Bimodal distribution of performance in discriminating major/minor modes

Charles Chubb,[a] Christopher A. Dickson, Tyler Dean, Christopher Fagan, Daniel S. Mann, Charles E. Wright, and Maime Guan
Department of Cognitive Sciences, University of California at Irvine, Irvine, California 92697-5100

Andrew E. Silva
Department of Psychology, University of California at Los Angeles, Los Angeles, California 90095-1563

Peter K. Gregersen and Elena Kowalsky
The Feinstein Institute for Medical Research, 350 Community Drive, Manhasset, New York 11030

(Received 2 November 2012; revised 6 July 2013; accepted 10 July 2013)

This study investigated the abilities of listeners to classify various sorts of musical stimuli as major vs minor. All stimuli combined four pure tones: low and high tonics (G_5 and G_6), dominant (D), and either a major third (B) or a minor third ($B\flat$). Especially interesting results were obtained using tone-scrambles, randomly ordered sequences of pure tones presented at ≈15 per second. All tone-scrambles tested comprised 16 G's (G_5's + G_6's), 8 D's, and either 8 B's or 8 $B\flat$'s. The distribution of proportion correct across 275 listeners tested over the course of three experiments was strikingly bimodal, with one mode very close to chance performance, and the other very close to perfect performance. Testing with tone-scrambles thus sorts listeners fairly cleanly into two subpopulations. Listeners in subpopulation 1 are sufficiently sensitive to major vs minor to classify tone-scrambles nearly perfectly; listeners in subpopulation 2 (comprising roughly 70% of the population) have very little sensitivity to major vs minor. Skill in classifying major vs minor tone-scrambles shows a modest correlation of around 0.5 with years of musical training.
 [http://dx.doi.org/10.1121/1.4816546]

PACS number(s): 43.75.Cd, 43.66.Lj, 43.66.Hg [DD] Pages: 3067–3078

I. INTRODUCTION

The distinction between the major vs minor diatonic scales is central to western music. It is often asserted that major melodies sound happy whereas minor melodies sound sad. This mysterious proposition has baffled composers, philosophers, and musicologists for centuries; despite the fact that it refers to nothing outside itself, a melody's "mood" can strike many listeners with stunning immediacy. And indeed, substantial research supports the claim that the mode of a melody conveys mood (happiness vs sadness).[1–11]

However, the evidence is far from unequivocal that the major and minor modes are as vividly distinctive as these observations might suggest. Halpern and colleagues[12,13] had listeners rate the similarity of melodies that differed in one or more of rhythm, contour, and mode. They found that (1) melodies that differed only in mode (but that had the same contour and rhythm) were rated as highly similar and (2) listeners had difficulty discriminating melodies with identical rhythm and contour but differing in mode. Leaver and Halpern[14] found that nonmusicians were unable to discriminate tunes from identical-except-for-mode tunes and showed only slight improvement with training. Moreover, musicians were also surprisingly far from perfect at this task. However, when nonmusicians were asked to classify tunes as happy vs sad instead of major vs minor, their performance improved, and they were subsequently able to make use of the happy = major and sad = minor association to improve their performance at classifying tunes as major vs minor. Blechner[15] used as stimuli triadic chords whose central component varied in different conditions over small steps between the minor vs the major third. Listeners were asked to classify stimuli as major vs minor. Strikingly, many listeners were unable to perform this task with success rates above chance even for pure major vs minor triads. Crowder[2] replicated this result and noted in addition that his listeners seemed to fall into two distinct classes: those who could do the task vs those who could not. It is this result that motivates the current study. Our primary aim is to determine if there really are two distinct classes of listener: those who can hear the difference between major vs minor modes and those who cannot.

A. Tone-scrambles

We use a new class of stimuli called "tone-scrambles" designed to isolate effects due to variations in mode from effects due to other aspects of musical structure. Examples of the tone-scrambles used in all experiments are provided in http://hdl.handle.net/10575/9881. Our tone-scrambles comprise 32 tones presented at the rate of 15.38 per second (each tone lasts 65 ms); thus, a given tone-scramble lasts 2.08 s. The tones in the tone-scramble occur in random order creating an effect akin to alien birdsong. The frequencies of the

[a]Author to whom correspondence should be addressed. Electronic mail: cfchubb@uci.edu

tones used in the stimuli were drawn from the equally tempered scale: G_5 : 783.99 Hz, $B_5\flat$: 932.33 Hz, B_5 : 987.77 Hz, D_6 : 1174.66 Hz, and G_6 : 1567.98 Hz.

A tone-scramble is a type of "sound texture"[16,17] that shares some characteristics of music, which often features fast flurries of notes. A sound texture is an auditory stimulus (such as the sound of galloping horses or wind in leaves) that displays variability at the micro-level, but also qualitative homogeneity at the macro-level. McDermott and Simoncelli[17] present a set of simple, neurophysiologically motivated statistics that suffice to determine the perceptual quality of a given sound texture. It is claimed that any two sound textures equated in these statistics will sound like different samples of the same sound texture. Thus, for example, if a sample of sound texture is synthesized to match a sample of raindrops-on-pond-surface in all of these statistics, then the synthesized sample will sound like another sample of raindrops-on-pond-surface. It should be noted, however, that although the set of statistics offered by McDermott and Simoncelli[17] may be sufficient, some statistics in this set may well not be *necessary* to determine the perceptual quality of a sound texture. One implication of the current results is that different listeners have different necessary sets of McDermott–Simoncelli statistics.

1. Details about the tones used in the stimuli

Each tone in a tone-scramble is a pure tone 65 ms in duration, comprising 3250 samples presented at 50 000 samples per second. To prevent clicking, each tone is windowed by the raised cosine function in Eq. (1) with $a = 1125$ and $b = 3250$,

$$W(t) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\frac{\pi t}{a}\right)\right), & 1 \le t \le a \\ 1, & a+1 \le t \le b-a \\ \frac{1}{2}\left(1 - \cos\left(\frac{\pi(t-b)}{a}\right)\right), & b-a+1 \le t \le b. \end{cases} \quad (1)$$

In Experiment 3, in addition to testing listeners in discriminating major vs minor tone-scrambles, we also test them in discriminating major vs minor chords. Each tone used in a chord is a pure tone 1 sec in duration windowed using Eq. (1) with $a = 1125$ and $b = 50\,000$.

II. EXPERIMENT 1

A. Method

1. Stimuli

This experiment used two types of tone-scrambles which were referred to as "type 1" and "type 2" stimuli. Both types contained 32 tones including 8 each of G_5 and G_6 (the low and high tonic of the scale) and eight D's (the dominant). In addition, type 1 stimuli contained eight B's (the major third) whereas type 2 stimuli contain eight $B\flat$'s (the minor third).

2. Participants

Eighty listeners, including three experimenters, participated. All listeners but one (who was 52) were between 18 and 26 years of age with self-reported normal hearing. Most were recruited through the UC Irvine School of Social Sciences Subject Pool. The UC Irvine Institutional Review Board approved the experimental procedures in this and in the other two experiments reported here.

3. Testing protocol

Each listener first filled out a questionnaire that included items that registered the age and sex of the listener and also whether or not he/she had received any formal training in singing or in playing an instrument, and if so, for how many years. The listener was then tested in discriminating type 1 vs type 2 tone-scrambles. Testing was conducted with headphones in a quiet lab; volume was adjusted to a comfortable level for each listener individually. Before testing, the listener heard ten tone-scrambles that alternated between type 1 and type 2; the listener initiated each presentation with a button-press and was visually informed which type of tone-scramble had just been presented. After listening to these 10 examples, the listener was tested in 90 experimental trials. The listener initiated each trial with a button-press; after the stimulus was presented, the listener responded by pressing a "1" or a "2" on the keyboard and received visual correctness feedback. The listener was not told that type 1 (type 2) tone-scrambles were "major" ("minor"), nor was the listener encouraged to associate type 1 (type 2) tone-scrambles with "happiness" ("sadness"). The tone-scrambles used in this study were produced ahead of time and saved as .wav files, and every listener heard the same 100 tone-scrambles. However, the 5 type 1 and 5 type 2 tone-scrambles which got used in the initial 10 examples were randomly determined for each listener, as was the order of presentation of the remaining 90 tone-scrambles.

