Cognition 212 (2021) 104712

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Do you hear what I hear? Perceived narrative constitutes a semantic dimension for music

J. Devin McAuley [a,*], Patrick C.M. Wong [b], Anusha Mamidipaka [a], Natalie Phillips [a], Elizabeth Hellmuth Margulis [c,**]

[a] Michigan State University, USA
[b] Chinese University of Hong Kong, China
[c] Princeton University, USA

ARTICLE INFO

Keywords:
Music cognition
Narrative
Musical meaning
Cross-cultural comparison
Intersubjectivity

ABSTRACT

Music has attracted longstanding debate surrounding its capacity to communicate without words, but little empirical work has addressed the topic. Here, 534 participants in the US and a remote region of China participated in two experiments using a novel paradigm to investigate narrative perceptions as a semantic dimension of music. Participants listened to wordless musical excerpts and determined which of two presented stories was the correct match. Correct matches were stories previously imagined by individuals from the US or China in response to each of the excerpts, while foils were correct matches to one of the other tested excerpts. Results revealed that listeners from Arkansas and Michigan had no difficulty matching the music with stories generated by Arkansas listeners. Wordless music, then, far from an abstract stimulus, seems to engender shared, concrete narrative perceptions in listeners. These perceptions are stable and robust for within-culture participants, even at geographically distinct locales (e.g. Arkansas and Michigan). This finding refutes the notion that music is an asemantic medium. In contrast, participants in both the US and China had more difficulty determining correct story-music matches for stories generated by participants from another culture, suggesting that a sufficiently shared pool of experiences must exist for strong intersubjectivity to arise.

1. Introduction

Instrumental music has beguiled thinkers for centuries, attracting vigorous debate surrounding its capacity to communicate without words. Absolutists hold that music is incapable of referencing entities outside itself, while referentialists hold that music can generate this kind of meaning (Meyer, 1956; see also Davies, 1994; Radocy & Boyle, 2012; Scruton, 1997). Composer Igor Stravinsky argued that "music, by its very nature, is essentially powerless to *express* anything at all, whether a feeling, an attitude of mind, a psychological mood, a phenomenon of nature, etc." (Stravinsky, 1936). Another twentieth-century composer, Witold Lutoslawski, referred to music as an "asemantic art" (Skowron, 2007, pp. 102–103). Psychologists have couched similar notions in these terms: "whereas language is understood by reference to an extralinguistic designated space, music is self-referential" (Besson & Friderici, 1998). Yet people often report a powerful sense that wordless music seems to be *about* something; Ian Cross (2012) argues that music is nevertheless non-referential: the aboutness is "floating" and doesn't actually bottom out at consistent concrete referents.

Although empirical evidence could presumably shed light on these arguments, investigations of music's semantic dimensions are surprisingly scant. Children have been shown to successfully match animal referents with themes from *Peter and the Wolf* (Trainor & Trehub, 1992), and adults unfamiliar with Wagner's operas provided consistent, non-arbitrary responses on referential rating scales (kindness-cruelty, natural-supernatural, etc.) to individual leitmotivs (HaCohen & Wagner, 1997). In addition, snippets of music have been shown to elicit the N400 semantic priming effect (Koelsch et al., 2004)—an example featuring pitch intervals set close together, for example, primed "narrowness." It is unclear how a semantic dimension might function within fuller musical excerpts extending beyond a few seconds.

Given the broader human propensity to narrativize (Abbott, 2008; Heider & Simmel, 1944; Sarbin, 1986) and the existence of a network of high-level brain regions that respond to narrative perception regardless

* Correspondence to: J. D. McAuley, Department of Psychology, Michigan State University, East Lansing, MI 48840, USA.
** Correspondence to: E. H. Margulis, Department of Music, Princeton University, Princeton, NJ 08544, USA.
E-mail addresses: dmcauley@msu.edu (J.D. McAuley), margulis@princeton.edu (E.H. Margulis).

https://doi.org/10.1016/j.cognition.2021.104712
Received 6 July 2020; Received in revised form 22 March 2021; Accepted 27 March 2021
Available online 10 April 2021

of modality (Nguyen, Vanderwal, & Hasson, 2019), one possible semantic dimension for longer musical examples involves narrative engagement—listening to wordless music as if it were telling a story (Huovinen and Kaila, 2015; Margulis, 2017; Margulis, Wong, Simchy-Gross, & McAuley, 2019; Tagg & Clarida, 2003). To this end, Margulis et al. (2019) developed a survey instrument to measure narrative engagement with instrumental musical excerpts that included questions on story vividness, clarity, and the extent to which the story was imagined during listening, while also gathering free response descriptions of associated stories that individuals imagined. For US participants in two different geographical locations (Arkansas and Michigan) and residents of a rural village in China (Dimen), a number of similarities emerged across cultures. Regardless of location, individuals showed high levels of narrative engagement with wordless Western and Chinese music. That is, whether an individual was from Arkansas, Michigan or Dimen, wordless Western and Chinese musical excerpts readily triggered vivid stories in people's minds while they were listening. Moreover, the content of the story descriptions generated in response to each excerpt were broadly consistent *within a culture* (soldiers marching off to battle for excerpt Y, a mouse frenetically trying to escape a cat for excerpt Z, etc.).

Results also revealed a number of interesting cross-cultural differences. First, although there was a high level of narrative engagement for both types of music in all groups, listeners tended to imagine stories to a greater degree for the musical style more familiar to them (Western for Arkansas and Michigan participants, and Chinese for Dimen participants). Second, although there was a high degree of within-culture consistency in the degree of narrative engagement and the stories generated in response to a particular musical excerpt, there was little between-culture regularity in either. Participants in Michigan and Arkansas, in other words, tended to most strongly narratively engage to the same individual excerpts, but there was no relationship between this set of excerpts and the set to which participants in Dimen tended to narratively engage the most.

