A Perceptual Study of Scale-degree Qualia in Context

Claire Arthur
Ohio State University

A perceptual study investigated the ability of scale degrees to evoke qualia, and the impact of harmonic context in shaping a scale degree's qualia. In addition, the following questions were addressed: What role does music training have in shaping qualia? Are listeners consistent in their descriptions? Are experiences similar across participants, or are they individual and subjective? Listeners with or without music-theoretic training were asked to rate the qualia of scale degrees following various chord progressions, each ending with a different final harmony. Scale degrees were found to exhibit relatively consistent musical qualia; however, the local chord context was found to significantly influence qualia ratings. In general, both groups of listeners were found to be fairly consistent in their ratings of scale-degree qualia; however, as expected, musician listeners were more consistent than nonmusician listeners. Finally, a subset of the musical qualia ratings were compared against Krumhansl and Kessler's (1982) scale-degree "profiles." While profiles created from the present data, overall, were correlated with the K&K profiles, their claim that tonal stability accounts for the high ratings ascribed to tonic triad members was found to be better explained by the effect of the local chord context.

Received: September 24, 2016, accepted June 21, 2017.

Key words: scale degree, phenomenology, key profiles, melodic-harmonic interaction, pitch perception

While there is frequently a shared commonality to many of our musical experiences, those who study music—and in particular those who study music perception—tend to agree that we do not all perceive musical events in the same way. Of course, we can never put ourselves into the mind of another so we can never *truly* know what it would be like to experience some musical moment from another's perspective. Nevertheless, understanding the range of musical experiences that exist, and the factors that might contribute to them remains one of the principal goals of the music cognition researcher. This goal sits at the heart of the present study, which takes an empirical approach to a rather challenging topic: the study of scale-degree *qualia*. Specifically, this paper attempts to answer several questions: 1) How does conceptual knowledge influence a scale degree's qualia?; 2) How different or similar are individuals' experiences of scale-degree qualia?; 3) How do scale-degree identification and qualia interact in "real time," or, more importantly, in real musical contexts, where typically—at least in Western music—a melody is most commonly embedded in a harmonic context? Before embellishing on these questions and the methodology for testing them, some discussion and background on the topic of qualia is warranted.

Qualia is a term borrowed from philosophy that generally refers to what it is *like* to experience something. The term is often used synonymously with "phenomenal character." In philosophy of mind, the concept of qualia is a hotly debated one, and features prominently in issues related to consciousness and the mind-body problem. The most common points of debate are "which mental states have qualia, whether qualia are intrinsic qualities of their bearers, and how qualia relate to the physical world both inside and outside the head" (Tye, 2015).

What qualia are elicited by music? There are many instances where a listener might recognize a distinctive feeling, character, or quality associated with some musical moment. For example, in the striving or effortful intensity of a sung high pitch, or in the paradoxical feel of both repose and impetus evoked by a deceptive cadence. In these and many other musical situations, listeners can experience seemingly ineffable yet characteristic subjective states that that would be appropriate to call "qualia." Yet, considering some of these experiences *as* qualia may raise concern for some philosophers who hold a strict viewpoint on what constitutes qualia.

While there are many claims made about qualia within the field of philosophy, it is commonly argued that qualia are intrinsic, nonrelational properties, and many believe that they are ineffable and nonrepresentational (Tye, 2015), and it is these claims that prove particularly problematic when describing musical qualia.

Music Perception, volume 35, issue 3, pp. 295–314, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2018.35.3.295

The claim of qualia as intrinsic is perhaps the single most difficult problem for musical qualia, in particular for the study of scale-degree qualia. Our experience of scale degree is not like that of color, or even of timbre, in that scale degrees are not "absolute" but in fact *only* exist in relation to something else. Specifically, scale degrees exist only in relation to a tonic, key, or other scale-step. Of course, the notion that scale degrees are capable of eliciting qualia does not appear to be under debate. Indeed, a trained musician can hear a single complex tone (e.g., a single key struck on the piano) and impose upon it the phenomenal experience of *any* scale degree entirely through mental "gymnastics." Importantly, this is not simply an artificial exercise, but occurs during real listening at moments of transposition, presumably not only for trained musicians, but for all listeners enculturated into the Western musical tradition. This problem, where seemingly the same physical stimulus can represent two things simultaneously, poses a serious challenge for philosophical theories that hold qualia to be intrinsic.

The problem of ineffability, if true, would certainly pose a problem for anyone wishing to study musical qualia empirically, since experiments typically rely on language as expressed by participants (e.g., "how does this sound?" or "how does this feel?"). Even when examinations of qualia attempt to avoid procedures that "collect" language as a dependent variable, language must nevertheless be used in instructions for experiments that implicitly measure qualia (e.g., "adjust the brightness until the object on the left is twice as bright"). A common argument by philosophers on the notion of qualia's ineffability is that it is considered impossible to describe a phenomenal experience in words to another individual who has *never had* that experience. While I agree that it is likely impossible to communicate the qualia of, for example, scale-degree 7 (or anything for that matter) to someone who has never experienced it, I argue that those of us who *have* experienced hearing scale-degree 7 are able to communicate with each other some aspect of that shared experience.

The concerns raised above are not unique. Despite differing points of view, Goguen (2004), Raffman (1993), Zentner (2012), and Dowling (2010) all discuss the problem with traditional accounts of qualia for empirical musicologists. Goguen, for instance, holds that qualitative experience is "situation dependent," and believes that "emotion . . . is the essence of qualia." Raffman believes that empiricism has much to contribute to aesthetics, and in particular criticizes the view of qualia as nonrepresentational, saying "some sensory states have . . . legitimate representational contents" (although she also believes these are "consciously accessible but not reportable"). Zentner argues that "contrary to the claim that musical experiences are ineffable . . . musical qualia may be amenable to linguistic description and objectification." Dowling, too, dismisses the stipulations of ineffability ("I tend to think something truly ineffable would resist being manipulated"), as well as determinacy ("there aren't any incorrigible facts"), and intrinsic properties ("there is almost no area of human perception which is not context dependent").

Like the authors above, I oppose various stipulations on the traditional definition of qualia, yet propose that it remains appropriate to use the term since, as Dowling (2010) states:

> This use of the term clearly is not consonant with some of the ways it has been used in the history of philosophy, but nevertheless it is difficult to see what term could better be used to point to the functions described.

In terms of an operational definition of qualia, mine can be taken as largely synonymous with Dowling's, who proposes qualia as so-called "intervening variables," or "inferred processes in the causal chain leading from stimulus to response." Specifically, I propose that while qualia are the resulting subjective experience of something, there are a multitude of factors that can *contribute to* that phenomenal aspect of musical experience, including sensory information (bottom-up information); conscious and unconscious (implicit) knowledge, memory, and awareness (top-down information); and the resulting inference or interpretation (including possible accompanying bodily changes, such as heartrate) that are a result of the synthesis of this sensory and cognitive processing. Furthermore, I informally define qualia to be: aspects of conscious experience, available via introspection, fleeting or temporary, extrinsic (relational), at least partially communicable, and mediated both by context and conceptual knowledge (explicit or implicit). In addition, while I regard qualia as necessarily subjective, the fact that persons can have similar experiences in response to the same stimulus suggests there may be (theoretically) measurable features of the stimulus that are seemingly "absorbed" by our senses and that may lead to these common, or overlapping, aspects of perception.

As mentioned above, qualia are, of course, subjective. Yet, it seems feasible that a majority of individuals might describe the qualia of a sunset, or of eating a pear, in similar ways. Perhaps the qualia of scale degrees, then, might similarly be described using common language across individuals? What it is like to undergo some

phenomenal experience is available to us by introspection (Tye, 2015), and I argue that we can study qualia by examining common language and observing common reactions to stimuli from multiple persons' experiences and introspections. In particular, I agree with Zentner (2012), who argues that "although the inner experience of an emotion is a private and subjective one, its expressions are amenable to scientific description, quantification, and analysis"; and Goguen (2004), who argues that "cognitive and qualitative aspects of experience are inseparable, even though first and third person approaches artificially separate them." In particular, as already mentioned, the converging evidence provided by similar descriptions independently offered by multiple individuals' reaction to the same stimulus surely points to *some* common component of an experience. This point of view is similarly shared by Zentner, who—talking about a similar problem in emotion research—says:

> Although we cannot know what people's inner experience . . . might feel like, the assumption is that the similarity in emotion expression is subtended by an interpersonally similar inner experience of the emotion. In other words, the similarity and dissimilarity of inner subjective experiences across individuals, while not directly accessible, can nonetheless be inferred.

In this way, I proceed by assuming that *some* aspects of musical qualia can be obtained indirectly using this "converging evidence" methodology, even if the resulting descriptions might be crude and/or incomplete, and that musical qualia are amenable to scientific analysis.

Theories of Musical Qualia

Where do musical qualia come from? What gives each note or passage a unique characteristic, or quality? For instance, what makes the scale degrees in each measure of Figure 1 "feel" different?

In Figure 1, the acoustic information in each measure is identical. Only the musical context—in this case, key—has changed. In order to perceive these three pairs of scale degrees as having different qualia, then, one must be able to imagine them within their appropriate positions within the given key (or scale). But what would cause one note within a scale to sound different from any other in the first place?

Huron (2006) argues that scale-degree qualia arise—at least in part—from statistical learning. As applied to melody, the theory of implicit learning assumes that through exposure we come to internalize the statistical

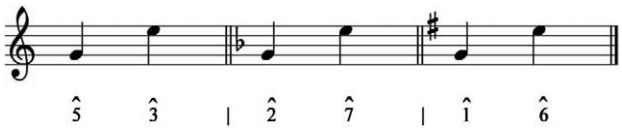

FIGURE 1. Three intervals are presented that are identical in pitch, but composed of different scale-degree pairs. Despite having the same acoustic information, each is capable of generating distinct qualia.