B. Results

To allow task performance to stabilize, we take proportion correct on the last 45 trials as our dependent variable. The histogram of this score across listeners is shown in Fig. 1. This histogram is strikingly bimodal, with an upper group (which includes 24 listeners—30%) showing high levels of competence in the task and a lower group (which includes 56 listeners—70%) showing much lower levels. Indeed, the mean score for the lower group is 0.5440, which is barely (but significantly, $t_{df=55} = 3.94$, $p < 0.001$) greater than 0.5.[36]

C. Discussion

Figure 1 shows that the task of classifying type 1 vs type 2 tone-scrambles is surprisingly effective at partitioning listeners into two distinct subpopulations. It is tempting to conclude that these two subpopulations consist of listeners who can hear the difference between major vs minor modes and listeners who cannot. There are, however, several reasons to doubt this conclusion.

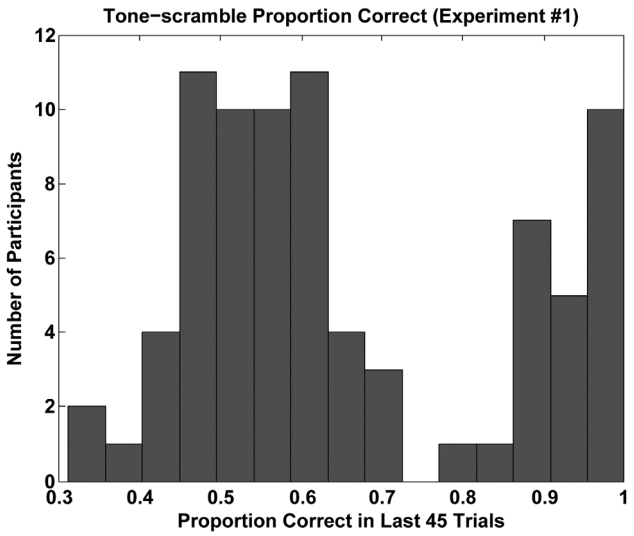

FIG. 1. The distribution across participants of proportion correct on the last 45 trials in Experiment 1. Each bar shows the number of listeners whose score fell within the given bar-bracket.

First, type 1 and type 2 tone-scrambles differ not only in mode but also in mean pitch-height. If we assign pitch height 0 to $G5$ and increase in steps of 0.5 up the chromatic scale, then the mean pitch-height of a major ("Type 1") tone-scramble in Experiment 1 is 3.0 whereas the mean pitch-height of a minor ("Type 2") tone-scramble is 2.875. It is thus possible that the superior performance of high-performing vs low-performing listeners is driven by sensitivity to tone-scramble mean pitch-height rather than tone-scramble mode.

Second, several features of the Experiment 1 protocol may have suppressed the performance of low-performing listeners: (1) Musically untrained listeners have difficulty classifying melodies as major vs minor; however, if instead they are asked to classify them as happy versus sad, they do better.[14] Low-performing listeners in Experiment 1 might show similar improvement if asked to classify tone-scrambles as "happy" vs "sad." (2) The 45 training trials used in Experiment 1 may have been too few; low-performing listeners might improve with more practice. (3) Low-performing listeners might also do better if they were (a) alerted to the particular trial on which testing was to begin and (b) given external incentive.

Experiment 2 was designed to investigate these issues.

III. EXPERIMENT 2

A. Method

1. Stimuli

This experiment used two main types of tone scramble which were identified to listeners as "HAPPY (major)" and "SAD (minor)" stimuli; we will refer to them as "major" and "minor." There were two types of major tone-scramble: high-pitch-height (low-pitch-height) major tone-scrambles had 7 (9) G_5's and 9 (7) G_6's as well as 8 D's and 8 B's. There were two corresponding types of minor tone-scramble. Recall that in Experiment 1, the mean pitch-height of a Type 1 (major) tone-scramble was 3.0 as compared to 2.875 for a Type 2 (minor) tone-scramble; the random variations in pitch-height injected into the stimuli in Experiment 2 were more dramatic. The mean pitch-heights of high- and low-pitch-height major, and high- and low-pitch-height minor tone-scrambles were 3.1875, 2.8125, 3.0625, and 2.6875.

2. Participants

One hundred four listeners participated. All were UC Irvine undergraduates with self-reported normal hearing recruited through the UC Irvine School of Social Sciences Research Participation Pool. The sample of listeners tested in Experiment 2 had no overlap with the sample tested in Experiment 1.

3. Testing protocol

The listener wore headphones with volume adjusted to a comfortable level and was then prompted to answer the questions, "How many years of training or serious practice do you have with music?" followed by, "How old were you when you began your musical training?" The listener then was presented with eight example stimuli (visually identified) alternating between major and minor tone-scrambles, including two each of the high- and low-pitch-height major and minor types. The listener was then tested in 4 experimental blocks of 50 trials each. The listener initiated each trial with a button-press; then after the stimulus, the listener pressed "1" on the keyboard for "major" or "2" for "minor" and received visual feedback. Across all 4 blocks there were exactly 50 each of high- and low-pitch-height major and minor tone-scrambles. These 200 stimuli were presented in random order across the 200 trials in the 4 blocks. A small amount of extra incentive was provided in each of the last three blocks. At the start of block k, for $k = 2, 3, 4$, the listener was informed of the proportion correct, p, that he/she had achieved in the previous block. If $p > 0.9$ ($p \leq 0.9$), he/she was further informed that if he/she attained a proportion correct greater than 0.9 (p) in block k, he/she would receive a bonus payment of a quarter at the end of the session.

B. Results

1. Evidence of learning across blocks

Do listeners improve across the four blocks? Any listener whose average performance across all four blocks is near perfect or near chance cannot show substantial improvement. Accordingly, we limit our consideration to the intermediate listeners ($N = 38$) whose average proportion correct across all four blocks was greater than 0.6 and less than 0.9.

For each intermediate listener, let d'_k be the d' achieved by that listener in block $k = 1, 2, 3, 4$, and let μ be the mean of d'_k across all four blocks. Figure 2 plots the mean across all 38 listeners of $d'_k - \mu$ for $k = 1, 2, 3, 4$. This plot reveals a general upward trend in sensitivity across the intermediate listeners. To assess the statistical significance of this trend, we derive the best-fitting (least squares) linear fit to the four values d'_k, $k = 1, 2, 3, 4$ and use the slope, s, of this line as a measure of learning. The sample mean of s across all 38 intermediate listeners was 0.172. That this value is significantly greater than 0 is confirmed by a (1-tailed) t-test:

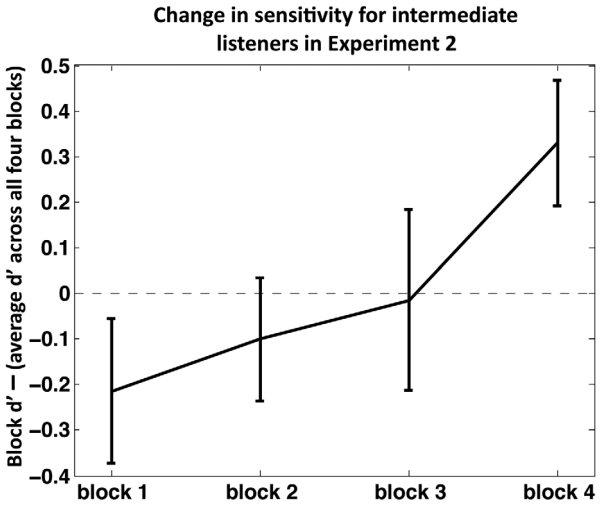

FIG. 2. The change in sensitivity of the 38 intermediate listeners in Experiment 2. A listener in Experiment 2 is classified as intermediate if he/she achieved an average proportion correct greater than 0.6 and less than 0.9 across all four blocks. For each intermediate listener, let d'_k be the d' achieved by that listener in block $k = 1, 2, 3, 4$, and let μ be the mean of d'_k across all four blocks. What is plotted is the mean across all 38 listeners of $d'_k - \mu$ for $k = 1, 2, 3, 4$. Error bars give 95% confidence intervals.