This within-, but not between-culture agreement about which excerpts led to high levels of narrative engagement and the similar within-, but not between-culture agreement in the free response descriptions of the stories participants imagined in response to individual excerpts, supports the referentialist view that music may have the capacity to convey relatively concrete narratives to listeners – at least for those with broadly shared cultural experiences. For example, in response to a specific atonal excerpt, participants in Arkansas and Michigan both described stories involving horror and murder, but participants in Dimen, China described stories involving happy games with friends.

One question that arises from this work is whether excerpts of wordless music imply stories with sufficient clarity that new listeners will be able to correctly identify the story that previous participants imagined for individual musical excerpts. Rephrased in response to the age-old debate on music's potential for referential meaning: does music possess a semantic dimension in the form of specific implied narratives for listeners with shared cultural experiences? To address this question, two experiments employed a novel experimental paradigm at research sites in the US (Arkansas and Michigan) and rural China (Dimen) where residents speak Dong, a tone language with no written version, and—critically—have little exposure to English-language media. Participants heard wordless musical excerpts (Western and Chinese) followed by two short story descriptions. One was the consensus imagined story generated for that excerpt by a participant from Margulis et al. (2019), while the other was a foil story that had been generated in response to another of the tested excerpts. Consensus Stories were selected on the basis that they had the most in common with other participants' stories for that excerpt. The participants' task was to determine which of the two stories had actually been imagined by another participant in response to the heard excerpt.

Two features of the stimuli are important to emphasize. One, the Consensus Stories were specific narratives generated by participants in Margulis et al. (2019), chosen because they best represented the larger pool of stories provided for that excerpt. Thus, the Consensus Stories were individual responses, not composite stories combining elements across multiple individuals. Two, the foil stories were comprised of correct Consensus Stories for other musical excerpts. Thus, inherent differences between correct and foil stories were controlled for by the fact that for each musical excerpt, (a) the foil story presented was the correct story for another musical excerpt in the study and (b) across participants, participants were presented with all possible correct-foil story pairings.

In Experiment 1, Arkansas and Dimen participants completed the matching task using Arkansas-generated narratives—Consensus Stories generated by Arkansas participants in Margulis et al. (2019). In Experiment 2, Michigan participants completed the matching task either for Arkansas-generated narratives or Dimen-generated narratives—Consensus Stories generated by previous Dimen participants (see Table 1 for examples). If listeners are able to identify the stories that other listeners imagine in response to wordless music, then they should be able to match stories to the corresponding musical excerpts. Moreover, if the narrative implication of each musical excerpt is shared broadly across cultures, then participants at all research sites (Arkansas, Michigan, Dimen) should perform well on the matching task for both Arkansas and Dimen-generated narratives.

At the opposite end of the continuum, if wordless music does not carry a semantic dimension at all (i.e., there is no common narrative triggered in response to individual musical excerpts), as the absolutists would have it, then participants in Experiment 1 and 2 should perform poorly on the story-music matching task for both Arkansas- and Dimen-generated narratives. Finally, if musical excerpts bring specific narratives to mind only for participants with broadly shared cultural experiences, then for Experiment 1, the Arkansas participants should be able to match Arkansas-generated stories to excerpts, but the Dimen participants should have more trouble. Similarly, for Experiment 2, which tests a single participant group (i.e., Michigan) on both the Arkansas- and Dimen-generated narratives, participants should be able to successfully match the Arkansas-generated stories to excerpts, but have more difficulty correctly matching the consensus Dimen-generated narratives.

2. Experiment 1

2.1. Methods

2.1.1. Participants and design

The experiment implemented a 2 (Group: Arkansas vs. Dimen) x 2 (Culture of Excerpt: Western vs. Chinese) mixed-factorial design. The Arkansas group consisted of 159 participants from the suburban community of Fayetteville, Arkansas ($n = 95$, female), ages 18–50 years ($M = 19.4$ years, $SD = 2.7$) who took part in the study at the Music Cognition Lab at the University of Arkansas in Fayetteville, AR.

Table 1
Examples of Consensus Arkansas- and Dimen-generated narratives for a Western musical excerpt and a Chinese musical excerpt.

Musical Excerpt	Arkansas Consensus Narrative	Dimen Consensus Narrative
Aaron Copland, Billy the Kid, II. Street in a Frontier Town	A little girl skips through town, when she sees a kitten and tries to catch it. Eventually, more children join the game.	People are dancing on the square, all together, in the evening.
Guan Pinghu, Strains of Spring Morning	A man walked around an empty town with only dirty roads staring at all the old run-down shops.	In the evening, a girl was a little sentimental, and quarreled with her boyfriend, but her boyfriend comforted her in various ways, and ultimately they reconciled.

Approximately 57% of participants reported no formal music training; the remaining 43% had between 1 and 23 years of music training (M = 4.66, SD = 3.5). Training for these participants most frequently consisted of choir or band classes at school. Nighty-8% of the Arkansas participants (n = 156) reported watching English language media (Mdn = 10 h per week, Range: 0.5–84). Conversely, only 10% of the Arkansas participants (n = 17) reported watching any Chinese language media (Mdn = 1 h per week, Range: 0.5–6). The Dimen group consisted of 104 participants (n = 86, female), ages 19–73 years (M = 41.4, SD = 12.6) who were from Dimen, Guizhou Province, China, which is a remote village in a rural part of China where people speak Dong, a tone language without a written version, and have little exposure to English-language media. Participants in the Dimen group took part in the study at the Dimen Dong Community Cultural Research Center. Approximately 68% of participants reported no formal music training; the remaining 32% of participants reported between 1 and 45 years of music training (M = 13.5, SD = 12.6). Training for these participants more frequently consisted of singing Big Song, a Dong tradition of polyphonic singing (Ingram, 2007). For the Dimen sample, approximately 60% of the Dimen participants (n = 62) reported no English-language media exposure. For the remaining 40% (n = 42), the median number hours per week that they reported watching English language media was <1 h per week (Range: 0.5–14). Conversely, 98% of the Dimen participants (n = 102) reported watching Chinese language media (Mdn = 14 h per week, Range: 0.5–50). Participants were told that by English-language media, we meant movies, television, and videos produced primarily in America for American audiences and by Chinese-language media, we meant movies, television and videos produced in China for primarily Chinese audiences. Sample sizes for both groups were sufficiently large to detect a small-to-medium effects with power = 0.80 for between-group comparisons.