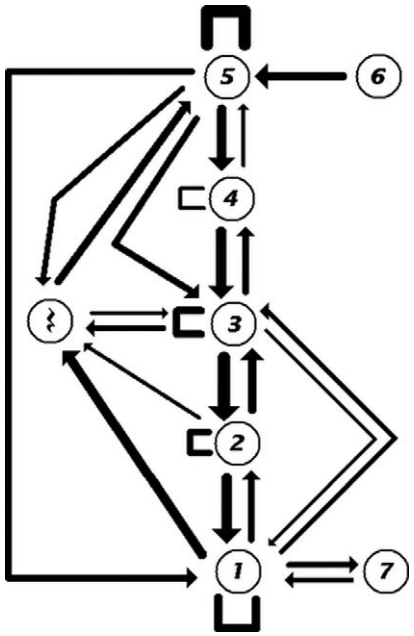

FIGURE 2. Flow chart of first-order diatonic scale-degree probabilities from Huron (2006). The strength of a scale degree's tendency is marked by the arrow's width. (Reprinted with permission from MIT Press.)

probabilities for where a given scale-degree will go next. In tonal music, of course, the progression of scale degrees is nonrandom, with scale degrees exhibiting a range of probable (or improbable) "behaviors," or tendencies to proceed in a predictable way. For instance, $\hat{7}$ tends to move to $\hat{1}$, and $\hat{4}$ tends to move to $\hat{3}$ (see Figure 2).

Huron argues that because the brain ceaselessly attempts to predict what will happen next, our sense of anticipation and the unconscious confirmation or denial of the arrival of some scale degree are, in part, what lead to scale-degree qualia. In other words, because of statistical regularities, the qualia of an individual scale degree become closely tied to feelings of tension or resolution that have become associated with it. Of course, the entire concept of scale degree requires a connection not only to the other scale degrees and where they might proceed, but to a particular position within the scale. Browne (1981) theorized about what has come to be known as

the "rare interval hypothesis." Browne proposed that it is the unique intervallic properties of the diatonic collection that allow for "position finding," or a sense of maintaining one's bearings with regard to a particular position within the scale or reference to a tonic. Specifically, the subset of all possible intervallic (dyadic) possibilities within the diatonic collection can be tabled into a set of interval-classes, where each interval class appears a unique number of times. Thus, the rare intervals—the tritone and semitones—function as the "position finding" elements. This theory was tested empirically by Butler and Brown (1981), who found that listeners were best able to correctly infer the tonic from three note subsets when they included the rare interval of the tritone. Shepard (2009) argued that we build an internalized mental representation of the pattern of the scale, which is what leads us not only to understand the music we hear, but also to the generation of qualia. Note that Shepard's position also implies that simple exposure (implicit learning) is at work in the generation of these internalized representations.

Huron (2006) describes an informal study in which he asked ten experienced musicians to imagine each scale degree, and to "free associate" words or phrases that they felt described that particular scale degree. He then analyzed the responses by grouping words and phrases that were alike in meaning, and found that there was a clustering of similar responses according to scale degree. This suggests that experienced musicians not only have the ability to hear qualia, but also that their personal experiences of the qualia of scale degree appear to be similar. However, Huron's study used a limited number of participants, all of whom were experienced musicians (music faculty and graduate students in music theory). Furthermore, his participants were introspecting about their memory about a scale degree, rather than immediately experiencing one. Lastly, by imagining a scale degree in isolation, the harmonic implications (if any) for these scale degrees are ambiguous, but likely represent the most probable harmonic association. For instance, someone imagining scale-degree 3 is likely to imagine it as a member of the tonic triad (as opposed to a member of the mediant or submediant triads). As will be discussed below, the present study builds on Huron's design, but attempts to address these concerns in order to further inquire into the processes involved in the acquisition and experience of scale-degree qualia.

Aims

If musical "objects" such as scale degrees evoke seemingly unique qualia that are largely brought about through implicit learning, do their qualia remain stable across different contexts, or would the statistical frequency of the context itself bear on the qualia of a scale degree? For instance, the quintessential scale-degree qualia are perhaps those evoked by scale-degree 7, which typically are characterized with terms such as "leading," "leaning," "pulling," and "restless" (Huron, 2006). However, scale-degree 7 is unique in that, within Western classical music, it is the only scale degree that is so strongly associated with chords of dominant function, which inherently carry strong tendency to resolve to tonic. Interestingly, however, if the so-called "leading tone" (i.e., $\hat{7}$) is placed in a mediant context, or, more rarely, in a tonic context, it appears to lose those leading-type qualia, and, in fact, in those contexts (iii and I7) scale-degree 7 tends to resolve downwards. Thus, the principle question addressed in this paper is how the role of harmonic context might shape certain aspects of scale-degree qualia. Specifically, this paper will evaluate a formal hypothesis: that changes to the immediate harmonic context will modify the perceived qualia of scale degrees.

When teaching scale-degree (or interval) identification in music, it seems natural to point out what are assumed to be some *shared* attributes of those experiences. For instance, one might discuss the "leading" or "leaning" quality of scale-degree 7, or the "clashing" of the dissonance in a major seventh. Identifying scale degrees, of course, comes more easily for some than for others. Perhaps students that have difficulties with scale-degree identification tend to confuse the qualia of scale degrees? Or perhaps scale degrees with similar functions (e.g., $\sharp\hat{4}$ and $\hat{7}$) tend to elicit similar qualia? However, there is a possibility that these labels that we attach to certain scale degrees do not arise organically from experience, but rather—perhaps from generations of teachers passing down these learned vocabularies—scale degrees become imbued with the representations we attach to them. Perhaps the qualia are not useful for identification, and instead we come to identify scale degrees by other means and then, once identified, gain access to all the associated features we have learned.

In addition to the primary research question noted above, I aim to investigate the following exploratory questions: First, is everyone capable of experiencing scale-degree qualia? It may be that certain listeners are incapable of hearing in this way. For those that can distinguish the qualia of scale degrees, do those listeners do so in consistently similar ways? And what about the role of learning and experience? Are the participants in Huron's study merely responding according to learned associations? Can individuals without any music-theoretic training distinguish the qualia of scale degrees

in consistent ways? And if so, are their responses similar to those who do have music training? Finally, do scale-degree qualia remain stable within real musical contexts, where typically—at least in Western music—a melody is most commonly embedded in a harmonic context?

Finally, this paper presents a partial replication of a well-known experiment conducted by Krumhansl and Kessler (1982). In a series of well-known experiments, Carol Krumhansl, along with Edward Kessler and Roger Shepard, tested participants' responses to perceived "goodness of fit" of a scale degree after a given harmonic progression in an attempt to investigate the properties of scale degrees as they relate to the overall key (Krumhansl, 1990; Krumhansl & Kessler, 1982; Krumhansl & Shepard, 1979). (This work directly led to the use of the famous Krumhansl and Kessler "key profiles" that are still widely in use today.) Krumhansl and Kessler claim that their goodness of fit ratings "confirm" a tonal hierarchy theory, in which they propose that what listeners are really responding to are the varying levels of tonal "stability" of the scale degrees within a given key, which fit into three hierarchical categories: tonic chord members, remaining diatonic tones, followed by remaining chromatic tones. However, the "various harmonic contexts" in the experiments that led to the key-profiles were all variations of a predominant-dominant-tonic progression, and therefore all ended with the tonic chord. Therefore, in order to ensure that Krumhansl and Kessler's key profiles truly represent effects of *key*, it would be prudent to examine "goodness of fit" ratings using a greater variety of chord progressions, and specifically, using chord progressions that end on a chord other than tonic.[1] Accordingly, this paper will also examine ratings related to "goodness of fit" in various harmonic contexts, and compare them to Krumhansl and Kessler's findings.

In sum, this paper ultimately has three goals. The first is to reinvestigate Huron's (2006) findings that musicians generally agree on qualitative terms for the various scale degrees, and then to explore whether listeners without music training describe the qualia of scale degrees using similar terms, or whether they can do the task at all, given that they cannot identify scale degrees. If descriptions are relatively consistent across all levels of experience, it would suggest that scale-degree qualia are not dependent on training. The second goal is to test the role of harmony in the perception of scale-degree qualia. Huron (2006) suggested that scale-degree qualia arise from statistical learning. Arthur (2017) found that changes in the harmonic accompaniment affected scale-degree transition probabilities, which, according to Huron's hypothesis, implies that a scale degree's qualia would be altered in different harmonic contexts. Said another way, if scale-degree qualia *are* found to be systematically altered by the harmonic context, then this finding would support Huron's hypothesis regarding the statistical learning of qualia. In addition, if scale-degree qualia are affected by changes to the underlying harmonic accompaniment, this might suggest a revision to current models of aural skills pedagogy, where melodic and harmonic activities are commonly segregated. Lastly, a final goal is to compare "goodness of fit" judgments for scale degrees in different harmonic contexts from the present study with the findings from Krumhansl and Kessler in order to re-examine the relation of scale degree to chord, and chord to key.

Method

Prior to the main experiment, an informal pilot study was conducted with 10 participants of mixed musical backgrounds. Some were undergraduate students pursuing a music degree, and others were persons with little to no musical experience and no training in music theory or aural skills. The primary purpose of the pilot study was to establish the descriptive terms that should be used in the main experiment, and to evaluate whether individuals without musical experience would be able to perform the task. Therefore, an approach similar to that of Huron (2006) was taken, where participants were asked to free-associate words with various scale degrees. In contrast to Huron's approach, however, the scale degrees were actually heard (as opposed to imagined) with key contexts first established by performing simple scales or short chord progressions at the keyboard. After setting up a key context, a single scale degree was played, after which the participant would respond with their free-associated terms that were then written down by the experimenter.

Given the goals of the present research, and that perceived scale-degree qualia may be a highly individual experience, an appropriate design for the main experiment might take the large list of adjectives given by these participants and ask listeners in the main experiment to check all that apply. However, since the task is rather abstract (according to feedback from the participants in the pilot study), responses are likely to contain a large amount of variation, and the more terms given, the more variation one is likely to find. Although there

[1] As is further discussed at the end of the results section, Krumhansl and Kessler's 1982 paper, in fact, did contain further experiments that used varied chord progressions. However, their original "key profiles" were not altered by these subsequent experiments.

may be subtle differences between words like "jarring" and "harsh," or "gloomy" and "sad," a primary goal of this research is to establish whether there are *general* similarities in descriptions across participants, especially given differences in music training. Therefore, the participants' responses from the pilot study were subjected to content analysis, in an attempt to come up with a representative set of descriptive words that would be able to be rated by participants in the main experiment.