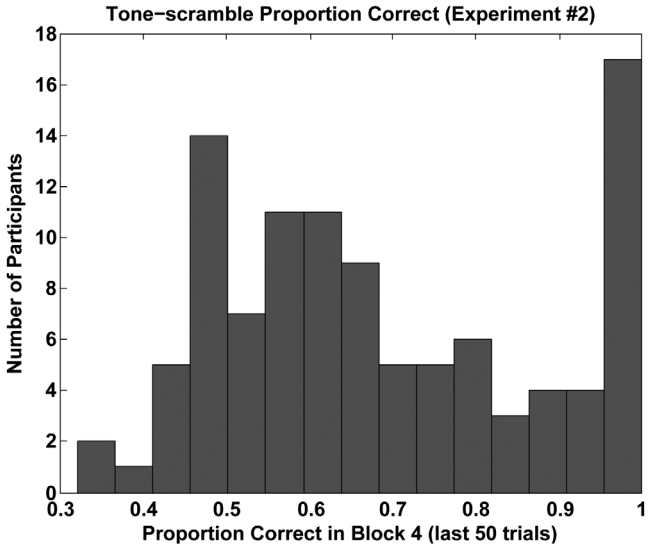

FIG. 3. The distribution of proportion correct on the last 50 trials in Experiment 2. Each bar shows the number of listeners whose score fell within the given bar-bracket. As in Experiment 1, the histogram shows a peak slightly above 0.5 and another peak near 1.0; however, the histogram also shows a substantial proportion of listeners who achieve proportions correct between these two extremes.

$t_{df=37} = 3.843$, $p = 0.00012$. To gauge the size of this effect, suppose a listener started out with $d' = 0$. If this listener improved at the rate of $0.172\,d'$ units per block, his/her d' would be 0.687 at the end of block 4. If this listener used an optimally placed criterion to make his/her judgments, his/her proportion correct would increase from 0.5 to 0.634 across the four blocks.

As Fig. 2 shows, the greatest increase in sensitivity occurred in block 4. Specifically, the average difference $d'_4 - d'_3$ was 0.344. That this jump is significant is confirmed by a t-test: $t_{df=37} = 2.623$, $p = 0.00628$. By comparison, the mean differences $d'_2 - d'_1$ and $d'_3 - d'_2$ were 0.113 and 0.086, neither of which was statistically significant when considered on its own ($t_{df=37} = 1.153$, $p = 0.128$ for $d'_2 - d'_1$, and $t_{df=37} = 0.587$, $p = 0.280$ for $d'_3 - d'_2$).

2. Bimodal histogram of proportion correct

In light of the jump in performance in block 4, we take proportion correct in this block as our measure of performance. Figure 3 shows the histogram of these scores (for all listeners). Although the histogram differs from that in Fig. 1, it also shows two dominant modes, one slightly above 0.5 and one near 1.0. Note that the peak near 1.0 might well aggregate listeners with a broad range of different sensitivities, all sufficiently high to enable near perfect performance.

3. Influence of pitch-height on performance

Are the judgments of our listeners influenced by stimulus pitch-height? We use the following probit model to address this question. Let $A_k = 1$ on trial k if the stimulus was major and -1 if it was minor, and let $B_k = 1$ ($B_k = -1$) if the stimulus had nine (seven) high tonics and seven (nine) low tonics. The model we fit assumes that the listener responds "major" on trial k if

$$\mu(k) + X_k > 0 \quad \text{for}$$
$$\mu(k) = W_A A_k + W_B B_k + W_{AB} A_k B_k + \text{Bias}, \tag{2}$$

where X_k is a standard normal random variable. The model parameter, W_A, reflects the strength with which the listener's response is influenced by the stimulus mode (major vs minor); W_B reflects the strength with which the listener's response is influenced by the pitch-height (high vs low) of the tone-scramble; W_{AB} reflects the influence of a possible interaction between A_k and B_k, and the model parameter, Bias, reflects the listener's baseline response tendency.

For each listener, using the Bayesian procedure described in the Appendix, we attempted to fit the model of Eq. (2) to the data derived from the last 150 trials in Experiment 2. For the 84 listeners who achieved proportions correct less than 0.9 (across the last 150 trials) and for whom stable estimates of the model parameters were available, the four panels of Fig. 4 plot the values of Bias, W_A, W_B, and W_{AB} as a function of the average proportion correct (error bars are 95% Bayesian credible intervals). Filled dots indicate listeners for whom the parameter estimates differ credibly from 0. The solid horizontal line in a given panel indicates the mean ordinate value.

The results for Bias, W_A and W_{AB} make sense. The panel of Fig. 4 for parameter W_A shows, as expected, that higher proportions correct correspond to increased influence of stimulus mode in determining the listener's response. The panel for parameter W_{AB} shows that stimulus mode and stimulus mean pitch-height do not interact systematically to influence judgments. The panel for parameter Bias in Fig. 4 shows an effect that at first seems surprising: although signal detection theory predicts that Bias should be 0 in this task situation, listeners show a baseline tendency to respond "minor" [mean Bias ($= -0.098$) differs significantly from 0;

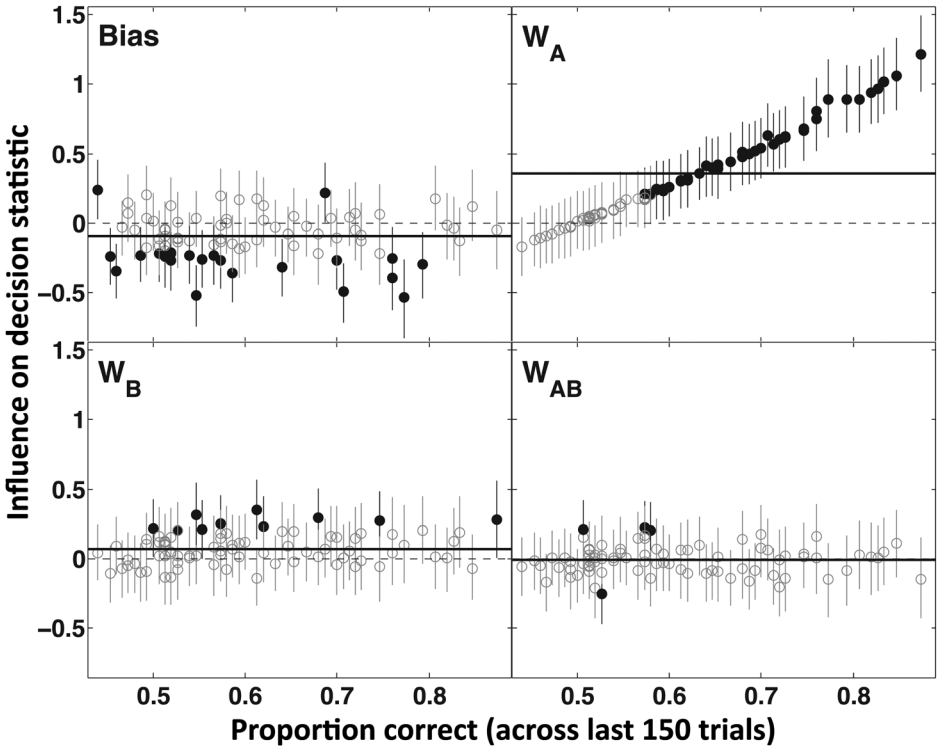

FIG. 4. Influence of pitch-height and mode on judgments in Experiment 2. The abscissa of each panel gives proportion correct across the last 150 trials of Experiment 2. The ordinates of the panels plot the estimated values of the parameters Bias, W_A, W_B, and W_{AB} of the model of Eq. (2) for the 84 listeners who achieved proportions correct less than 0.9 and whose data enabled stable estimates of all four parameters. Error bars are 95% Bayesian credible intervals. Filled circles mark parameter estimates whose credible intervals do not include 0. The solid horizontal line in each panel gives the mean value of the plotted parameter across listeners.