2.1.2. Materials

Stimuli were 12 one-minute excerpts, 6 drawn from Chinese art music and 6 from Western art music, which had elicited broadly similar free response descriptions of imagined stories from previous participants at the Arkansas research site. Participants in a previous study involving both research sites had rated these specific excerpts as unfamiliar (Margulis et al., 2019). To generate the matching Consensus Narrative for each excerpt, free-response narratives previously provided by participants in Arkansas in response to each excerpt (Arkansas-generated narratives) were first coded for story content by two coders blind to the hypotheses of the study. The coders first generated a list of possible story features to use throughout the coding process, and then tagged each narrative for every story feature that was present. Only story features tagged by both coders were retained in the list of that narrative's story content. The largest set of free-response narratives that shared story content (as determined by feature tags) formed the set of potential consensus narratives for the musical excerpt. The selected matching narrative (the Consensus Narrative) was the free-response narrative (i.e., individual participant story) from the set of potential consensus narratives that was determined by the coders to be most representative of the shared story content (where most representative was defined subjectively as the story that seemed most prototypical and illustrative of the set of consensus narratives; when the coders disagreed on which story was most representative, they corresponded to ensure that their choices were equally representative, and then selected one randomly. This happened in cases where separate stories were almost exactly the same, differing only slightly in wording). The pool of 'incorrect' (Foil) narratives for that excerpt consisted of the Consensus Narratives for the 11 other musical excerpts tested in the study. Thus, the experiment tested matching performance for 12 musical excerpts with 12 Consensus Narratives based on Arkansas-generated narratives. Translations of narratives were made and refined by using extensive back-translation to ensure accuracy, and native Dong speakers evaluated all the translated stories to ensure they sounded equally natural and understandable to the Dimen participant group.

2.1.3. Procedure

Participants listened to a musical excerpt, followed by the presentation of two short narrative descriptions in a two-alternative forced-choice (2AFC) task. Their task was to select which of two narratives someone else had previously imagined in response to the musical excerpt. The two possible narratives consisted of the Consensus Narrative to that excerpt, which was the correct match to the excerpt, plus a Foil Narrative, which consisted of a Consensus Narrative for a different excerpt in the set of 12 tested excerpts. Consensus-Foil pairings were rotated so that across all participants, all possible foils were paired with each excerpt. For each participant, an equal number of matching Consensus Narratives appeared in the first position and in the second position, and whether the correct matching narrative occurred in the first or second position for a given excerpt was counterbalanced across participants. The excerpt (trial) order was randomized for each participant. Participants completed 12 trials (one for each of the target musical excerpts; half of the excerpts were Western in origin, while the other half of the excerpts were Chinese in origin). Across the 12 trials, participants experienced each Consensus Narrative twice – once as the correct response and once as a foil for another excerpt's Consensus Narrative.

Participants at the Arkansas site selected the correct narrative from a visual display of the two story texts. The display remained on screen until they chose a response. Participants at the Dimen site, whose language lacks a written version, selected the correct narrative from an auditory presentation of the two auditory recordings. Dimen participants were able to re-listen to the audio recordings as many times as they wanted before choosing a response. The entire experiment took ~20 min.

2.2. Results and discussion

Fig. 1 shows box-and-whisker plots of matching performance of the Arkansas participants and the Dimen participants for the Arkansas-generated narratives for the Western and Chinese excerpts. A 2 (Group: Arkansas vs. Dimen) x 2 (Culture of Excerpt: Western vs. Chinese) mixed-measures ANOVA revealed a main effect of group, $F(1,261) = 510.7$, $p < 0.001$, $\eta^2_p = 0.66$, a main effect of Culture of Excerpt, $F(1,261) = 13.97$, $p < 0.001$, $\eta^2_p = 0.05$, but no interaction between Group and Culture of Excerpt, $F(1,261) = 2.01$, $p = 0.16$, $\eta^2_p = 0.008$. Matching performance for the Arkansas participants on the Arkansas-generated narratives was close to ceiling for both the Western excerpts (M = 0.90, 95% CI = 0.88–0.93) and the Chinese excerpts (Arkansas, M = 0.87, 95% CI = 0.84–0.90) with performance for the Western excerpts slightly better than for the Chinese excerpts, $t(158) = 2.30$, $p = 0.02$, Cohen's d = 0.18. In contrast, matching performance for the Dimen participants on the Arkansas-generated narratives was much worse than the performance of the Arkansas participants. Dimen performance on the Western excerpts was slightly above chance for the Western excerpts (M = 0.56, 95% CI = 0.53–0.59, $t(103) = 2.89$, $p < 0.01$, Cohen's d = 0.28) and no different than chance for the Chinese excerpts (M = 0.48, 95% CI = 0.45–0.52, $t(103) = -0.91$, $p = 0.36$, Cohen's d = 0.09).

Next, we considered factors that might account for individual differences in Dimen participant matching performance on the Arkansas-generated narratives. One factor we considered was self-reported exposure to Western media. Although exposure to Western culture in Dimen is minimal, some participants did report watching English-language media. Fig. 2 shows a box-and-whisker plot of matching Dimen participants with and without Western media exposure for the Arkansas-generated narratives for the Western and Chinese excerpts. Consistent with the hypothesis that exposure to Western media would afford extra-musical associations similar to those of the Arkansas participants for Western musical excerpts, Dimen participants' matching performance on Arkansas-generated narratives for Western excerpts was