Thus, the complete set of vocabulary obtained in the pilot study were subjected to categorization according to whether they could be considered synonyms (or at least similar in meaning) or antonyms, and then those categories were given names. For example, the following words were grouped into a category designated as relating to the concepts of "strength and/or stability": confident, centered, weak, heavy, strong, stable, light, unsure, unsteady, airy, tipsy, fragile, unstable. Remarkably, despite a few unusual and unexpected responses from the nonmusician group, the results of the pilot study suggest that even when participants have no music training, listeners came up with similar descriptive terms, the majority of which could be put into similar categories as found in Huron (2006). (See Appendix A for the table of unique responses to each scale degree in the pilot study.) The content analysis initially resulted in five categories: movement, strength/stability, emotional valence (primarily happy/sad), lightness/darkness, and tense/relaxed. After comparing these results with Huron's categories (certainty, tendency, completion, mobility, stability, power, and emotional valence), it was acknowledged that the "movement" category held the largest number of terms, and could easily be divided into the subcategories: tendency and completion. Since the majority of descriptions given by participants could fit within these categories, it was decided that these seven category names themselves would be the terms used as dependent variables in the main experiment.

Notice that many of the terms given by the participants in the pilot study, as well as the resulting vocabulary that make up the dependent variables, can be considered metaphorical. Some scholars (especially those who believe in the strict definitions of qualia discussed in the introduction) might consider the reliance on metaphor in describing aspects of music as providing a supporting argument for the claim that musical qualia are ineffable. However, I have already noted that the language gathered and provided in these experiments is in no way meant to *directly* represent their qualia. Rather the vocabulary used (and collected) are thought of as aspects contributing to the overall qualia, and are simply used as a way of measuring converging evidence. In addition, metaphors of motion, energy, brightness or darkness, tension or relaxation have permeated music-pedagogical language for centuries (as evidenced by historical pedagogical treatises,) and obviously carry (indirect) communicative value (Lakoff & Johnson, 2003). They would therefore appear perfectly appropriate to use as descriptive elements of music in these studies. (See Roholt, 2014, and Schiavio & van der Schyff, 2016, for discussions of the importance of metaphor, comparison, and shared experience in music.)

PARTICIPANTS

Sixty-five participants were recruited for this study. Of the total participants, 43 were second-year undergraduate music students, and 22 were graduate and undergraduate students from various other disciplines. All participants were given the Ollen Musical Sophistication Index (OMSI), primarily to distinguish those who had received formal music-theoretic training from those who had not. For the purposes of this study, only two questions from the OMSI questionnaire were considered: one that asks them to self-identify as either a nonmusician or a musician (of varying degrees of competency) and another that asks how many years (if any) of college-level coursework in music they have completed. The main factor for including persons in the "nonmusician" group was that they had not been exposed to the language and terminology commonly used in music theory pedagogy, and that they would not have developed the skill of identifying scale degrees by name. Thus, in addition, participants were informally asked during pre- and post-experiment interviews whether they had ever received any music theory or aural skills training of any kind, and whether they believed they possessed absolute pitch. Based on their responses to the above questions, participants were then divided into two groups: "musicians" and "nonmusicians." Note that a few individuals in the "nonmusician" group did claim to have vocal or instrumental training, but did not consider themselves musicians, nor had they had any experience with music theory or aural training. Of the total participants, five identified themselves as having absolute pitch. These participants' data were first examined in comparison with the musician group, to see if, as a group, their responses were markedly different from the rest of the musician group. *A priori*, it was determined that if the absolute-pitch group responded in a significantly different way, they would be considered separately from the other two groups. If not, their responses would be included in the musician group. In the end, two participants were excluded from the

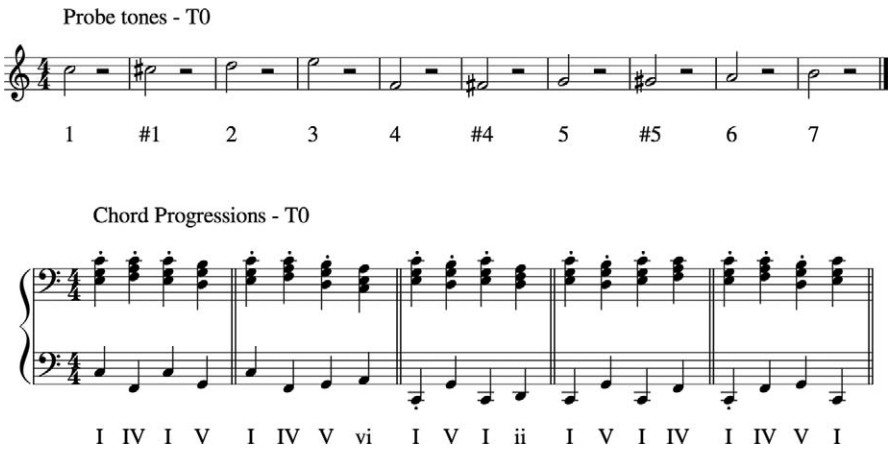

FIGURE 3. Sample experiment stimuli. This figure shows the original transposition level (T0) and voicing for each possible chord progression and probe tone. All probe tones are at least a perfect fourth higher than the upper-most note in the chord progression in order to avoid proximity effects. Each trial consisted of one possible chord progression, followed by a slight pause, followed by one of the possible probe tones. Stimuli were randomly transposed to a new key after every ten trials. The order of stimuli and order of key presentation was randomly assigned.

experiment due to technical malfunctions leading to incomplete data collection. Of the remaining participants, 41 music students fell in the "musician" group and the other 22 students fell in the category of the "nonmusician" group.

STIMULI

The study was broken into two blocks of 35 trials each. The stimuli followed a context+ probe design. In each block, participants were presented with a key-defining progression in a major key. All progressions were four chords in length, began with the tonic triad, implied the same key, but each ended with a different harmony. The possible chord progressions (contexts) consisted of: I–IV–V–I, I–V–I–IV, I–IV–V–vi, I–IV–I–V, or I–V–I–ii. Thus, each progression contained the tonic and dominant chord, but did not necessarily end on the tonic harmony. Each trial consisted of one such key defining progression, followed by a single probe tone, which could be any one of the seven diatonic (major) scale degrees, or one of three chromatic scale degrees: #1, #4, or #5. (The full chromatic scale was reduced to these ten possibilities to limit the combinatorial possibilities and therefore reduce the length of the experiment.) Note that since these chromatic notes are being presented aurally, they may equally be interpreted as their enharmonic equivalents. However, for the sake of clarity these chromatic scale degrees will only be given a single representation (as sharp scale degrees) in all figures and text.

It was decided that the sounds should be as realistic as possible. As such, all stimuli were recorded using a Yamaha P-90 keyboard, Sonar sound editing software, and VST instruments by Roland's *Sonic Cell* (VST used was the "ultimate grand" 003). The stimuli were not quantized, however, they were recorded with the use of a metronome. Although this meant that the inter-onset intervals and coordination of voice onsets within chords would be inexact, it was decided that any effects resulting from these liberties would likely be negligible for this particular experiment; and given the already abstract nature of the task, would in fact be preferred over more highly constrained and less realistic sounding stimuli. However, the velocity of the stimuli was equalized such that all chord progressions and scale degrees would have equal velocity (and so roughly equivalent apparent loudness). Chords were voiced in a traditional "keyboard" spacing, with the left hand playing a single bass note and the right hand playing a close-position triad approximately one octave above the left. See Figure 3 for stimuli examples.

In order to prevent qualia responses from being biased towards describing pitch height, the original stimuli—which were all recorded in the key of C—were randomly transposed to a new key after every 10 trials. This was preferred over changing key after every trial, since it was informally noted in the pilot study that participants (especially those in the nonmusician group) found the task more difficult immediately following a key change. Thus, this design introduced a randomization of keys while still allowing participants enough time in one key to familiarize themselves with it before moving on to a new one. Key presentation order, and the order of stimuli within each key group, were

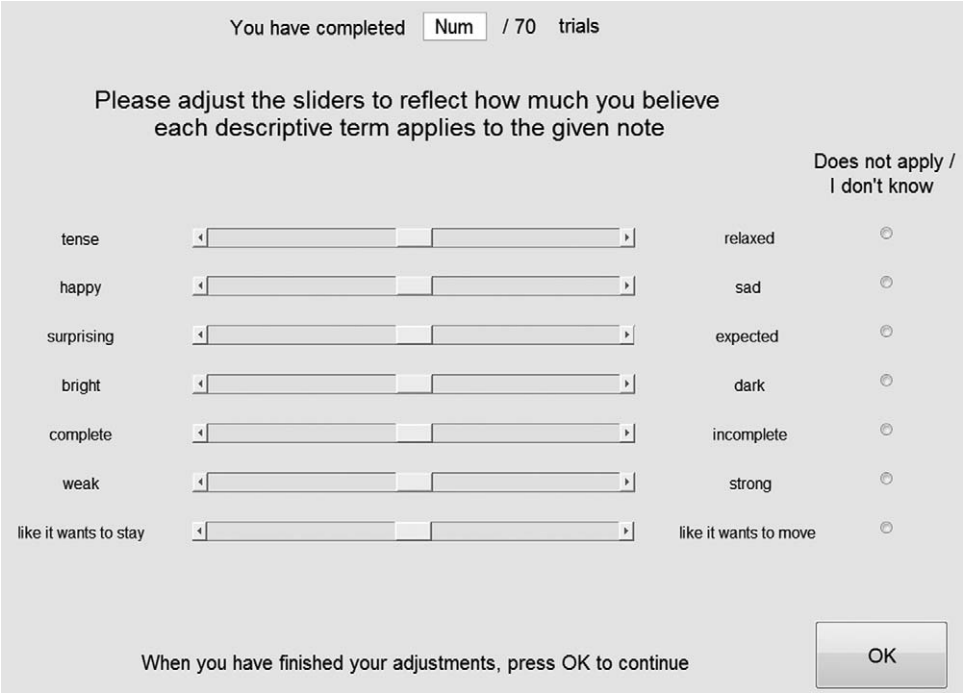

FIGURE 4. Image of digital interface used in experiment. Rating terms are listed in opposite-facing pairs with a slider in-between. Participants were instructed to move the sliders to indicate the degree to which they believed a particular term described the qualia of the probe tone.

randomly assigned for each participant. The scale degrees were all transposed along with the chord progression in order to maintain equivalent spacing from the chord progressions across transposition levels. Scale degrees (probes) were always at least a perfect fourth higher than the upper-most tone of the chord progression, and were presented following a slight pause/silence in order to minimize response biases due to pitch proximity (see Krumhansl & Shepard, 1979, p.359). Participants heard each scale degree (total = 10) in each key context (total = 5) once each for a total of 50 trials, plus an additional 20 trials which were repeated and arranged randomly throughout the two blocks in order to gather a measure of within-subject variability.