$t_{df=83} = -5.40$, $p < 0.00001$]. Recall, however, that all listeners used the "1" key to respond "major" and the "2" key to respond "minor." A slight preference for the "2" key over the "1" key (e.g., because the "2" key is on the right) might well produce the pattern seen in the Bias panel of Fig. 4.

The lower left panel of Fig. 4 shows that W_B tends to be positive, implying that high (low) mean pitch-height tone-scrambles tend be heard as major (minor). The mean of W_B ($= 0.0763$) differs significantly from 0: $t_{df=83} = 6.17$, $p < 0.00001$. Note that the mean effect of 0.0763 d' units is produced by replacing two low tonics by two high tonics. (This is the difference between high vs low mean pitch-height stimuli.) By comparison, the mean value of W_A ($=0.3573$) is produced by replacing eight minor thirds by eight major thirds. Thus, the effect (in d' units) produced by replacing one low tonic by one high tonic ($=0.0382$) is 85% of the effect produced by replacing one minor third by one major third ($=0.0447$), a strikingly large effect.

Finally, we note that the correlation between W_B and years of musical experience (across the 84 listeners who achieved proportions correct less than 0.9 and whose data enabled stable estimates of all four parameters) was 0.062, which is not significant ($t_{df=82} = 0.5636$, $p = 0.2873$).

Substantial previous research supports the claim that higher pitches tend to be heard as "happier" than lower pitches.[19–22] It thus seems natural that listeners who are explicitly cued to classify tone-scrambles as "HAPPY (major)" vs "SAD (minor)" might be influenced by the mean pitch-height of the scramble as seen in Fig. 4.

4. Contextual influences

On a given trial in Experiment 2, the response of a listener who is highly sensitive to the difference between major vs minor tone-scrambles will be determined predominantly by the mode of the stimulus presented on that trial. However, one might wonder what factors operate to determine the responses of less sensitive listeners. Other than (i) the mode of the current stimulus, the salient first-order features of the response situation are (ii) the mode of the previous stimulus (which the listener can infer from the feedback from the previous trial) and (iii) the response made by the listener on the previous trial. To assess the influence on the listener's current response of these 3 features and their higher-order interactions, we attempt to fit a probit model to the data from the last 150 trials of each listener. Under this model the listener responds major on a given trial, k, if

$$\mu(k) + X_k > 0 \quad \text{for} \quad \mu(k) = W_A A_k + W_C C_k + W_D D_k + W_{AC} A_k C_k + W_{AD} A_k D_k + W_{CD} C_k D_k + W_{ACD} A_k C_k D_k + \text{Bias}, \tag{3}$$

where X_k is a standard normal random variable, A_k [as in Eq. (2)] takes the value 1 (-1) if the mode of the stimulus on trial k is major (minor), $C_k = A_{k-1}$, and D_k takes the value 1 (-1) if the listener's response to stimulus $k-1$ was "major" ("minor"). The model parameters are the constant Bias and the weights W_A, W_C, W_D, W_{AC}, W_{AD}, W_{CD}, W_{ACD} reflecting the influences exerted on the listener's decision statistic by the three first-order features, A, C, D, and their interactions.

Stable estimates of the model parameters were available for 69 listeners. Across this set of listeners, as expected, the mean of W_A ($=0.2130$) was significantly greater than 0 ($t_{df=68} = 7.73$, $p < 0.000\,01$). We also observed a significant mean Bias favoring "minor" responses; however, as

discussed in Sec. III B 3, we suspect that this effect is due to the fact that all listeners pressed the "1" key to respond "major" and the "2" key to respond "minor."

More interestingly, mean W_C (=0.2115) was significantly greater than 0 ($t_{df=68} = 5.58$, $p < 0.000\,01$), implying that these listeners tend to respond on each trial by mimicking the correct response to the previous trial. Why should this be? Suppose that on each trial the listener compares the current stimulus to the previous one, seeking to latch hold of some qualitative difference between the two; if (as is likely to be true for these low-performing listeners) the two stimuli sound the same, then the listener has no option except to issue a response reflecting his/her judgment that the current stimulus is the same type as the last one was.

Two other statistically significant effects remain mysterious: the means of W_{CD} (= −0.1185) and W_{ACD} (=0.0483) both deviate significantly from 0 (for W_{CD}, $t_{df=68} = -6.00$, $p < 0.00001$; for W_{ACD}, $t_{df=68} = 3.49$, $p < 0.0004$). We have no good account of either of these effects.

Finally, we note that although these results provide interesting clues about how low-performing listeners make their judgments, they do not alter the main result that the tone-scramble task yields a cleanly bimodal distribution in performance.

C. Discussion

Experiment 2 corroborates the main finding of Experiment 1: The task of classifying major vs minor tone-scrambles yields a strongly bimodal distribution, with one peak slightly above chance performance and another near perfect performance; however, the histogram does not split the population so cleanly into high- vs low-performing subpopulations. The peak at the high-performance end of the distribution aggregates only 17 listeners, all of whom achieved proportions correct of 0.96 or better throughout blocks 3 and 4 (the last 50 trials). The lower peak seems to be around 0.6; however, the spread of this part of the distribution is broad. Thus, Experiment 2 reveals a substantial population of listeners whose performance is intermediate between chance and perfect.

In addition, an analysis of the 38 listeners who compiled proportions correct (across all 200 trials) between 0.6 and 0.9 revealed that collectively these listeners improved significantly over the course of the four blocks, implying that skill in the classification task is at least partially learnable (at least for some listeners).

One of our motivations for Experiment 2 was to investigate the advantage of the high-performing over the low-performing listeners is driven by heightened sensitivity to tone-scramble mean pitch-height rather than mode. If this were true, then the subpopulation achieving near-perfect performance should have been disrupted in Experiment 2. Although it is possible that some listeners who might otherwise have performed near perfectly were thrown off by the pitch-height variations, the histogram of proportion correct continues to show a prominent peak near perfect performance suggesting that a substantial proportion of listeners who perform well in this task are able to base their judgments on a mode-sensitive statistic distinct from tone-scramble mean pitch-height.

Another motivation for Experiment 2 was to investigate whether low-performing listeners might do better if encouraged to classify tone-scrambles as "happy" vs "sad" instead of as "type 1" vs "type 2"; the current results show no clear evidence of such an effect. We continue to observe a large peak in the histogram near chance performance, and the proportion of listeners who performed well in the task in Experiment 2 is no greater than the proportion who performed well in Experiment 1. However, it is possible that the suppression of performance due to the pitch-height variations introduced in Experiment 2 may have obscured potential benefits in performance due to referring to tone-scrambles as "happy" vs "sad" instead of as "type 1" vs "type 2."

IV. EXPERIMENT 3

Experiments 1 and 2 document that the task of classifying major vs minor tone-scrambles yields a strongly bimodal distribution. In Experiment 3, we ask whether a similar bimodal distribution can be obtained with a task using major vs minor chords instead of tone-scrambles.[23]

A. Method

1. Tasks and stimuli

This experiment involved three tasks. The first was identical to the tone-scramble classification task used in Experiment 2. The stimuli for this task are described in Sec. III A 1. The other two tasks used chords instead of tone-scrambles. Chord stimuli lasted one second and were generated by taking a weighted sum of four tones with pitches G_5, D, G_6, and either B (for major chords) or $B\flat$ (for minor chords). The chords were varied across trials in a quality that we shall call "brightness" by adjusting the relative amplitudes of the low (G_5) and high (G_6) tonics. (This use is nonstandard; typically the term "brightness" is used to refer to a timbric quality of a single-voiced sound.) In the "high-variation" chord classification task, stimulus brightness varied strongly across trials; in the "low-variation" task, stimulus brightness varied much less strongly.