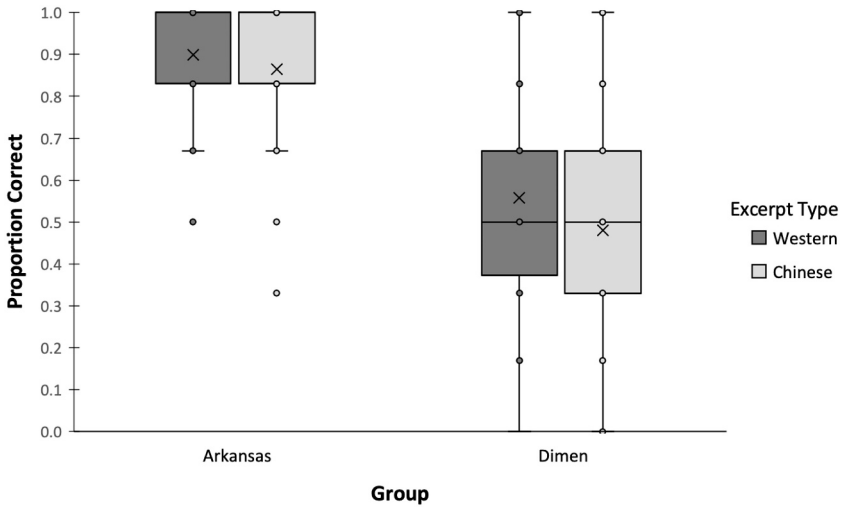

Fig. 1. Box-and-whisker plot showing story-to-excerpt matching performance (Experiment 1) for the Arkansas participants (left pair of boxes) and Dimen participants (right pair of boxes) for the Arkansas-generated narratives for Western musical excerpts (dark gray) and Chinese musical excerpts (light gray). Chance performance corresponds to PC = 0.5. The Arkansas participants could successfully identify the imagined stories of other individuals from the same geographical location in response to wordless musical excerpts with a very high degree of accuracy. The Dimen participants, in contrast, performed, on average, at chance or slightly greater than chance levels on the same task, but also revealed large individual differences.

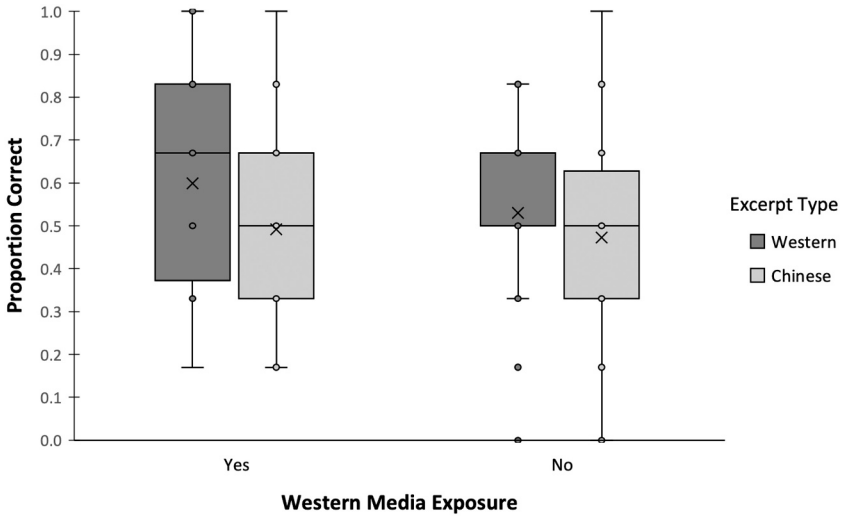

Fig. 2. Box-and-whisker plot showing story-to-excerpt matching performance (Experiment 1) for the Dimen participants who reported any Western media exposure (left pair of boxes) and those who reported no Western media exposure (right pair of boxes) for the Arkansas-generated narratives for Western musical excerpts (dark gray) and Chinese musical excerpts (light gray). Chance performance corresponds to PC = 0.5. Dimen participants' matching performance on Arkansas-generated narratives for Western excerpts was better if they reported exposure to Western media than if they did not. Western media exposure, in contrast, did not reliably affect matching performance for the Chinese excerpts.

better if they reported exposure to Western media ($M = 0.60$, $SD = 0.22$) than if they did not ($M = 0.53$, $SD = 0.18$), $t(102) = 1.73$, $p < 0.05$, one-tailed, Cohen's $d = 0.34$. Western media exposure, in contrast, had no impact on matching performance for the Chinese excerpts, $t(102) = 0.44$, $p = 0.66$, Cohen's $d = 0.09$. Exploratory analyses additionally revealed no reliable correlations between either age or music training (MT) in years with matching performance for either the Western excerpts (Age, $r(102) = -0.007$, $p = 0.94$; MT, $r(102) = 0.012$, $p = 0.90$) or the Chinese excerpts (Age, $r(102) = -0.043$, $p = 0.67$; MT, $r(102) = 0.045$, $p = 0.65$).

Next, we considered differences in the Dimen participants' matching performance across excerpts. Although Dimen matching performance was only slightly above chance for Western excerpts and did not exceed chance for Chinese excerpts, there was substantial variability across excerpts. For example, for one Western excerpt, Dimen participants correctly matched the Arkansas-generated narrative 69% of the time. To consider what might be driving performance differences across excerpts, we considered the cross-cultural similarity between the Arkansas-generated narratives and Dimen-generated narratives that had been provided for each of the tested excerpts by participants in separate previous studies at the same two research sites.[1] That is, if previous Arkansas and Dimen participants had generated similar stories to the same excerpt, we might then expect Dimen participants in this study to be better able to determine the correct Arkansas-generated Consensus Narrative match for that excerpt. To assess this, we calculated the semantic distance between the set of Arkansas-generated and Dimen-generated narratives for each excerpt in the matching task using the cosine similarity with the term frequency–inverse document frequency (TF-IDF) vectors method described in Sitikhu, Pahi, Thapa, and Shakya (2019). Consistent with the hypothesis that more cross-cultural semantic similarity between Arkansas- and Dimen-generated narratives should lead to better matching performance for the Dimen-participants on the Arkansas-generated narratives, results revealed that Western excerpts, which yielded overall better matching performance for Dimen participants than the Chinese excerpts, show greater cross-cultural similarity (less semantic distance) in the generated narratives than the Chinese excerpts, $t(10) = 3.20$, $p < 0.01$. Thus, although the set of excerpts for these analyses is relatively small, the fact that across-excerpt variability in Dimen performance is to some extent predictable from the cross-cultural semantic similarity of the Dimen- and Arkansas-generated

[1] The study that produced the Dimen-generated narratives took place later and did not permit analysis and use within Experiment 1.

narratives supports the conclusion that the overall poor matching performance does not arise from a broader inability of Dimen participants to complete a narrative matching task, but rather from real cross-cultural differences in the imagined stories.