PROCEDURE

As explained in the Method section, it was decided that the best method of approach was to use a condensed version of the most commonly used adjectives to describe the qualia of scale degrees, and have participants rate the appropriateness of the terms. Limiting the variables also makes the somewhat daunting task a bit easier for those with no music training. For statistical purposes, a rating task simplifies the analysis by having numerical values on a continuous scale, rather than counts of categorical variables. In addition, using a continuous rating scale rather than yes/no categories allows possible subtle differences in scale degrees to come to the fore that might otherwise be described with similar terms. For example, using a checkbox approach, one might find that $\hat{1}$ and $\hat{3}$ both tend to be described as "stable," "relaxed," "solid," etc. However, by using a continuous rating scale it becomes possible to discern whether, for example, $\hat{1}$ might be *more* "stable," "relaxed," "solid," etc., than $\hat{3}$.

Participants were asked to rate the qualia of each scale degree using a digital interface with sliders that could be dragged to one side or the other indicating that a given scale degree was heard as either more x or more y, where x and y represent terms of opposite qualia (e.g., happy or sad). As can be seen in Figure 4, the interface contained 14 terms, arranged in opposite-facing pairs, and participants were instructed to move sliders to indicate the degree to which they believed each particular term described the qualia of the probe tone. Participants were also given the option to "opt-out" of using any particular rating scale on any given trial by checking "does not apply / I don't know" if they felt that it was not a useful or applicable descriptive term.

Although it is common in experimental designs to list terms on a unidirectional scale (e.g., "tense—not tense"), here a bidirectional scale was preferred because

TABLE 1. *Result Statistics by Dependent Variable*

	Strong/ Weak	Complete/ Incomplete	Bright/ Dark	Expected/ Surprising	Happy/ Sad	Tense/ Relaxed	Move/ Stay
Scale degree	***	***	***	***	***	***	***
Progression	*ns*	*ns*	***	*ns*	***	*ns*	*ns*
Training	*ns*	*	*ns*	*ns*	*ns*	*ns*	***
S-deg*Prog	***	***	***	***	**	***	***
S-deg*Train	***	***	***	***	***	***	***
Train*Prog	*ns*	*ns*	*ns*	*ns*	*ns*	*	*
S-deg*Prog*Train	*ns*	*ns*	*ns*	*	*ns*	*ns*	*ns*

Note: Columns list the dependent variables, or sliders. Rows list the main effects and various interactions from the regression analysis. Each cell reports a p value category (i.e., $p < .001^{***}$, $p < .01^{**}$, $p < .05^{*}$, *ns*).

it introduced more descriptive terms (and therefore, theoretically, more options). That is, it might avoid participants thinking of scale degrees along only one parameter (e.g., "tense"). Also, by presenting the terms in this binary manner, it was believed that it would simplify an already abstract and difficult task, and make the midway point on the sliders more likely to represent "complete neutrality." An attempt was made to distribute "positive" and "negative" terms on either side of the scale to avoid creating unintentional associations.

The stimuli were presented in two blocks of 35 trials each. In between blocks participants were given a break during which they completed the OMSI questionnaire. The experiment was conducted in a sound-attenuated booth, and stimuli presented through stereo headphones (Sennheiser, HD 280 pro), with the volume adjusted to a comfortable listening level. Prior to beginning the experiment, participants were shown the list of terms to be used in the experiment and were given basic explanations of how they were meant to be interpreted.[2] Participants were encouraged to respond as quickly as possible to avoid "over thinking" their responses, and to avoid forgetting the probe tone. In addition, it was explained that the chord progressions were only setting up the context for the probe tone and that they were to rate the note that followed and not the progressions themselves. The instructions to participants read as follows:

> Please rate, using the sliders provided, how much each term describes the given note. If you feel that a term is not useful for describing the note, or you are unsure, please check "does not apply / I don't know."

[2] For example, the first participant interrupted the experiment to ask if she should be rating every note as "incomplete" since in real music one never encounters a single note by itself! After this, it was determined that in order for the results to be meaningful, participants would have to be using the sliders in roughly equivalent ways. Therefore, a basic explanation for how to interpret each term was given at the beginning of the experiment.

Results

Recall that the main hypothesis predicts that the harmonic context will influence scale-degree qualia ratings. The remaining exploratory questions asked: How consistent are individuals in their qualia judgments? And, do those with music theory training perceive scale-degree qualia in similar ways as those that do not? Mixed effects regression was used to evaluate the main hypothesis and exploratory questions. A separate test was conducted for each dependent variable (i.e., qualia rating or slider). The main factors were: the level of training (musician/nonmusician), scale degree (10 possibilities), and progression (5 possibilities). Note that responses from individuals with self-reported absolute pitch were not found to be different from the remaining musicians and so were included in the musician group. The main result of interest was that of the interaction between scale degree and progression. Indeed, for each dependent variable, a significant interaction was found for scale degree and progression (all $p < .01$). Even with a Bonferroni correction for multiple comparisons, all p values remained significant. In other words, this is consistent with the notion that harmonic context can alter a scale degree's qualia. While every comparison in the regression analysis does not necessarily yield a *meaningful* result, the complete statistics are listed for each dependent variable in Table 1.

Looking at Table 1, one can see a clear main effect for scale degree across all dependent variables. This can be interpreted to mean that despite all differences in training, and despite the changes in harmonic context, participants tended to rate at least some scale degrees in consistently different ways from other scale degrees. That is to say, while scale-degree qualia can clearly be mediated by the harmonic context, they appear to elicit a certain degree of unique qualia in and of themselves. All tests of the main effect for progression were not significant except for in two cases: Bright/Dark and

Happy/Sad. This means that, for at least one of the five progressions, the progression itself was having the effect of causing participants to alter their Bright/Dark and Happy/Sad ratings of the probe tones in a consistent way. Indeed, as will become evident from examining Figure 5, when the progression ends with a minor chord (vi or ii), it tends to cause more scale degrees (in particular, those that are chord tones) to be rated as both sadder and darker.

Recall also that participants were not given a forced-task, but rather could "opt-out" of using any particular rating scale if they felt that it did not apply to the particular scale degree in question. A simple tally by dependent variable (shown in Table 2) shows that Happy/Sad and, to a lesser degree, Dark/Bright received the highest number of opt-outs. In post-experiment interviews it was commonly reported that chromatic tones proved difficult to rate as happy or sad. It should be noted that Happy/Sad and Dark/Bright are the two dependent variables that are not (obviously) associated with tendency, strength, or melodic completion. It should also be noted that in post-experiment interviews, participants regularly commented on how they tended to associate notes that were higher in pitch with "bright," and notes lower in pitch with "dark" (recall that the stimuli were transposed so that the various scale degrees appeared at a number of pitch heights.) The work of Huron (2006) and Arthur (2017) suggests that scale-degree qualia may largely be attributed to statistical regularities associated with tendency or motion. However, there appear to be some qualitative aspects that are not learned statistically. Specifically, the association of both happy/sad and bright/dark with high/low (respectively) have been associated with ethological cues (Huron, 2015; Huron & Davis, 2013; Kraepelin, 1899/1990; Morton, 1977). Studies have suggested that the association with bright/dark and high/low (respectively) may be an intrinsic characteristic of perception (e.g., Marks, Hammeal, & Bornstein, 1987).

Two oddballs were also found in the main effects for training. While it is possible that the two sliders where this effect was found to be significant (Complete/Incomplete and Move/Stay) may simply reflect an oddity in the data, it appears that those in the nonmusician group were using these sliders in different ways than those in the musician group. Interestingly, of the correlations between all dependent variables, Move/Stay and Incomplete/Complete had the strongest correlation. Said another way, it seems that what a musician deems to be musically "incomplete" may not be heard as such by a nonmusician. Somewhat surprisingly, in addition to the significant interaction for scale degree and harmony, a significant interaction was also found for scale degree and training across all dependent variables. This finding is slightly more difficult to interpret, but suggests that at least some scale degrees are being rated as qualitatively different for the musician group compared to nonmusician group.

Recall that these tests merely compare means. Given the task that was required of participants, a large amount of variation in the responses is expected. In other words, what is of interest is not necessarily the means of the two groups, but how much their responses vary. For instance, perhaps individuals in the musician group tend to be in more agreement with each other compared to those in the nonmusician group, in which case the main difference would be the spread in variation (or standard deviation). Thus the remaining graphs and figures attempt to re-examine the data in order to clarify and expand upon the results found from the main statistical tests reported above.

Figure 5 highlights differences in perceived scale-degree qualia across the various harmonic contexts for each slider (with one graph per slider). Recall that each progression ended with a different final harmony. Thus, the harmonies listed along the *x*-axis represent each progression's final chord. Each circle represents the average qualia rating for that particular scale degree +harmony pair. Shading is used to designate which side of the slider (left or right) the average value falls on, and the size of the circle represents the strength of the rating. The first thing to notice is that there is a clear tendency for some scale degrees to have somewhat similar qualia regardless of harmonic context. This can be seen from looking for consistent horizontal patterns. For instance, there are clear "stripes" for chromatic tones compared with diatonic ones; and scale-degree 1 tends to be rated as the more positive of the two options across all dependent variables (i.e., more happy, strong, bright, etc.), compared with all other scale degrees, whose ratings tend to have more variation. These graphs also illuminate the main effects for chord progression (see Table 1), mentioned earlier, where the progressions ending with minor chords (vi and ii) tend to have the effect of making many of the ensuing scale degrees sound darker and sadder.