In each of the high- and low-variation tasks, there were five levels of brightness for each of the major and minor chords. Examples of all 20 chord stimuli are provided in http://hdl.handle.net/10575/9881. Suppose the amplitudes of G_5 and G_6 in a given chord are A_5 and A_6, then we shall refer to (i) $A_5 + A_6$ as the combined amplitude of G_5 and G_6 in the chord, and (ii) $\alpha_5 = A_5/(A_5 + A_6)$ and $\alpha_6 = A_6/(A_5 + A_6)$, as the relative amplitudes of G_5 and G_6 in the chord. The five possible values of α_6 (note that $\alpha_5 = 1 - \alpha_6$) in the high-variation task were 0, 0.1875, 0.5, 0.8125, and 1 (corresponding to db differences between A_6 and A_5 of $-\infty$, −12.74, 0, 12.74, ∞); and the five possible values of α_6 in the low-variation task were 0.375, 0.4375, 0.5, 0.5625, 0.625 (corresponding to db differences of −4.44, −2.18, 0, 2.18, 4.44). For a given listener, the combined amplitude of G_5 and G_6 was adjusted to a comfortable level and fixed across all trials in both the high-variation and low-variation conditions. In all chords used in both tasks, the

amplitude of the D (the dominant) and of the B (for major chords) and of the $B\flat$ (for minor chords) was always equal to $(A_5 + A_6)/2$. As in Experiment 2, major and minor chords (and also tone-scrambles) were referred to as "HAPPY (major)" and "SAD (minor)."

2. Participants

Ninety-two listeners participated. All were UC Irvine undergraduates with self-reported normal hearing recruited through the UC Irvine School of Social Sciences Research Participation Pool; one had previously participated in Experiment 1; another had previously participated in Experiment 2.

3. Testing protocol

As in Experiment 2, the listener was first prompted to answer the questions, "How many years of training or serious practice do you have with music?" followed by, "How old were you when you began your musical training?" The listener then completed each of the three tasks. The sequence in which tasks were performed was balanced across listeners with all six permutations of the three tasks occurring approximately equally often.

In each of the three tasks, the listener then was first presented with eight example stimuli (of the type used in the given task) alternating between major and minor.[24]

The listener was then tested in 3 blocks of 40 trials each. In each of the chord tasks, from trial to trial, brightness followed the sequence $1, 4, 2, 5, 3, 1, 4, \ldots$, where 1 stands for the chord (either major or minor) in which the relative amplitude of G_5 was maximal and 5 stands for the chord in which the relative amplitude of G_6 was maximal. This manipulation insured that brightness always differed moderately strongly between successive trials. The listener initiated each trial with a button-press , then heard the stimulus, then pressed "1" for "major" or "2" for "minor" and then received visual feedback. For each of the two chord tasks, across all three blocks there were 12 each of the 10 different types of stimuli. For the tone-scramble task there were 30 each of the 4 different types of stimuli. These 120 stimuli were presented in random order across the 3 blocks.

As in Experiment 2, at the start of block k, for $k = 2, 3$, the listener was informed of the proportion correct, p, that he/she had achieved in the previous block. If $p > 0.9$ ($p \leq 0.9$), he/she was further informed that if he/she attained a proportion correct greater than 0.9 (p) in block k, he/she would receive a bonus payment of a quarter at the end of the session.

B. Results

The histograms of proportions correct in the last 2 blocks of each of the three tasks are shown in Fig. 5. As in Experiments 1 and 2, the histogram for the tone-scramble task is bimodal with distinct peaks near 0.5 and 1. By contrast, neither of the chord tasks yields a strongly bimodal histogram: the high-variation chord task yields a histogram with a peak slightly above 0.5 and what might be construed as a peak around 0.75. However, few listeners achieve proportions correct near 1.0. By contrast, the low-variation chord task yields a histogram with a single dominant peak near perfect performance subsuming nearly half of all listeners, with the remaining listeners spread more or less uniformly between 0.4 and 0.9.

C. Discussion

Our main reason for testing the chord versions of the major/minor classification task was to ascertain whether these more standard stimuli might produce the same bimodal distribution in performance that we observed with the tone-scrambles. With that said, there are many reasons to think that chords might yield different performance in the major/minor classification task than tone-scrambles. There almost certainly exist (i) chord-specific neural mechanisms whose activation requires the simultaneous occurrence of different tones as well as (ii) sequence-specific neural mechanisms activated only by sequential variations in tone. Sensitivity to the difference between major vs minor musical modes might be conferred more or less effectively by mechanisms in these different classes.

Neither of the two variants of the chord task we have tested yields a strongly bimodal distribution in performance across the same set of listeners that show a bimodal

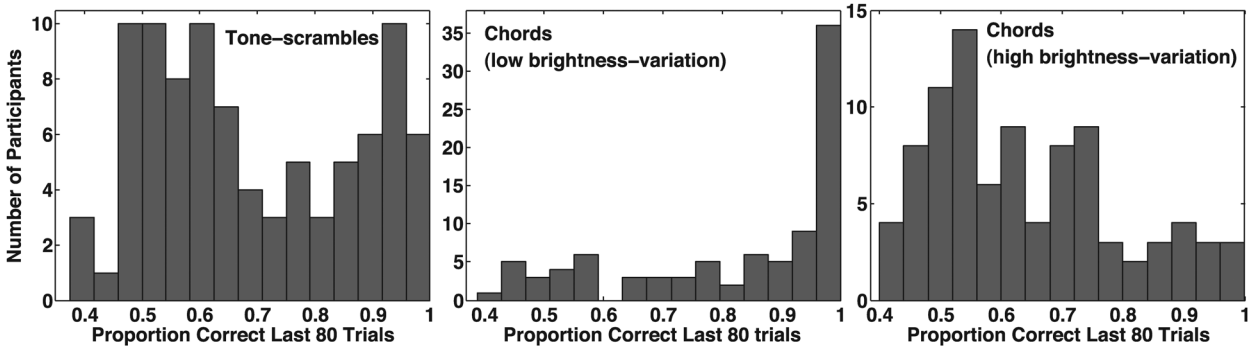

FIG. 5. The distribution of proportion correct on the last 80 trials in all three tasks of Experiment 3. Each bar shows the number of listeners whose score fell within the given bar-bracket. In the tone-scramble task (left histogram), as in Experiments 1 and 2, the histogram shows a peak slightly above 0.5 and another peak near 1.0. This histogram for the low brightness-variation chord classification task (center) shows a peak near 1.0, but few listeners achieve proportions correct near 0.5. The histogram for the high brightness-variation chord classification task (right) shows a peak near 0.5, but few listeners achieve high proportions correct.

distribution in the tone-scramble task. Apparently, the low-variation chord task is too easy and the high-variation chord task is too hard.

It is easy to understand the high performance of many listeners in the low-variation chord task. In the limit, if brightness variation is reduced to 0, then the chord classification task uses only two stimuli, a single major chord and a single minor chord. In this case, the listener can perform perfectly if he/she can differentiate these two stimuli in any way whatsoever. Suppose, for example, that the listener possesses a neuron sensitive to a nonlinear distortion product that is more prominent in the minor than in the major chord. If the listener can base his/her judgments on the responses of this neuron, then he/she should be able to perform perfectly. Crucially, this neuron need not be sensitive either to the "happiness" ("sadness") characteristic of a major (minor) chord. Strategies of this sort are less likely to be available in the tone-scramble task because of the physical differences between stimuli produced by the random sequencing of the tones.

The high-variation chord task is surprisingly challenging. The chords used in the high-variation chord task vary (over five levels) between a root-position triad and an inverted triad. However, all of these chords have the same tonic *G*; all of the major variants should, therefore, be equivalent in their harmonic properties, and the same is true of all of the minor variants. Strikingly, however, most of our listeners were unable to enact response strategies that isolated these harmonic properties unperturbed by the brightness variations with which they were required to contend.

The failures of both the low- and high-variation chord tasks to produce bimodal histograms underscores the delicateness of the qualitative difference isolated by the tone-scramble task. It is, of course, possible that a chord classification task that used a level of brightness variation somewhere between the two levels we have tested might yield a bimodal histogram similar to the histogram observed with the tone-scramble task. However, the current results suggest that (despite what music theory might lead us to expect) brightness variation tends to corrupt the signal we seek to isolate.