Finally, to ensure that the Dimen participants' overall poor matching performance was not due to a general difficulty with matching tasks, two control tasks were administered that asked Dimen participants to make other types of matching judgments about the same excerpts. The format of the control tasks was identical to the narrative-matching task, except that instead of selecting the story description that best matched the musical excerpt, participants instead selected the acoustic description that best matched the excerpt (loud vs. soft or fast vs. slow). After hearing short clips extracted from each excerpt, they determined which of two acoustic descriptors best matched the clip (fast vs. slow or loud vs. soft). The clips were selected so that on half of the trials one of the acoustic descriptors (e.g., fast) was the correct descriptor and the opposite descriptor (e.g., slow) was the incorrect one. On the other half of the trials the correct and incorrect descriptors were reversed (i.e., slow was correct and fast was incorrect). The loud vs. soft control task was constructed in the same way. Performance was close to ceiling for both control tasks (loudness, $M = 0.89$, 95% CI = 0.83–0.94; tempo, $M = 0.93$, 95% CI = 0.87–0.98). Thus, the inability of Dimen participants to match stories generated by individuals from Arkansas to the corresponding musical excerpt does not appear to be due to a general problem with matching, but rather is consistent with the absence of linkages between excerpts and stories that were, in contrast, quite obvious to the participants in Arkansas.

3. Experiment 2

Experiment 1 showed that participants living in a particular geographic region (Arkansas), with broadly shared patterns of everyday experience (as members of a public university community in suburban middle America), perceive specific shared stories to individual musical excerpts. However, these stories do not extend to the same degree to participants from a different geographical region (Dimen) with different patterns of everyday experience (as residents of a rural village in China). To assess whether the concrete story imaginings depend on a set of experiences that are restricted to a specific community within a specific geographical locale, or whether they depend instead on a broader set of experiences that might characterize a similar community (another public university in suburban middle America) in a different geographic locale, a second experiment was conducted in East Lansing, Michigan.

In Experiment 2, participants in Michigan either participated in a replication of the study from Experiment 1 using the same set of Arkansas-generated narratives, or using a set of Dimen-generated narratives. If the ability to match stories to the corresponding excerpt depends on experiences specific to a particular location, then individuals from Michigan should perform poorly on the matching task for both the Arkansas-generated and the Dimen-generated narratives. If, on the other hand, matching performance depends on more broadly shared within-culture experiences, then individuals from Michigan should perform at a similar level to the Arkansas participants for the Arkansas-generated narratives, but significantly worse for the Dimen-generated narratives.

3.1. Methods

3.1.1. Participants and design

Two-hundred seventy-one individuals ($n = 204$, female), ages 18–59 years ($M = 19.0$, $SD = 2.3$) completed the experiment in return for course credit in an undergraduate psychology course at Michigan State University. All participants were native speakers of American English. Approximately 38% of participants reported no formal music training; the remaining 62% of participants had between 1 and 14 years of music training ($M = 5.8$, $SD = 3.5$). Training for these participants most frequently consisted of choir or band classes at school. English language and Chinese language media exposure was similar to the Arkansas sample in Experiment 1. The experiment implemented a 2 (Culture of Story: Arkansas vs. Dimen) x 2 (Culture of Excerpt: Western vs. Chinese) mixed-factorial design. Culture of Story was a between-subjects variable ($n = 102$, Arkansas-generated narratives; $n = 169$, Dimen-generated narratives), while Culture of Excerpt was a within-subjects variable. Sample sizes in each condition were sufficiently large to detect a small-to-medium effect with power = 0.80.

As an additional element of the design, we varied the presentation format of the Arkansas-generated narratives to investigate whether the poor matching results of the Dimen participants in Experiment 1 was not somehow due to the different (audio) presentation format for the narratives compared to the Arkansas participants, who viewed written versions of the narratives. To consider this possibility, Michigan participants matching the Arkansas-generated narratives were divided into two subsets. Fifty-three participants viewed written versions of the Arkansas-generated narratives (matching the presentation format of the narratives for the Arkansas participants in Experiment 1), while forty-nine of the Michigan participants heard audio versions of the Arkansas-generated narratives (matching the presentation format of the narratives for the Dimen participants in Experiment 1).

3.1.2. Materials

Stimuli consisted of twelve one-minute excerpts (6 Western, 6 Chinese) as in Experiment 1. Correct Consensus Narratives for the Dimen-generated narrative condition were selected using the same procedure as Experiment 1 from the set of narratives previously generated by the group of participants in Dimen reported in Margulis et al. (2019). The correct Consensus Narratives for the Arkansas-generated narrative condition were the same as in Experiment 1.

3.1.3. Procedure

The same 2AFC task and general procedure was used in Experiment 2, except that the Consensus and Foil Narratives used in the experiment had been generated *either* from responses by Arkansas participants (Arkansas-generated narrative condition) or by Dimen participants in a previous study (Dimen-generated narrative condition). Participants were only enrolled in one condition, so all Consensus and Foil Narratives for an individual participant were either Arkansas-generated or Dimen-generated. The experiment was administered using Eprime 2.0 with participants making responses using the computer keyboard and with musical excerpts presented at a comfortable listening level over headphones. Participants were tested in small groups of up to 8 participants, each seated at a separate Dell PC computer. The entire experiment lasted ~20 min.

3.2. Results and discussion

An initial comparison of matching performance for participants who heard audio recordings of the narratives compared to those who viewed written descriptions of the narratives revealed no difference in performance between the two conditions (audio: $M = 0.87$, $SD = 0.12$; written: $M = 0.88$, $SD = 0.11$), $t(100) = 0.49$, $p = 0.63$; Cohen's $d = 0.09$). Because mode of presentation had little-to-no impact on matching performance, all subsequent analyses reported below were collapsed over mode of presentation.