The most interesting effect is the change in scale-degree ratings as the harmonic context changes. For many scale degrees, changes in color and/or circle size can be seen in the graphs as one moves laterally through the harmonic changes. This tendency is especially marked, perhaps unsurprisingly, for scale degrees that are chord tones of the final harmony. For example, scale-degree 4 shows many changes in size and color

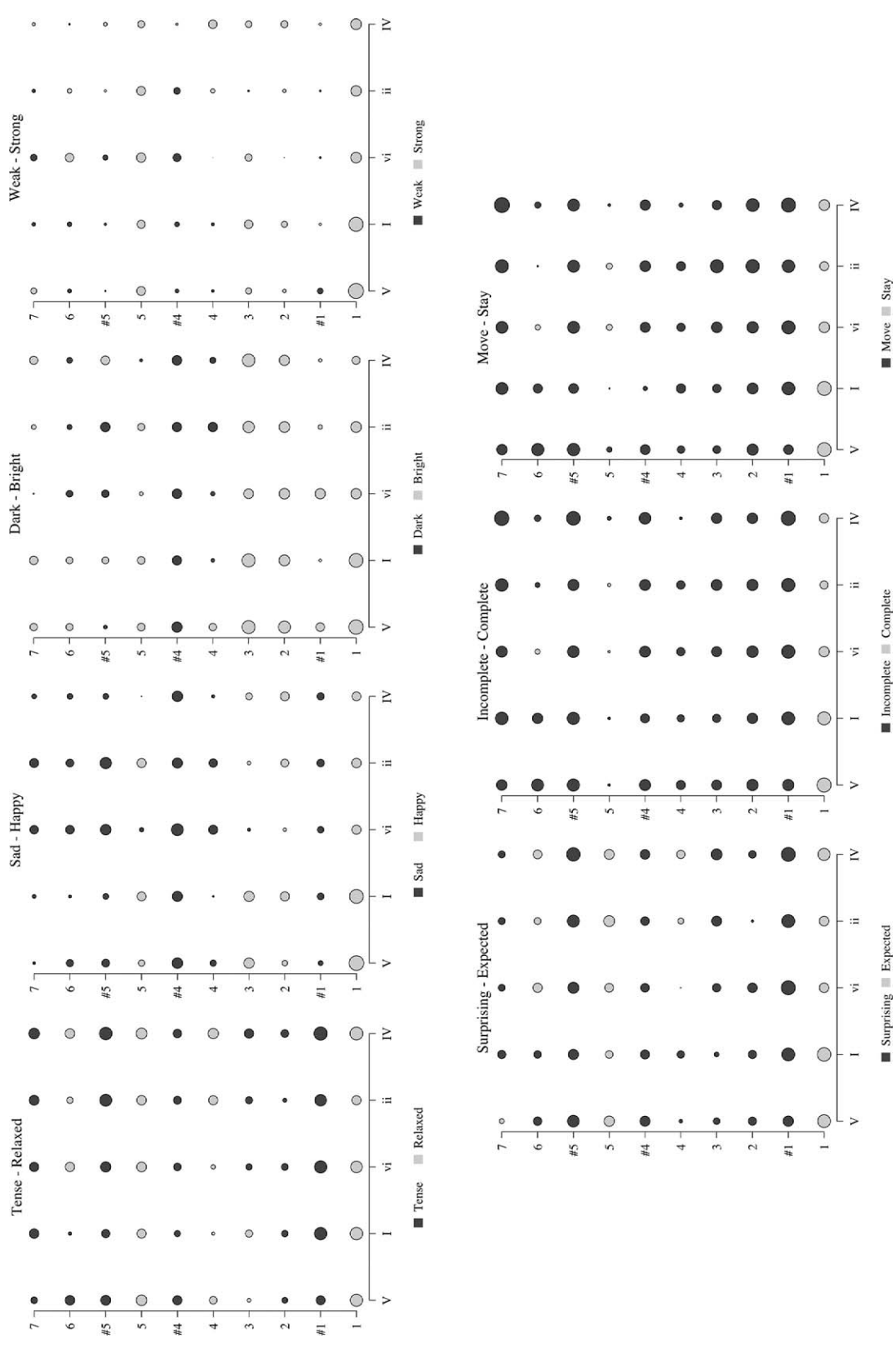

FIGURE 5. Scale-degree qualia ratings for each scale degree across harmonic contexts. There is one graph per slider. On the *x*-axis, each Roman numeral represents the *final* chord of each possible progression. Scale degrees are listed on the *y*-axis. Each circle represents the average rating across all participants for that particular scale degree following the given harmonic context. Shading indicates which side of the slider the value falls on. The larger the circle, the more extreme the average rating.

TABLE 2. *Tally of "Opt-outs" by Dependent Variable*

Strong/Weak	Complete/Incomplete	Bright/Dark	Expected/Surprising	Happy/Sad	Tense/Relaxed	Move/Stay
223	108	314	215	544	230	279

Note: Values indicate the number of times participants opted-out of rating any particular scale degree on any of the dependent variables.

across the levels of harmony for almost all graphs; and scale-degree 6 shows clear changes in qualia when it becomes a chord tone in the graphs for Surprising/Expected, Dark/Bright, Weak/Strong, and Tense/Relaxed.

One caveat to mention is that probes in this experiment were presented *after* the harmonies, not concurrently. This was done for two reasons. First, as will be explained below, one goal of this experiment was to replicate a component of Krumhansl and Kessler (1982), which used the harmony-followed-by-probe design, in which case it was preferred to use as similar a methodology as possible. Second, it was preferred to first investigate the effect of harmony on scale-degree qualia in isolation (rather than use complete melodies) and if a single scale degree were presented simultaneously with the last chord of the progression, it would be impossible to know whether the evoked qualia was due to the chord, the scale degree, or the extended harmony that would result from the combination of the two. Therefore, the present design was adopted in an attempt to isolate the individual scale degrees, and it appears from these results that the final harmony was indeed still capable of having an effect. Future research, however, will be needed in order to address the effects of concurrent sounding harmonies and scale degrees.[3]

Since the circles in the graphs from Figure 5 represent the average rating across *all* participants, it will be useful to know how much the participants are in agreement with each other. Recall that these were the main exploratory questions: How similar is qualia perceived across individuals, and how much does it depend on one's music training?

Figure 6 illustrates how well participants were in agreement with each other, as well as how consistent they were when rating repeated stimuli. One way of evaluating agreement is to examine correlation values. Accordingly, a string of values was assigned to each participant representing their responses for all 70 trials across all seven dependent measures. This string of values was then compared with every other participant's values. In addition, since each participant heard a subset of the stimuli twice, their first responses to those (20) stimuli were compared to their second responses to those same stimuli to evaluate consistency. For the between-subject comparisons, since there were 63 participants in total, and each participant's responses were compared with every other participant's, there are a total of 3,906 correlation values. Figure 6b shows the distribution of these correlations (Pearson's r values) centered around a mean of .18. However, in order to evaluate these correlations, they must be compared against something. Thus a randomized set of correlations was created by comparing each participant's string of responses to randomized versions of the other participant's responses. Since there should be no relationship between random pairs of responses, one would expect the distribution of these random correlations to be centered around zero. Indeed, randomized correlations of all responses produces the distribution as seen in Figure 6a, which has a mean of .03 with 95% of the values falling below .15.[4] In comparison, then, while the average between-subjects correlation of .18 may not appear very strong, when those ratings are scrambled and reassessed for the amount of correlation, less than 5% of the r values reach .18. This can be taken to mean that, overall, participants were in moderate agreement with each other in their qualia ratings. What about the within-subject agreement? How consistent were participants on repeat qualia judgments? Figure 6d shows the actual within-subject correlations (one value per subject) for the 20 repeated stimuli across all dependent variables, which can again be compared against correlations made from a randomized version of the data (Figure 6c). In this case, the actual within-subject correlations have a mean of .39, whereas the correlations from the randomized set of data show 95% of values falling under .15.

Thus, participants are showing a good level of consistency in their qualia judgments. Recall that these graphs so far have included the results from *all* participants, regardless of music training. It might be expected,

[3] In fact, Toiviainen and Krumhansl (2003) use this approach. See the Discussion section for more detail.

[4] Fisher's z transformation was used to convert the correlation values to an interval scale (in order to compute an average) before being converted back to r values. See Meyers, Gamst, and Guarino (2013, p. 298).

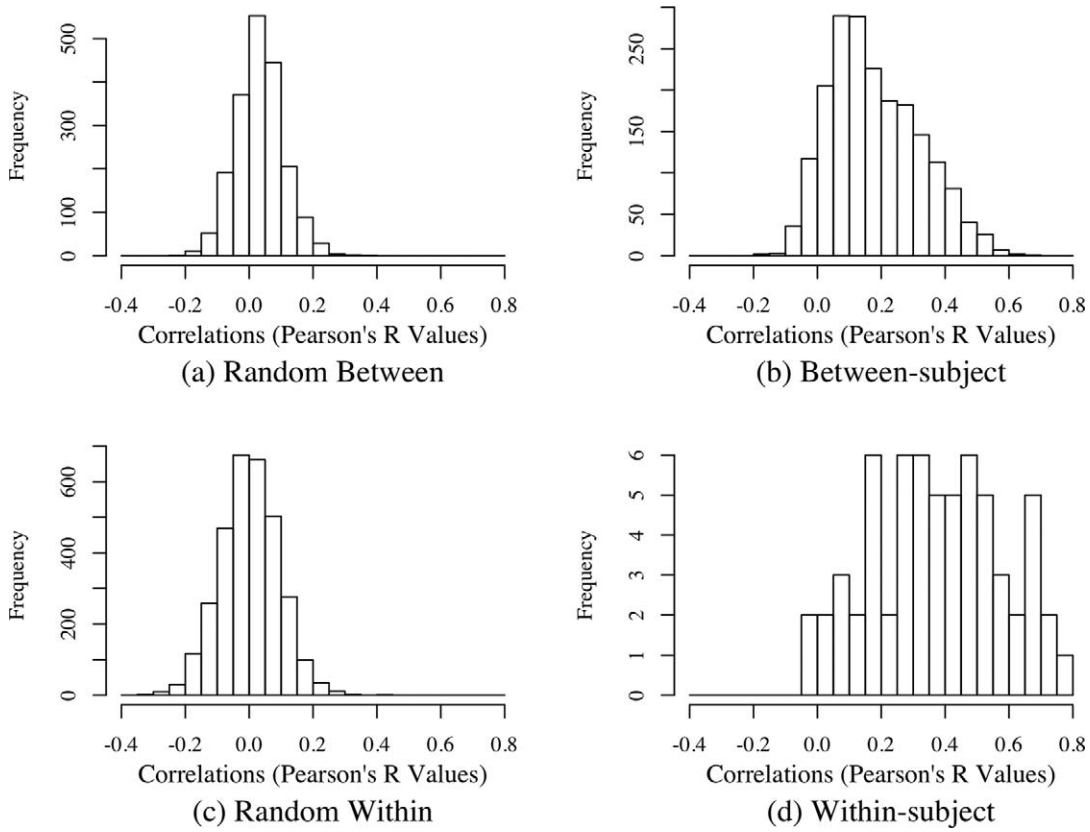

FIGURE 6. Amount of intra- and inter-subject agreement as measured by correlation. Pearson's *r* values are listed along the *x*-axes. (a) A distribution of random correlation values obtained by randomizing the between-subject ratings, with 95% of the values falling under $r = .15$ (b) The distribution of actual between-subject correlations (each participant's response correlated with every other participant's response) with a mean of $r = .18$ (c) A distribution of random correlation values obtained by randomizing within-subject ratings, with 95% of the values falling under $r = .15$ (d) The distribution of actual within-subject correlations (each participant's first response correlated with their second response to the same stimuli) with a mean of $r = .39$.

