of neural networks. arXiv [Preprint] (2014). https://arxiv.org/abs/1312.6199 (Accessed 18 December 2021).
50. E. H. Margulis, P. C. M. Wong, C. Turnbull, B. M. Kubit, J. D. McAuley, Narratives imagined in response to instrumental music. Open Science Framework. https://osf.io/43nqy/. Deposited 17 December 2021.