Fig. 3 shows a box-and-whisker plot of the Michigan participants' proportion correct matching performance for Western and Chinese excerpts in response to the Arkansas-generated narratives (left set of boxes) and to the Dimen-generated narratives (right set of boxes). A 2 (Culture of Narrative) x 2 (Culture of Excerpt) mixed-measures ANOVA on proportion correct matching scores revealed a main effect of Culture of Narrative, $F(1, 269) = 57.96$, $p < 0.001$, $\eta^2_p = 0.18$, and a main effect of Culture of Excerpt, $F(1, 269) = 22.49$, $p < 0.001$, $\eta^2_p = 0.08$, but no interaction, $F(1,269) = 1.07$, $p = 0.14$, $\eta^2_p = 0.008$. Michigan participants' matching performance for the Arkansas-generated narratives was

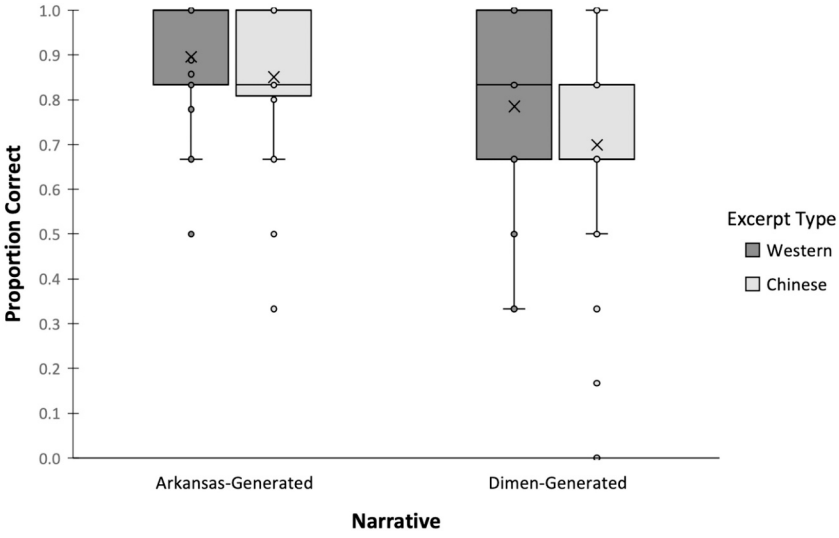

Fig. 3. Box-and-whisker plot showing story-to-excerpt matching performance for Michigan participants for the Arkansas-generated narratives (left pair of boxes) and the Dimen-generated narratives (right pair of boxes) for the Western musical excerpts (dark gray) and the Chinese musical excerpts (light gray) in Experiment 2. Chance performance corresponds to PC = 0.5. Paralleling the pattern observed in Experiment 1, Michigan participants' matching performance for the Arkansas-generated narratives was close to ceiling and no different from the Arkansas participants in Experiment 1. Matching performance for the Dimen-generated narratives, in contrast, was significantly worse for both the Western and Chinese excerpts.

better than for the Dimen-generated narratives (Arkansas-generated, M = 0.87, 95% CI = 0.85–0.90; Dimen-generated, $M = 0.74$, 95% $CI = 0.72$–0.76). Similar to Experiment 1, matching performance was also better for the Western excerpts ($M = 0.84$, 95% $CI = 0.82$–0.86) than for the Chinese excerpts ($M = 0.78$, 95% $CI = 0.75$–0.80) for both the Arkansas and Dimen-generated narratives, possibly due to greater familiarity with the musical style.

A direct comparison to the Arkansas participants in Experiment 1 reveals there is no difference in the matching performance for the Arkansas and Michigan participants, $t(210) = 0.26$, $p = 0.80$, even though the original stories had been generated exclusively by participants in Arkansas. Moreover, the by-excerpt correlation in matching performance across the two locations was highly positive, $r(10) = 0.84$, $p < 0.001$. Thus, the practically identical level of matching performance for the two separate locations and the highly positive per-excerpt matching correlation between locations reveals a remarkably high degree of within-culture agreement about the specific stories musical excerpts tell, extending to two different research sites in middle-America university towns.

On the other hand, the weaker matching performance of the Michigan participants for the Dimen-generated narratives is indicative of an important between-culture difference. Michigan participants had more difficulty identifying Dimen-generated stories in response to Western and Chinese musical excerpts, echoing the difficulty Dimen participants had with identifying Arkansas-generated stories in response to Western and Chinese musical excerpts. However, although performance was worse for both cross-cultural matching tasks, Michigan participants were comparatively more successful at matching Dimen-generated narratives (Experiment 2) than Dimen participants were at matching Arkansas-generated narratives (Experiment 1). We consider possible reasons for this performance asymmetry in the General Discussion.

4. General discussion

Individuals from the US and China listened to wordless musical excerpts followed by two short narrative descriptions and were asked to identify which had been imagined by another listener in response to the music. Experiment 1 revealed that participants at the same research site where the stories had previously been generated (Arkansas) were resoundingly able to match the correct story to the corresponding musical excerpt. In contrast, individuals from Dimen, a remote rural village in China, performed at chance or slightly above chance levels on the same task using the same excerpts.

The poor matching performance of the Dimen participants for the Arkansas-generated narratives is unlikely to be due to a general problem with matching tasks, as Dimen participants had no difficulty performing identically-formatted control tasks using the same excerpts. Moreover, it is unlikely to be due to a specific problem matching narratives to music (or a general inability to narrativize to music) since 1) the Dong community's most distinctive genre of music revolves around storytelling (Ingram, 2007) and 2) a separate study with the same participants revealed high scores on the Narrative Engagement Scale—a measure of how narratively engaged people are while listening to an individual excerpt—for similar excerpts of Western and Chinese instrumental music (Margulis et al., 2019). Finally, it is unlikely to be attributable to a difference in naturalness among the translated versions of the narratives, since the team of native Dong speakers who served as translators evaluated them for linguistic naturalness and identified no differences, and because the same stories that served as the correct matches for one excerpt served as the foil story for another. Rather, what seems most likely as the dominant factor driving Dimen participants' poor matching performance is that the specific story-to-excerpt associations of Dimen participants differed from those in the minds of Arkansas listeners.