however, that for the nonmusician group—given the abstract nature of the task—their level of consistency might be worse than that of the musician group. Thus it is appropriate to examine these correlation values by splitting the participants according to their level of training.

Figure 7 shows the consistency of each participant, divided by level of training. Each circle graphs the single correlation value (Pearson's *r*) for each participant, determined by comparing their first and second responses to the repeated set of 20 stimuli across all dependent variables. The solid horizontal line drawn through each graph represents the mean of each group's correlation values (musician group: $M = .46$, $SD = .26$; nonmusician group: $M = .27$, $SD = .15$). Recall that a distribution of random, between-subject correlations showed that 95% of values fell under $r = .15$ and that the actual mean of *all* participants' between-subject correlations was $r = .39$. If participants were performing at the chance level, we would expect only 5% (or 1 person in 20) to have a correlation value greater than .15. From Figure 7 one can see that 17/22 of the nonmusicians have correlation values greater than .15 (represented by the dotted line), and 34/41 musicians have correlation values above .15. In comparing the two sets of intra-subject correlations, it can be seen that the musician group are indeed, on average, more consistent than the nonmusician group. However, there are also more participants in the musician group than the nonmusician group, making it a slightly unfair comparison. In addition, if one looks at the scatter of individual values, it is clear that there are individuals in the nonmusician group who are actually more consistent than individuals in the musician group. Also to be noted are the few individuals whose correlation values hover around zero (note there are an equal number from both the musician and nonmusician group). As suggested in the introduction, it is possible that some individuals are not capable

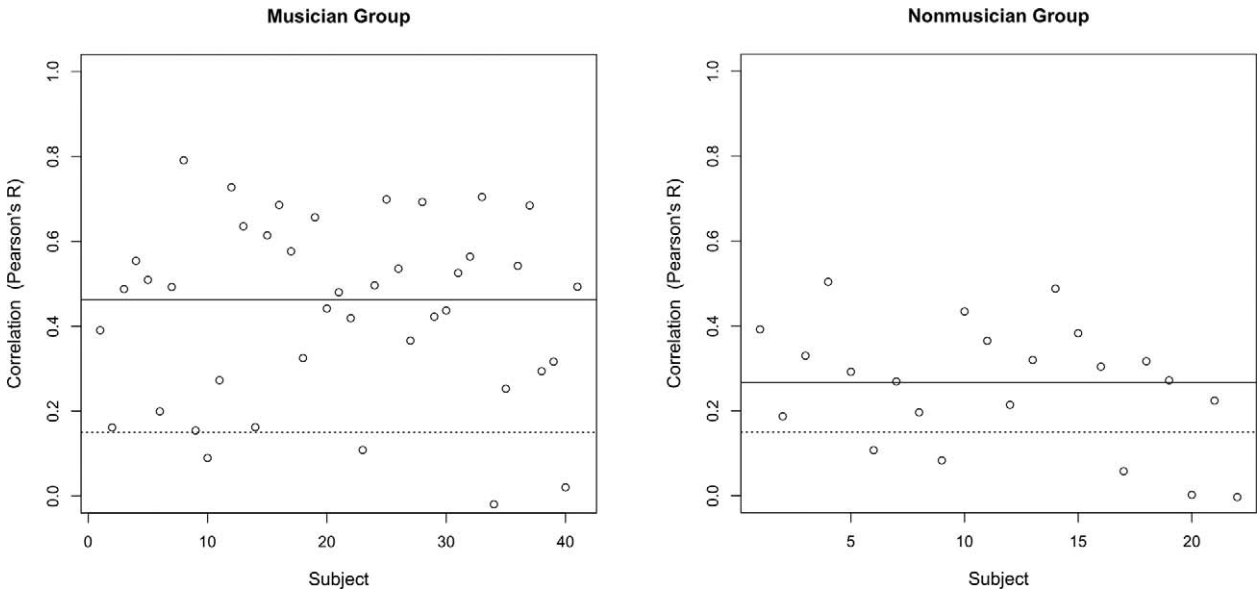

FIGURE 7. Intra-subject (within-subject) consistency across groups as measured by correlation. Each circle represents a single correlation value for an individual participant, based on the comparison of their first and second responses to the 20 repeated stimuli across all dependent measures. A solid horizontal line is drawn through the mean for each group (musician group: $M = .46$, $SD = .26$; nonmusician group: $M = .27$, $SD = .15$). A dotted horizontal line shows the 95% cutoff range from the distribution of randomized within-subject correlation values (shown in Figure 6c.)

of hearing scale-degree qualia in this way. Although it is a possibility that participants were simply tired or unmotivated, the results as presented in these graphs suggest that there may indeed be a minority of listeners who do not hear scale-degree qualia in a consistent manner. Nevertheless, those in the nonmusician group demonstrated a fair level of consistency, despite the fact that they do not possess the skills to identify scale degrees by ear. This finding is further expounded upon in the Discussion section.

Recall that some of the rating scales used in the experiment related to the concepts of resolution and closure (i.e., Surprising/Expected, Move/Stay, Complete/Incomplete). These ratings can be compared to the "goodness of fit" ratings (henceforth "key profiles") from Krumhansl and Kessler (1982; these profiles will be referred to as K&K). Since the methodology used in the present study was very similar to that of Krumhansl and Kessler (1982), a portion of the present study can be thought of as a conceptual replication (and expansion) of their original study. Since Krumhansl and Kessler only used experienced musicians in their study, in order to make a fair comparison with the present study, only the results from the musician group will be considered. The K&K key profiles were obtained by taking the average "goodness of fit" rating (across all participants) of a probe tone (one of the 12 possible chromatic scale degrees) following a predominant-dominant-tonic (PDT) progression in a major or minor key. Krumhansl and Kessler do not provide the unique profiles for the individual progressions, they only average them together (see Krumhansl & Kessler, 1982, p. 343).

In order to compare findings, a similar set of scale-degree "profiles" were created for the three rating scales related to resolution and closure: Complete/Incomplete, Wants to Move/Wants to Stay, and Surprising/Expected. Since the present study only used 10 scale degrees, the extra two scale-degrees from the K&K profiles were removed so that the number of scale degrees in all profiles would match. In Figure 8, the Krumhansl and Kessler major key profile is compared with profiles created using data from the present study, with a separate profile (see legend) for each harmonic context. Similar to Figure 5, each point represents the average rating across participants (in this case, only musician participants) for that scale degree +harmony pair. Each of these profiles were then correlated with Krumhansl and Kessler's original profile.

As can be seen from Figure 8, the current profiles are all relatively well matched to the overall contour of the K&K profile, with the same dips for chromatic tones, and peaks for diatonic tones. However, the strongest correlations tend to be found between the K&K profile and the profiles for tonic context (see Table 3 for complete list of Pearson's *r* values). As already noted, when non-tonic chord contexts precede the scale-degree