This is certainly true for chords; it also seems to be true for tone-scrambles. Although the random pitch-height variations introduced in Experiment 2 were slight, they nonetheless exerted systematic influence on the responses produced by many of our listeners. These pitch-height variations were uncorrelated with the correct response; it follows that they operated in Experiment 2 to suppress the sensitivities of our listeners to the major–minor difference. In fact, of the three tone-scramble tasks we have run, the task used in Experiment 1 yielded the most cleanly bimodal distribution. This may be because this was the only task variant that did not include random variations in mean pitch-height.

V. GENERAL DISCUSSION

The variants of the tone-scramble task used in our three experiments differ in several ways. In the three variants, listeners received different numbers of training trials before testing (45 in Experiment 1, 150 in Experiment 2, and 40 in Experiment 3). They were also tested in different numbers of trials (45 in Experiment 1, 50 in Experiment 2, and 80 in Experiment 3). In addition, the tone-scrambles used in Experiments 2 and 3 were perturbed by random variations in average pitch-height, whereas those in Experiment 1 were not. Finally, bonuses were used to boost incentive in Experiments 2 and 3 but not in Experiment 1.

Despite these differences, it is instructive to look at the histogram of proportions correct achieved by the 275 listeners across all three task variants: Fig. 6. This histogram shows very clearly the bimodality we have seen in the histograms from all of the task variants individually. The question we must now ask is: What does this mean?

A. How are sensitivities to the difference between major vs minor distributed?

Until now we have restricted our consideration to histograms showing distributions of *proportions correct*. Although we expect proportion correct generally to increase with listener sensitivity, proportion correct is not the same thing as sensitivity. A standard measure of sensitivity in a classification task is d'. In the current situation, this statistic is based on a model in which it is assumed that on each trial the listener extracts from the stimulus a statistic S that is corrupted by additive noise, which is assumed to be normally distributed with some standard deviation σ. On trials in which the tone-scramble is minor, S is assumed to have mean μ_1, and on trials in which the stimulus was major, S is assumed to have mean μ_2. It is further assumed that the listener makes his/her decision by comparing S to some fixed internal criterion C (which is typically placed near the midpoint between μ_1 and μ_2). If $S > C$, the listener says "major"; otherwise, he/she says "minor." The listener's overall sensitivity in the task is gauged by $d' = (\mu_2 - \mu_1)/\sigma$ because it is this statistic that determines the potential

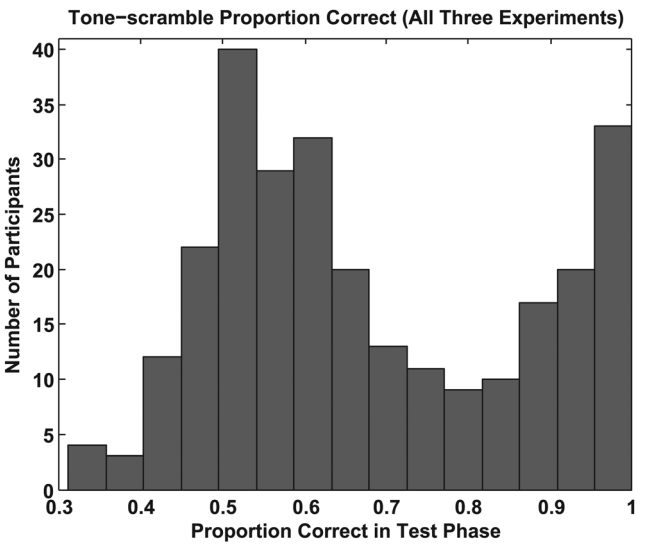

FIG. 6. The distribution of proportion correct pooled across all 275 listeners who participated in the tone-scramble tasks in Experiments 1, 2, or 3. Each bar shows the number of listeners whose score fell within the given bar-bracket. This histogram shows very clearly the bimodality seen in each of Experiments 1, 2, and 3.

effectiveness of the strategies open to the listener. It should be noted, however, that the proportion correct the listener achieves depends not only on his/her value of d', but also on where he/she places the criterion C.

The histogram of d' values achieved by our 275 listeners is shown in Fig. 7. This histogram looks similar to others we have plotted. In particular, it seems to be bimodal with a peak on the left near $d' = 0$ and another on the right. However, the peak on the right is artificial; it collects all listeners with d' levels greater than or equal to 3.5. We impose this cutoff because our data do not allow us to accurately estimate d' values higher than this. It is possible that there does, in fact, exist a high concentration of listeners with d' levels near 4; if so, then the distribution of sensitivities will actually be bimodal as it appears in Fig. 7. Plausibly, however, the actual values of d' of the listeners subsumed in this spike spread broadly between 3.5 and ∞. If so, then the distribution of d' values is not bimodal; rather, it concentrates a large peak of d' values near 0 and spreads out the remainder of d' values in a long positive tail. It should be noted, however, that any d' value greater than around 2 suffices to enable proportion correct of 0.84 or higher.

B. The relation between musical training and sensitivity in the tone-scramble classification task

To what extent does musical training increase sensitivity in the tone-scramble classification task?[25] The current study offers hints but no definitive answer to this question. In Experiment 2, listeners who achieved average proportions correct greater than 0.6 and less than 0.9 showed significant improvement over the four blocks (see Fig. 2). This suggests that, at least for some listeners, sensitivity can be heightened by some sort of training. It should also be noted, however, that 22 out of 104 listeners achieved proportions correct equal to 0.5 or lower in block 4 of Experiment 2 suggesting that there may exist listeners for whom training fails to increase sensitivity.

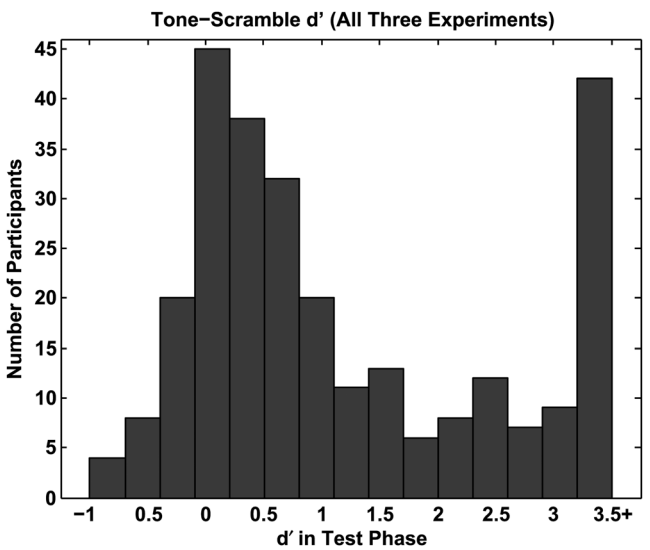

FIG. 7. The distribution of d' values pooled across all 275 listeners who participated in the tone-scramble tasks in Experiments 1, 2, or 3. This histogram seems to be bimodal with a peak on the left near $d' = 0$ and another on the right. However, the peak on the right is artificial; it collects all listeners who achieved d' levels greater than or equal to 3.5. This cutoff is imposed because it is impossible to accurately estimate d' values higher than this from the current data. Plausibly, however, the actual values of d' characterizing the different listeners subsumed in this spike are spread out across a wide range of values between 3.5 and ∞. If so, then the distribution of d' values is not bimodal; rather, it concentrates a large peak of d' values near 0 and spreads out the remainder of d' values in a long positive tail. Note that any d' value greater than around 2 suffices to enable proportion correct of 0.84 or higher.

Further hints are provided by Fig. 8, which plots d' in the tone-scramble task as a function of years of musical training. Each circle corresponds to one of our 275 listeners. The solid line is the best fitting linear regression line. The correlation between years of musical training and d' is 0.488, implying that around 24% of the variance in d' can be accounted for by years of musical training.

It is tempting to conclude that musical training heightens sensitivity in classifying major vs minor tone-scrambles; however, such a conclusion is unwarranted. One can imagine, for example, an alternative scenario in which the sensitivity of most listeners to the difference between major vs minor tone-scrambles is acquired early in life (through some combination of genetic predisposition and early experience) and is invariant with respect to subsequent musical training. Under this scenario, the positive correlation we see in Fig. 8 would be due to the fact that people who are sensitive to the difference between major vs minor modes are more likely than others to seek out musical training.