Experiment 2 revealed that individuals from Michigan, geographically removed from Arkansas, were similarly able to identify the correct matching stories for the Arkansas-generated narratives; indeed, there was no difference in matching performance between the Arkansas and Michigan participants on the Arkansas-generated narratives. The Michigan participants, however, performed more poorly (though still better than chance) on matching the Dimen-generated narratives than they did on matching the Arkansas-generated narratives. Thus, although the ability to identify specific story-excerpt associations extends very well across research sites where participants share a macroculture (Michigan participants matching Arkansas-generated narratives), it extends only to a lesser degree across research sites where participants share fewer cultural experiences (Michigan participants matching Dimen-generated narratives and Dimen participants matching Arkansas-generated narratives). The fact that Michigan participants were still able to match Dimen-generated narratives, although not as well as the Arkansas-generated narratives, highlights that some associations convey cross-culturally, necessitating future work that can identify the mechanisms for such transfer and the kinds of experiences and exposures on which they depend. For example, future studies could strive to generate novel media-driven associations via controlled exposure to specially designed video stimulus sets, and assess the degree to which specific exposures drive narrative response to subsequently presented musical excerpts.

With respect to the cross-cultural matching conditions, an important

outstanding question is why Michigan participants were better able to match Dimen-generated narratives than Dimen participants were able to match Arkansas-generated narratives. As already detailed, it's unlikely that (1) Dimen participants had general problems with the matching task itself, (2) Dimen participants had specific problems forming music-narrative associations, or (3) there were issues with the translation. There are at least two possible additional explanations for this asymmetry. First, issues surrounding data collection in the field within a non-literate community almost certainly added noise to the results collected in Dimen, which would have lowered their performance relative to the Michigan sample. Second, differences in the semantic similarity in the sets of Arkansas-generated and Dimen-generated Consensus Narratives could have contributed to a performance asymmetry. Presumably, greater similarity among the stories forming the Consensus Narratives set makes it more difficult to discriminate between the correct Consensus Narrative and the paired Foil (since the Foils are drawn from the same set of Consensus Narratives). To consider this possibility, we had naïve raters judge the story similarity of all consensus/foil story pairs for the Arkansas-generated narratives and the Dimen-generated narratives. We found that the consensus Arkansas-generated narratives had greater overall story similarity within the set than the Dimen-generated narratives, $t(22) = 2.52$, $p = 0.02$; thus, with respect to the cross-cultural comparison, it should be somewhat easier for US participants to discriminate between the less internally-similar set of Dimen-generated narratives than for Dimen participants to discriminate between the more internally-similar set of Arkansas-generated narratives: exactly what we observed.

More broadly, following research linking short musical passages to specific referents (Koelsch et al., 2004; Trainor & Trehub, 1992), this research suggests that longer musical excerpts can robustly cue stories for listeners with shared backgrounds. This finding is inconsistent with an absolutist view of music, prevalent overtly in many theoretical perspectives on art music (see Scruton, 1997 for an overview), and tacitly in many empirical ones (see Besson & Friderici, 1998). It suggests that instrumental music carries robust associations with significant inter-subjective agreement, constituting a previously under-acknowledged (and often vigorously argued against) semantic dimension to music. Future work should investigate how these narrative links form and what acoustic features drive them. Research on other aspects of music perception should consider narrative as a mediating factor. Excerpts that trigger robust stories may be remembered better (Delis, Fleer, & Kerr, 1978), and both perceptions of tension and relaxation across musical excerpts (Steinbeis & Koelsch, 2008) as well as music's emotional and expressive effects (Gabrielsson, 2015) may connect to the dramatic events imagined to unfold in the music's story.

The mechanisms behind these narrativizations seem to include topicality, whereby individual patterns of features come to connote a shared extramusical referent (Mirka, 2014), as well as sonic analogues of dynamic processes (Zbikowski, 2017), whereby dynamic event changes in the sound get mapped onto dynamic changes in the event structure of the perceived story. As an example of topicality, a cluster of rumbling in the low strings might have come to connote a shark following the 1975 movie *Jaws*. As an example of sonic analogues of dynamic processes, an alternation back and forth between this kind of rumbling and a higher passage might translate into a narrative of escalating attacks and narrow escapes. An important agenda for future and ongoing work is to manipulate acoustic features and examine changes in the resultant narratives.

In addition to acoustic features of the music itself, the previous experiences of the listeners seem to be playing an important role in narrative generation as well. What precisely constitutes the shared background that allows listeners in the relatively distant locales of Arkansas and Michigan to identify with ease the stories that other individuals imagine in response to wordless musical excerpts? One possibility is that listeners abstract a set of conventions surrounding music's narrative functions from its use in mass media (Tagg & Clarida, 2003)—deriving the topicality of particular sound patterns through repeated exposure. Some support for this hypothesis is found in the analysis of individual differences in Dimen participants' matching performance reported in Experiment 1. Even though the exposure of Dimen participants to Western media was minimal, Dimen participants with some Western media exposure performed better on the matching task than Dimen participants with no Western media exposure. As a further test of this hypothesis, we assessed matching performance on Arkansas-generated narratives for a fourth sample of participants from Hong Kong who had substantial Western media exposure, unlike the participants in Dimen. Hong Kong participants showed similar matching performance to Arkansas and Michigan participants, performing the matching task at a high level for both the Western excerpts ($M = 0.89$, 95% $CI = 0.83$–0.95) and Chinese excerpts ($M = 0.81$, 95% $CI = 0.71$–0.90). Thus, shared cultural experiences in the form of Western media exposure seems to allow geographically dispersed participants in Arkansas, Michigan, and Hong Kong to predict the stories participants in Arkansas imagined in response to wordless music with a high degree of accuracy and also confers some advantage to participants in Dimen who, in general, had minimal Western media exposure.