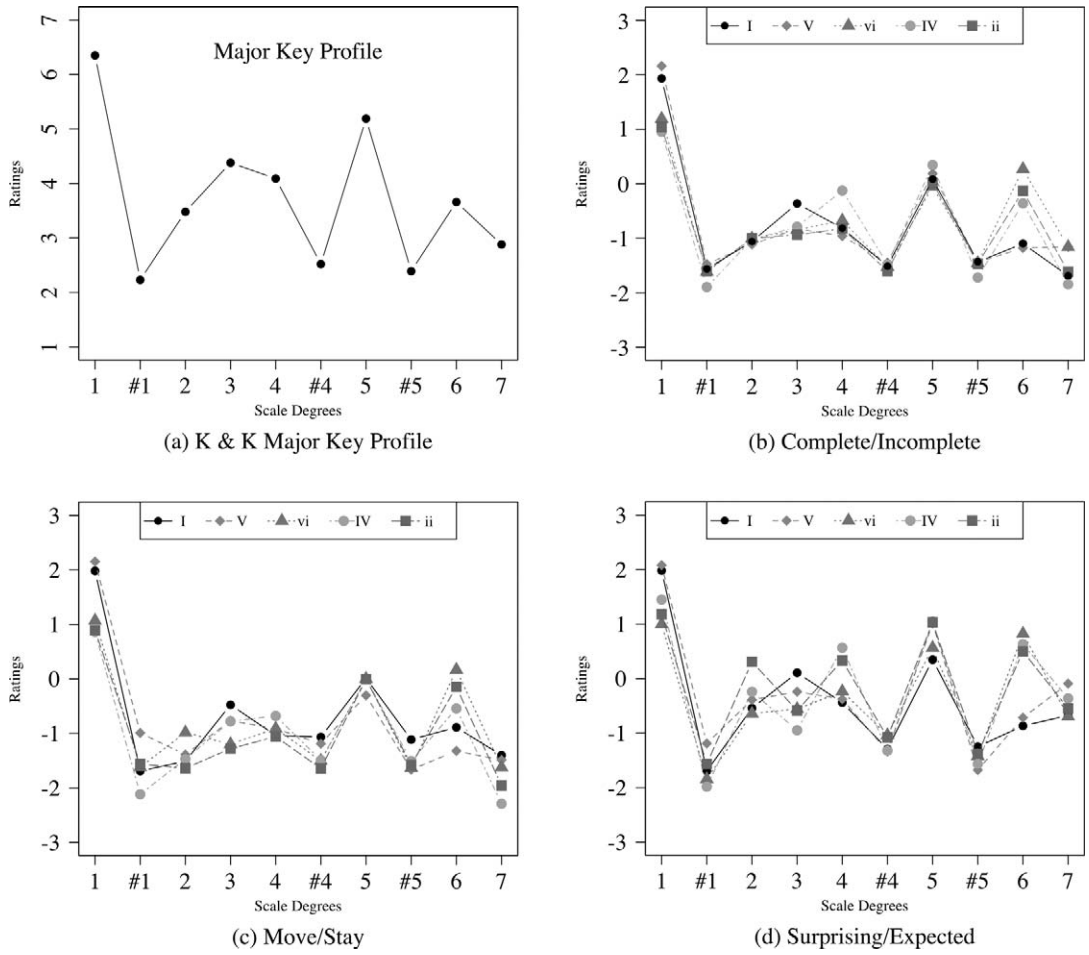

FIGURE 8. Present data compared to Krumhansl and Kessler's (1982) key profiles. The *y*-axes represent the rating scale (note that Krumhansl and Kessler used 1 to 7 while the present study used -3 to +3. The *x*-axes represent scale degrees. (a) K&K major key profile with two chromatic scale degrees removed (to be consistent with present data). b) Ratings across all chord contexts for "Complete/Incomplete." c) Ratings across all chord contexts for "Move/Stay." d) Ratings across all chord contexts for "Surprising/Expected." To be consistent with Krumhansl and Kessler, the data reported in these graphs shows only the responses from the "musician" group. Roman numerals in the legend represent the last sounding chord in each progression.

TABLE 3. *Correlations Between the K&K Major Key Profile and Each of the Current "Chord" Profiles*

	I	V	vi	IV	ii
Expected/Surprising	.96	.92	.84	.84	.86
Stay/Move	.88	.84	.86	.92	.85
Complete/Incomplete	.94	.90	.88	.94	.92

Note: Each Roman numeral represents the final chord of the "context" progression. Correlations represented by Pearson's correlation coefficients (*r*).

probe, the ratings for the scale degrees are altered, with the most marked effects occurring for chord-tones. This can be most easily seen in Figure 8 by looking at the ratings for scale degrees 2, 4, and 6, which tend to "pop out" of the texture when those notes become chord tones. This "chord-tone" effect caused the correlations between the K&K profile and the non-tonic context profiles to decrease.

Recall that the progressions used by Krumhansl and Kessler all end with the tonic, while in the present experiment they all imply the same key but end on different harmonies. Since in this experiment the interaction of harmony and scale degree was significant, it implies that Krumhansl and Kessler's key profiles might be better explained as "chord profiles" rather than "key profiles." Said another way, while possible that they are testing tonal stability, as they claim, it appears at least a strong possibility that they may have largely been testing which scale degrees are a good fit with a tonic triad context.

Krumhansl and Kessler argue that their profiles reflect scale-degree relations within a key, since the ratings for each scale degree consistently matched an expected scale-degree "hierarchy," with tonic being the highest rated, followed by dominant, followed by the mediant. In the present experiment, however, the only profile that exactly matched the contours in the K&K profile (and therefore their hierarchical claim) was that of tonic context. Again, this suggests that the K&K profile may not represent scale-degree relations to a key, but rather scale-degree relations to a harmonic context.

Aarden (2003, p. 66) proposed that, "the probe-tone method [from which the key profiles were derived] encourages listeners to hear the tone being tested as occurring in a phrase-final position . . . since musical phrases typically end with harmonic 'cadences' or with stereotyped melodic figures." Indeed, Aarden found that the K&K profile closely matched the zeroth-order distribution of scale degrees at phrase-final positions. Obviously, the probe-tone method was also used here, implying that listeners may have heard the probes as an "ending." However, unlike Krumhansl and Kessler's main experiment, the present experiment consisted of lead-in progressions that ended with various harmonies (not only tonic), and as such, while possibly heard as a "phrase final position," some of these contexts would be unusual at a cadence, providing at least some minor discouragement to listeners for hearing the final note as an "ending." For instance, it would be very uncommon for a phrase to end with the supertonic chord. Moreover, the increase in chord tones rated as "expected" in some chord contexts is not consistent with what would be expected for melodic tonal completion. For example, $\hat{4}$ and $\hat{6}$ are two of the least likely scale degrees to end a phrase (Aarden, 2003); yet, in the present experiment, when those scale degrees were preceeded by IV or vi, the expectancy and completeness ratings for those scale degrees increased. Since phrases are simply more likely to end on tonic (I), and the K&K profiles were created entirely from ratings following a tonic harmony, it makes sense that the K&K profiles more closely resemble the distribution of phrase-final scale degrees. Thus, while Aarden's proposal that the K&K profiles are more apt to be representative of phrase-final positions is appropriate, the findings from the present experiment are at least *also* consistent with the interpretation that participants are responding to effects from the local chord context.

First, it is important to note that, in fact, *part* of the hierarchy proposed by Krumhansl and Kessler *can* still be seen in the present data, despite the changes in harmonic context. Specifically, the "hierarchy," as demonstrated by the present research, only seems to involve the first and fifth degrees of the scale, with the effect of the first degree being the strongest. It should also be noted that even in Krumhansl and Kessler's original experiment, the difference between ratings for scale degrees 3 and 4 was only 0.3 (on a seven-point scale), which could easily have arisen by chance. Thus, it seems that Krumhansl and Kessler's claim that the ratings of scale degrees match a tonal hierarchy (as influenced by Meyer, 1956) has some merit; however, the hierarchy may only involve the tonic and dominant notes of the scale. In fact, in the majority of profiles from the present data, scale-degree 4 tends to be rated (moderately) higher than the mediant on all three graphs, suggesting the possibility that a hierarchy (if it exists) may have more to do with the cycle of fifths (i.e., a fifth above and below the tonic) than the tonic triad.

While the K&K key profiles remain widely in use (e.g., Aarden, 2003; Bigand, 1997; Butler, 1989; Huron & Parncutt, 1993; Pauws, 2004; Takeuchi, 1994; Temperley, 2004, 2007; Sapp, 2005), the remainder of their 1982 paper actually detail additional, and more complex, experiments that attempt to demonstrate perceived key distances, as well as show converging evidence for their tonal stability theory. In these later experiments, the authors in fact do use a multitude of progressions; the majority of which do *not* end on tonic. The authors' primary claim is that the overall results of these subsequent experiments show that the effect of the key overrides that of the local chord; however, the evidence they provide in support of this claim is poor. First, their progressions ending with tonic *always* provided the strongest correlation with the key profile. Second, in nonmodulating progressions with a tonic triad in the middle, the correlations with the key *following* the tonic harmony get worse, suggesting that either participants believed that they were moving to a new key, or they were simply rating the goodness of fit with the local chord. Moreover, Krumhansl and Kessler do not provide any data profiling the scale degrees in each unique chord context (what I call "chord profiles"), nor do they examine the relationship between the profile at each point in the sequence and the key profile for that local chord (as a tonic). Rather, they only show correlations between the profile at each point in the sequence and the profile of the *intended* overall key, and correlations between chord profiles (taken from an earlier experiment in the same paper) and the intended overall key. Note also that all of these correlations are averaged over multiple progressions.

Given the strong correlations between the present paper's "chord profiles" and Krumhansl and Kessler's

original key profiles, one could argue that there is simply a difference in emphasis between the two papers in that here I emphasize the effects of the local chord while Krumhansl and Kessler emphasize the effect of the overall key. While this may be true, Krumhansl and Kessler do acknowledge multiple times the effects of the local chord (e.g., Figure 9, pp. 359-360), yet they did not subject those local chord effects to significance testing, nor did they attempt to update their original key profiles based on the data from these subsequent experiments, potentially overlooking its importance.

Krumhansl and Kessler's key profiles directly led to the creation of the widely used Krumhansl-Schmuckler key-finding algorithm (described in Krumhansl, 1990, Chapter 4). This algorithm is frequently referenced as the only key-finding algorithm to be based on perceptual data. A component of the present paper can be seen as an attempt to replicate a portion of their 1982 study using new perceptual data. Others have criticized Krumhansl and Kessler's work both in terms of the probe-tone methodology, as well as the ability of the profiles to accurately predict key (e.g., Aarden, 2003; Butler, 1989; Leman, 2000) The present work, then, can be seen, not as adding a new criticism, but perhaps adding evidence in support of an existing criticism; in that it provides empirical support for the notion that the K&K profiles may be underestimating the effects of the local chord, and therefore may not be as accurate a representation of perceived scale-degree structure within a key.

It should be pointed out that Toiviainen and Krumhansl (2003) carried out an experiment investigating tonality induction and scale-degree "fit" using a *concurrent* context and probe in an ecologically valid context (Bach's organ *Duetto* BWV 805). Listeners judged how well a scale-degree probe fit with the music during various "slices" of time. This methodology was clearly chosen in an effort to improve upon earlier probe-tone methods and collect data that better represents "real time" experiences of music. Nevertheless, Toiviainen and Krumhansl relied extensively on the K&K profile data (1982) in evaluating their results (e.g., pp. 761-762).

One final caveat relates to the language used in Krumhansl and Kessler's experiment compared with the present one; which is, there may be a difference between rating "fit" and "expected" (or "complete," etc.). First, it may be that the interpretations by the participants cannot be controlled in the experiment. Even when instructed to rate how a given scale-degree "fits" in a given context, participants may change their rating (unconsciously or unconsciously) to adopt how a given scale degree might "complete" a given context. Thus, although "completing" and "fitting" are not the same, participants may be changing their interpretations such that, in the end, they are largely used in the same way. In post-experiment interviews this question of interpretation was raised to participants with the example of how they rated scale-degree 7 following a dominant versus tonic context. Some participants reported being inconsistent, while others seemed unaware of any differences while completing the experiment. In the profiles from the data broken down by chord context, it is clear that a double usage was occurring: Across ratings for "Surprising/Expected" following a dominant context, results found that the highest rating went to $\hat{1}$, followed by $\hat{5}$ and $\hat{7}$. This means that participants were rating notes that *fit* the current context as expected (i.e., $\hat{5}$ and $\hat{7}$), but also notes that *followed* the current context as expected (i.e., $\hat{1}$). Of the three dependent variables that were thought to most likely match Krumhansl and Kessler's "goodness of fit" task, when the context ended with tonic, "Surprising/Expected" had the highest correlation to the K&K profile. However, of the dependent measures *overall* (averaged across all chord contexts), "Tense/Relaxed" and "Complete/Incomplete" showed the strongest correlations to the K&K profile at .92, while "Surprising/Expected" was .86. Overall, the relatively good correlations with the K&K profile suggest that, despite inconsistent interpretations of the terms, the difference in language used between the two experiments (i.e., "fit" versus "expected" or "complete") is probably trivial.

Discussion

In this paper, an experiment was detailed that attempted to continue a line of investigation on scale-degree qualia, following Huron (2006). The results of the pilot study suggest that even when listeners have limited musical experience and are presented with physical stimuli (as opposed to using their imagination), listeners came up with similar types of descriptive terms (that could be categorized in similar ways) as found in Huron (2006). Throughout the main experiment, the same stimuli in the same context was repeated, to test how consistent listeners were in their judgments, and it was found that listeners were not only capable of performing the task, but showed relatively good consistency. With regard to the question of music training, it appears that those with music training certainly appear more consistent. However, understanding why musicians might perform "better" at this task is not straightforward. As Hansberry (2017) claims, one learns only about "what the participants believe their experience should be like"

and thus "a participant's concepts of scale degrees may well outrun their qualia." It is true that participants are asked to reflect upon some experience, and attempt to convey that experience in words (or evaluate it with a set of given terms), which is admittedly a difficult and rather crude task. For those participants with music theory training, it is impossible to know for certain whether they might be identifying a scale degree and responding in a way that simply echoes their conceptual knowledge of that scale degree. Recall that there was a significant interaction for scale degree and level of training, suggesting that at least some scale degrees were being rated as qualitatively different for the musician group compared to the nonmusician group. It is possible, however, that it is not the conceptual knowledge that creates the difference, but perhaps some other factor, such as an increased amount of statistical learning. Presumably—even when considering statistical listening arising from listening—someone who regularly practices an instrument and/or regularly participates in musical activities will have acquired more statistical knowledge. However, as pointed out, individuals in the nonmusician group were still capable of rating scale-degree qualia, with a level of consistency greater than would be expected by chance. This is consistent with the notion that scale-degree qualia can exist apart from mere identification, although it is unclear how much conscious understanding of the concept of scale degree might *add* to that qualia. One avenue for future research might attempt to pry apart these empirically entwined factors through the use of implicit measures, such as reaction time studies.

The most basic but perhaps the most critical component of this research was the inclusion of harmonic progressions that end with harmonies other than tonic. As it relates to scale-degree qualia, indeed, the harmonic context was shown to have significant influence over listeners' judgments. Although this finding itself may seem unsurprising, scale-degree qualia have until now only been considered in isolation, and given the strong influence of chord-tones on qualia judgments, this suggests that a large part of a scale degree's ability to influence qualia may be tied to a harmony (or perhaps harmonic function) with which it is most closely attached.

Krumhansl and Kessler's work on the perception of key structures has become seminal in the field of music cognition. The methodology used in the present experiment allowed for a replication of a part of Krumhansl and Kessler's (1982) perceptual study. To the author's knowledge, no replication of this perception experiment has previously been carried out. In the present study, the regression results showing a main effect for scale degree, in conjunction with the scale-degree profiles across multiple chord contexts, support the notion that scale degrees have a unique "identity" within a key, despite changes in local chord contexts, consistent with Krumhansl and Kessler's claim. However, Krumhansl and Kessler's discounting of the role of local chord effects may have been premature. The present work showed a significant interaction for scale degree and progression, with subsequent analyses showing an effect for chord-tones being rated as "better fitting." The primary claim of Krumhansl and Kessler that scale-degrees 1, 5, and 3 form a tonal-stability "hierarchy" (in that order), was only replicated following a tonic context. While scale-degrees 1 and, to an extent, 5 were consistently rated high, the remaining scale-degree ratings varied according to chord context.

Author Note

Claire Arthur is now at the Schulich School of Music, McGill University.

This research was supported in part by a doctoral fellowship from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Correspondence concerning this article should be addressed to Claire Arthur, Music Technology Area, Schulich School of Music, McGill University, 555 Sherbrooke St. W., Montreal, Quebec, H3A 1E3. E-mail: claire.arthur@mcgill.ca

References

Aarden, B. J. (2003). *Dynamic melodic expectancy* (Unpublished doctoral dissertation). The Ohio State University.

Arthur, C. (2017). Taking harmony into account: The effect of harmony on melodic probability. *Music Perception*, *34*, 405-423.

Bigand, E. (1997). Perceiving musical stability: The effect of tonal structure, rhythm, and musical expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 808-822.

Browne, R. (1981). Tonal implications of the diatonic set. *In Theory Only*, *5*(6-7), 3-21.

Butler, D. (1989). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception*, *6*, 219-241.

Butler, D., & Brown, H. (1981). Diatonic trichords as minimal cue cells. *In Theory Only*, *5*(6-7), 39-55.

Dowling, W. J. (2010). Qualia as intervening variables in the understanding of music cognition. *Musica Humana*, *2*, 1-20.

Goguen, J. (2004). Musical qualia, context, time and emotion. *Journal of Consciousness Studies*, *11*(3-4), 117-147.

Hansberry, B. (2017). What are scale degree qualia? *Music Theory Spectrum, 39*, 182-199.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Huron, D. (2015). Affect induction through musical sounds: An ethological perspective. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *370*(1664): 20140098.

Huron, D., & Davis, M. (2013). The harmonic minor scale provides an optimum way of reducing average melodic interval size, consistent with sad affect cues. *Empirical Musicology Review, 7*, 103-117.

Huron, D.m & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology, 12*, 154-171.

Kraepelin, E. (1990). *Psychiatry: A textbook for students and physicians* (J. M. Quen, Ed., H. Metoiu & S. Ayed, Trans.). Canton, MA: Science History Publications. (Original work published 1899)

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford, UK: Oxford University Press.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*, 334-368.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 579-594.

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by.* Chicago, IL: University of Chicago Press.

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, *17*, 481-509.

Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, *52*(1), 1-102.

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

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Los Angeles, CA: Sage.

Morton, E. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, *111*, 855-869.

Pauws, S. (2004). Musical key extraction from audio. *Proceedings of the 5th International Society of Music Information Retrieval (ISMIR) conference* (pp. 96-99). Barcelona, Spain: ISMIR.

Raffman, D. (1993). *Language, music, and mind*. Cambridge, MA: MIT Press.

Roholt, T. C. (2014) *Groove: A phenomenology of rhythmic nuance*. New York: Bloomsbury Academic.

Sapp, C. (2005). Visual hierarchical key analysis. *Computers in Entertainment*, *3*(4), 1-19.

Schiavio, A., & van der Schyff, D. (2016). Beyond musical qualia. Reflecting on the concept of experience. *Psychomusicology*, *26*, 366-378.

Shepard, R. N. (2009). One cognitive psychologist's quest for the structural grounds of music cognition. *Psychomusicology*, *20*, 130-157.

Takeuchi, A. H. (1994). Maximum key-profile correlation (mkc) as a measure of tonal structure in music. *Perception and Psychophysics*, *56*, 335-346.

Temperley, D. (2004). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Toiviainen, P., & Krumhansl, C. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*, 741-766.

Tye, M. (2015). Qualia. *Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/entries/qualia/.

Zentner, M. (2012). A language for musical qualia. *Empirical Musicology Review, 7*, 80-83.

Appendix A
List of Unique Responses to Each Scale Degree from All Participants in the Pilot Study.

The words were later sorted and grouped according to synonyms/antonyms, and then those groups were given labels resulting in the 7 category names (i.e., slider labels) used in the main experiment.

"Do"	**"Di/Ra"**	**"Re"**	**"Ri/Me"**	**"Mi"**	**"Fa"**
confident	unsettled	unfinished	upscale	excited	happy
complete	dark	almost finished	bright	unsure	unsettling
happy	sad	almost com-	happy	merry	not final, has to-
centered	weak	plete	unsure	happy	go somewhere
finalizing	grounded	bright	questionable	stationary	definite
whole	stationary	not complete	unsteady	white	knows where it's -
beginning	gloomy	out of place	off-sounding	descending	going
bark (like a dog)	dark	warm	out of the box	lighter	core
generic		stationary	high		abrupt
thin		generic	out of place		happy
low		light			airy, light
peak		repetitive			basic, boring
heavy		depth			yellow
harmony		almost the end			bright
strong		needs to be re-			graceful
stable		solved to tonic			suspension
		waiting			needs to be -
		common			resolved down
		rising			more interesting -
		different			than other notes
					predominant

"Fi"	**"Sol"**	**"Si/Le"**	**"La"**	**"Te"**	**"Ti"**
tension	playful	light	happy	deep	incomplete
incomplete	up and down	unsettling	wants to go up	dark	tension, unfinished
unsettled	excited	tension	neutral	sad	darker
wants to go up	mournful	stable	excited	hope	wants to go up
changing	sad and longing	moderately -	wants to go up	heavy	tense
tipsy	has an edge-	sad	suspended	water	incomplete
angry	to it, spiky		bright	jazzy	jabby, punchy
sharp, jagged	bright		changing tone	stable	uplifting
fragile	happier note		dark	incomplete	sad or angry
showtune	static, boring		light, fluffy		unresolved
green	ditzy		larger, old man		playful
needs to be-	energized		unnatural		uncomfortable
resolved	blue		loose		
doesn't fit in-	white		calm		
the scale	feels like it fits-		high		
unstable	but needs to-		pink		
	keep going		doesn't fit		
	passing		different		
	fitting, but un		doesn't fit in-		
	finished		minor		
	unexpected		could resolve up-		
	common		or down		
			dark		