One additional finding, consistent across both experiments, is that cross-cultural matching performance was better for the Western than the Chinese excerpts—that is, Dimen participants matched the Arkansas-generated narratives better for Western than Chinese musical excerpts, and Michigan participants matched the Dimen-generated narratives better for Western than Chinese musical excerpts. This likely reflects an asymmetry in exposure patterns to Western and Chinese music. Due to the globalized influence of Western media, Dimen participants had some minimal exposure to the use of Western music in Western media; thus, they were able to pick up faintly on some of the conventional associations that guided Arkansas responses. It is a different picture, however, for Chinese music. Arkansas participants had hardly any exposure to the use of Chinese music in the Chinese media and cultural contexts that would shape Dimen participants' narratives for these excerpts. Instead, Arkansas responses were conditioned by the use of Chinese music in *Western* media and cultural contexts. Accordingly, when Dimen participants were presented with Chinese excerpts, they were drawing on a completely different pool of exposures, resulting in almost no ability to match the Arkansas associations, whereas when they were presented with Western excerpts, that had *some* shared pool of experience to draw in—even if small—due to the globalized influence of Western media.

Taken together, these experiments provide evidence for narrative perceptions in a non-linguistic auditory domain, add to the growing evidence for a semantic dimension to music in line with the referentialist perspective, and highlight the role of shared cultural experiences in the construction of these tacit but widely understood meanings. Without the aid of any accompanying lyrics or text, strings of sound argued by some historical and contemporary thinkers to be inherently abstract in fact seemed to clearly convey specific narratives. This finding constitutes perhaps the clearest evidence yet for a referentialist versus absolutist position on musical meaning, opening new lines of research on musical semantics, and warning that psychological studies which position music as abstract and devoid of meaning need to reevaluate those assumptions. These imagined stories tie together diverse listeners, such that they formulate similar envisaged stories while listening to sequences of supposedly semantics-free sound. They also, however, have the potential to divide listeners with different prior sets of experiences, underscoring the importance of tacit and infrequently acknowledged semantic resonances in defining cultures. Music's power cannot be fully understood if we continue to assume that it operates independently from the shared stories it generates in listeners' minds.

Acknowledgements

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. Rhimmon Simchy-Gross, Lauren Shepherd and Lucas Bellaiche helped with data collection in Arkansas and Jewelian Fairchild and Gabby Kindig helped with data collection in Michigan. Benjamin Kubit and Cara Turnbull helped with the semantic similarity measures. Special thanks to Mr. LEE Wai Kit and the staff at the Dimen Dong Eco-Museum for making data collection possible, and to the people in Dimen who participated in this research. This research was supported by the Division of Behavioral and Cognitive Sciences of the National Science Foundation, Award Numbers 1734063 (PI: JDM) and 1734025 (PI: EHM).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104712.

References

Abbott, H. P. (2008). *The Cambridge introduction to narrative*. Cambridge, UK: Cambridge University Press.

Besson, M., & Friderici, A. D. (1998). Language and music: A comparative overview. *Music Perception, 16*, 1–9.

Cross, I. (2012). Music as a social and cognitive process. In P. Rebuschat, M. Rorhrmeier, J. A. Hawkins, & I. Cross (Eds.), *Language and music as cognitive systems* (pp. 315–328). Oxford, UK: Oxford University Press.

Davies, S. (1994). *Musical meaning and expression*. Ithaca, NY: Cornell University Press.

Delis, D., Fleer, J., & Kerr, N. H. (1978). Memory for music. *Perception & Psychophysics, 23*, 215–218.

Gabrielsson, A. (2015). The relationship between perceived structure and musical expression. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 215–232). Oxford, UK: Oxford University Press.

HaCohen, R., & Wagner, N. (1997). The communicative force of Wagner's leitmotifs: Complementary relationships between their connotations and denotations. *Music Perception, 14*, 445–476.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*, 243–259.

Huovinen, E., & Kaila, A. K. (2015). The semantics of musical topoi: An empirical approach. *Music Perception, 33*, 217–243.

Ingram, C. (2007). If you don't sing, friends will say you are proud': How and why Kam people learn to sing Kam big song. *Journal of Musicological Research, 32*, 85–104.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience, 7*, 302–307.

Margulis, E. H. (2017). An exploratory study of narrative experiences of music. *Music Perception, 35*, 235–248.

Margulis, E. H., Wong, P. C. M., Simchy-Gross, R., & McAuley, J. D. (2019). What the music said: Narrative listening across cultures. *Palgrave Communications, 5*, 146.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Mirka, D. (2014). *The Oxford handbook of topic theory*. Oxford, UK: Oxford University Press.

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage, 184*, 161–170.

Radocy, R. E., & Boyle, J. D. (2012). *Psychological foundations of musical behavior* (5th ed.). Springfield, IL: Charles C. Thomas.

Sarbin, T. R. (1986). *Narrative psychology: The storied nature of human conduct*. Westport, CT: Praeger.

Scruton, R. (1997). *The aesthetics of music*. Oxford, UK: Clarendon.

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A comparison of semantic similarity methods for maximum human interpretability. In *Proceedings of the IEEE international conference on artificial intelligence for transforming business and society*. https://arxiv.org/abs/1910.09129.

Skowron, Z. (2007). *Lutoslawski on music*. Plymouth, UK: Scarecrow Press.

Stravinsky, I. (1936). *Igor Stravinsky: An Autobiography*. London: Simon and Schuster.

Steinbeis, N., & Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cerebral Cortex, 18*, 1169–1178.

Tagg, P., & Clarida, B. (2003). *Ten little title tunes: Towards a musicology of the mass media*. New York and Montreal: The Mass Media Musicologist's Press.

Trainor, L. J., & Trehub, S. E. (1992). The development of referential meaning in music. *Music Perception, 9*, 455–470.

Zbikowski, L. M. (2017). *Foundations of musical grammar*. Oxford, UK: Oxford University Press.