Indeed, Fig. 8 offers support for the dual claims that musical training is neither a necessary nor a sufficient condition for a listener to have high sensitivity in the tone-scramble classification task. First, there exist listeners in our sample with zero years of musical training who have very high sensitivity in the tone-scramble classification task; second, there exist other listeners in our sample who have many years of musical training whose sensitivity is very near 0.

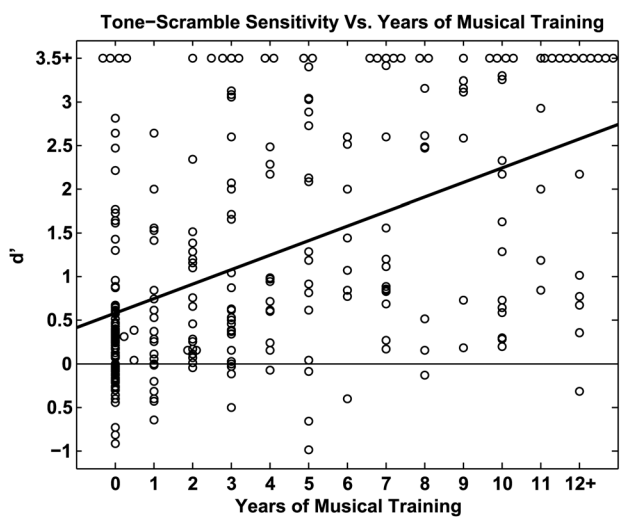

FIG. 8. Scatter plot of d' vs years of musical training pooled over experiments. Each circle corresponds to one of our 275 listeners. The solid line is the best fitting linear regression line. Although there is an obvious concentration of listeners with little or no musical training who also have low values of d', we also find some listeners with many years of musical training who have low values of d' and other listeners with no musical training but very high values of d'.

C. Which stimuli work most effectively to partition the population?

Across all three experiments, Experiment 1 yielded the most cleanly bimodal histogram. As we have noted, the major and minor tone-scrambles used in this experiment differed slightly in pitch-height, making this cue potentially useful for the task, and we cannot rule out the possibility that this cue did in fact help the listeners in Experiment 1. It is clear, however, that this pitch-height cue did not win the game for the large group of listeners who achieved proportions correct near 0.5. Experiment 2 implies, moreover, that it is not this pitch-height cue alone that supports the high levels of performance observed in the upper group of listeners in Experiment 1; even when pitch-height is removed as a useful cue, a large group of listeners is still able to perform the tone-scramble task nearly perfectly.

There is no doubt that the random pitch-height variations injected into the tone-scrambles used in Experiments 2 and 3 did in fact hurt performance. (Whether there were 7 or 9 G_5's in a given tone-scramble was uncorrelated with the correct response, yet this feature of the stimulus exerted a significant influence on responding.) It thus seems likely that the relative messiness [compared to Experiment 1 (Fig. 1)] of the histograms observed in Experiments 2 and 3 (Figs. 3 and 5) is due, at least in part, to the random pitch-height variations used in the tone-scrambles of Experiments 2 and 3.

These observations suggest that the most effective stimuli thus far discovered for partitioning the population into those listeners who can vs those who cannot hear the difference between major vs minor modes are the tone-scrambles used in Experiment 1.

VI. FINAL REMARKS

The current results show that there exist listeners for whom the tone-scramble classification task is very easy; these results also suggest that there may exist some listeners for whom this task is difficult or impossible. How should we interpret this finding? Assuming these two classes of listeners do indeed exist, let us call people who can (cannot) discriminate major vs minor tone-scrambles Ts^+'s (Ts^-'s).

It is possible that Ts^+'s possess a dimension of auditory sensitivity that Ts^-'s lack. Under this scenario, the neuronal system that confers this sensitivity operates in a pervasive fashion to enrich the musical experience of Ts^+'s in comparison to that of Ts^-'s. This would make Ts^-'s (Ts^+'s) analogous to colorblind (color-normal) visual observers. If indeed Ts^-'s (for many of whom music is very important) are impaired in sensing musical qualities that are available to Ts^+'s, they may be limited in their musical pursuits.

This is not the only possibility, however. Different listeners show remarkable agreement in the emotional coloration they ascribe to different musical passages,[26] and assessments of the emotional qualities of music are strongly influenced by a range of different factors including the music's rhythmic properties, the complexity of its harmonies, and also its mode (major vs minor). Most importantly, these emotional qualities seem to be strongly rooted in the hierarchical organization of the music's structure of tension and release, features of the music that are well-described by the tree-diagrams of Lerdahl and Jackendoff.[27] Recent research into the emotional qualities of music has tended to focus on these deeper structural aspects.[28–30]

If, as suggested by Jackendoff and Lerdahl,[28] a listener's understanding of a piece of music is embodied by "the cognitive structures (grouping, metrical, and tonal/reductional) that the listener unconsciously constructs in response to the music," then the emotional qualities of music should be most naturally accessible through structures of this sort as the music summons them into existence within the listener.

From this perspective, tone-scrambles are musically degenerate in the sense that they fail to afford the construction of any cognitive structure that might embody the meaning of the music for the listener. The notes come in a random flurry at a high, constant rate, and then they stop. Perhaps the only special skill that differentiates Ts^+'s from Ts^-'s is the ability to extract certain musical properties from these structureless sequences. To draw an analogy from visual perception, Ts^-'s may be akin to viewers with normal stereo vision who cannot extract the three-dimensional forms that others can see in the autostereograms[31] popularized in the "Magic Eye" books [*Magic Eye New Way of Looking at the World* (Andrew and McMeel, Kansas City, 1993)]. This scenario suggests the possibility that Ts^-'s may be equally sensitive to the entire gamut of emotional qualities that occur in music, but are able to access these qualities only through the cognitive structures that arise within them in response to actual music.

ACKNOWLEDGMENTS

We are grateful to Jon Sprouse and Lisa Pearl for helpful discussions.

APPENDIX

Here we describe the details of the Bayesian method used to estimate the joint posterior density characterizing the parameters W_A, W_B, W_{AB}, and Bias of the model of Eq. (2) as well as the parameters W_A, W_C, W_D, W_{AC}, W_{AD}, W_{CD}, W_{ACD}, and Bias of Eq. (3).

1. The likelihood function

In this paper, we have estimated parameters for two probit models, the models of Eqs. (2) and (3). In each case, some functions $G_j, j = 1, 2, \ldots, N$ are used to gauge different properties of the stimuli experienced by the listener that vary trial by trial. In the model of Eq. (2), these functions are $G_1(k) = A_k$, $G_2(k) = B_k$, $G_3(k) = A_kB_k$, for A_k and B_k given in Sec. III B 3. In the model of Eq. (3), these functions are $G_1(k) = A_k$, $G_2(k) = C_k$, $G_3(k) = D_k$, $G_4(k) = A_kC_k$, $G_5(k) = A_kD_k$, $G_6(k) = C_kD_k$, $G_7(k) = A_kC_kD_k$, for C_k and D_k given in Sec. III B 4. In each case, the specified model stipulates that the listener responds "major" on a given trial, k, if

$$\mu_V(k) + X_k > 0 \quad \text{for} \quad \mu_V(k) = \sum_j W_j G_j(k) + \text{Bias}, \tag{A1}$$

where X_k is a standard normal random variable. For current purposes, it is convenient to subscript "μ" by the vector $V = (W_1, W_2, \ldots, W_N, \text{Bias})$ that keeps track of the values of the model parameters. Thus, given a particular assignment of values to the coordinates of V, the probability of the response R_k on trial k is

$$P_V(k) = \begin{cases} \Phi(\mu_V(k)), & \text{if } R_k = \text{"major"} \\ 1 - \Phi(\mu_V(k)), & \text{if } R_k = \text{"minor"} \end{cases} \quad \text{(A2)}$$

for Φ the standard normal cumulative distribution function, and the likelihood function is

$$\Lambda(V) = \prod_{\text{all trials } k} P_V(k). \quad \text{(A3)}$$

2. Markov chain Monte Carlo simulation

The estimation method uses Markov chain Monte Carlo (MCMC) simulation. For simplicity, we use uniform prior distributions on all parameters. In any MCMC process, one starts with some arbitrary guess at the parameter vector V (which will ultimately be thrown away) and sets ${}_1S = V$. [Note: (i) We use pre-subscripts to indicate parameter vector sample number in the MCMC process; (ii) In the current applications of this method, V comprises guesses at the parameters $W_1, W_2, \ldots, W_N$ and Bias.] Then, one iterates the following steps some large number, N_{iter}, of times:

(1) Pick a candidate parameter vector, C, in the neighborhood of the last sample, ${}_{n-1}S$, added to the list. Then
(2) for[32]

$$R = \frac{\Lambda(C)}{\Lambda({}_{n-1}S)}, \quad \text{(A4)}$$

if $R \geq 1$, set ${}_nS = C$; otherwise, set

$${}_nS = \begin{cases} C & \text{with probability } R \\ {}_{n-1}S & \text{with probability } 1 - R. \end{cases} \quad \text{(A5)}$$

The classical result[34] is that (provided that the procedure for selecting candidates C satisfies certain conditions) in the limit as $N_{\text{iter}} \to \infty$ this algorithm yields a sample from the posterior density. In practice, one typically throws away the first several thousand samples from the list which are usually not representative of the samples accumulated after the MCMC process has stabilized.

3. Priors

The bounds of the uniform priors matter very little provided they are wide enough to include the posterior density. In the current simulations, the prior density of each coordinate of V was taken to be uniform between -10 and 10.

4. Adaptive candidate selection

The sampling window used to select the candidate parameter vector C on each iteration of the MCMC process dramatically influences the efficiency with which one can estimate the posterior joint density of the parameters. We adjust this sampling window adaptively after each 1000 iterations of the MCMC process. Specifically, let S_{Last1000} be the matrix whose columns are the 1000 most recent parameter vectors added to the list by the MCMC process. We first subtract the mean of these 1000 parameter vectors from each vector in S_{Last1000} to generate a matrix Δ_{Last1000}. We use singular value decomposition to extract (i) the matrix Q whose columns are the (orthonormal) principal components of Δ_{Last1000} as well as (ii) the diagonal matrix E whose kth diagonal entry is the eigenvalue of the kth column of Q. In each of the subsequent 1000 iterations of the MCMC process, we draw each successive candidate parameter vector, C, by setting $C = {}_{n-1}S + QEX$, where X comprises a vector of independent normal random variables with mean 0 and standard deviation $1/30$. In essence, we use the last 1000 parameter vectors to approximate the posterior density as an elliptical cloud, and take steps scaled to the axes of this cloud. This method succeeds in achieving an MCMC process that moves efficiently to scribble in the joint posterior density.

5. Starting values, burn-in, and number of iterations

In the current application, all values of ${}_1S$ are initialized to 0. Thirteen thousand iterations of the MCMC process were performed, the first 3000 of these were used to allow the MCMC process to "burn in," and the last 10 000 were taken as a representative sample of the posterior joint density characterizing the parameters W_j, $j = 1, 2, \ldots, N$ and Bias. In each case, the last 10 000 samples were plotted and inspected by eye to insure stability of parameter estimates.

[1]R. G. Crowder, "Perception of the major/minor distinction: I. Historical and theoretical foundations," Psychomusicology **4**(1/2), 3–12 (1984).
[2]R. G. Crowder, "Perception of the major/minor distinction: II. Experimental investigations," Psychomusicology **5**(1/2), 3–24 (1985).
[3]R. G. Crowder, "Perception of the major/minor distinction: III. Hedonic, musical, and affective discriminations," Bull. Psychon. Soc. **23**(4), 314–316 (1985).
[4]S. Dalla Bella, I. Peretz, L. Rousseau, and N. Gosselin, "A developmental study of the affective value of tempo and mode in music," Cognition **80**, B1–B10 (2001).
[5]L. Gagnon and I. Peretz, "Mode and tempo relative contributions to 'happy-sad' judgments in equitone melodies," Cognit. Emotion **17**(1), 25–40 (2003).
[6]G. M. Gerardi and L. Gerken, "The development of affective responses to modality and melodic contour," Music Percept. **12**(3), 279–290 (1995).
[7]C. P. Heinlein, "The affective character of the major and minor modes in music," J. Comp. Psychol. **VIII**(2), 101–142 (1928).
[8]K. Hevner, "The affective character of the major and minor mode in music," Am. J. Psychol. **47**, 103–118 (1935).
[9]M. P. Kastner and R. G. Crowder, "Perception of the major/minor distinction: IV. Emotional connotations in young children," Music Percept. **8**(2), 189–202 (1990).
[10]R. Whissell and C. Whissell, "The emotional importance of key: Do Beatles' songs written in different keys convey different emotional tones?," Percept. Mot. Skills **91**, 973–980 (2000).
[11]D. Temperley and D. Tan, "Emotional connotations of diatonic modes," Music Percept. **30**(3), 237–257 (2013).
[12]A. R. Halpern, "Perception of structure in novel music," Mem. Cognit. **12**, 163–170 (1984).
[13]A. R. Halpern, J. C. Bartlett, and W. J. Dowling, "Perception of mode, rhythm, and contour in unfamiliar melodies: Effects of age and experience," Music Percept. **15**, 335–356 (1998).
[14]A. M. Leaver and A. R. Halpern, "Effects of training and melodic features on mode perception," Music Percept. **22**, 117–143 (2004).

[15]M. J. Blechner, "Musical skill and the categorical perception of harmonic mode," Haskins Laboratories Status Report on Speech Perception (SR-51/52), pp. 139–174 (1977).
[16]N. Saint-Arnaud and K. Popat, "Analysis and synthesis of sound textures," in *Proceedings of the AJCAI Workshop on Computational Auditory Scene Analysis* (1995), pp. 293–308.
[17]J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," Neuron **71**(5), 926–940 (2011).
[18]S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," Ann. Math. Stat. **9**, 60–62 (1938).
[19]K. Hevner, "The affective value of pitch and tempo in music," Am. J. Psychol. **49**, 621–630 (1937).
[20]K. B. Watson, "The nature and measurement of musical meanings," Psychol. Monogr. **54**, 1–43 (1942).
[21]S. E. Trehub, "The music listening skills of infants and young children," in *Psychology and Music: The Understanding of Melody and Rhythm*, edited by W. J. Dowling and T. J. Tighe (Lawrence Erlbaum Associates, Hillsdale, NJ, 1993), Chap 7, pp. 161–176.
[22]W. G. Collier and T. L. Hubbard, "Judgments of happiness, brightness, speed and tempo change of auditory stimuli varying in pitch and tempo," Psychomusicology **17**, 36–55 (2001).
[23]The reader may wonder what would happen to performance in the major/minor classification task if the tone-scramble tonic were roved randomly from trial to trial. Pilot studies show that this makes the task much harder. None of the listeners we tested performed better in the roved-tonic version of the task; many who are near perfect in the basic tone-scramble task were severely impaired in the roved-tonic condition. Part of the reason for the increased difficulty of the roved-tonic condition may have to do with the fact that in this condition the tonic needs to be established anew in each tone-scramble; if tones at the third degree of the scale register as such only after the tonic is established, then these critical tones may exert little impact on the listener's decision if they occur early in the tone-scramble (before the tonic has been established). By contrast, in the un-roved condition, the tonic persists across trials; thus, tones at the third degree of the scale can exert powerful impact even when they occur very early in the stimulus.
[24]For chords, these eight example stimuli consisted of (i) a major chord followed by (ii) a minor chord, both of the lowest of the five possible brightnesses, followed by (iii) a major chord followed by (iv) a minor chord, both of the second-highest of the five possible brightnesses, followed by (v) a major chord followed by (vi) a minor chord, both of the second-lowest of the five possible brightnesses, followed by (vii) a major chord followed by (viii) a minor chord, both of the highest of the five possible brightnesses. For tone-scrambles, as in Experiment 2, the eight example stimuli included two each of all four types of tone-scrambles (high and low pitch-height major and minor) alternating between major and minor.
[25]Previous studies have addressed similar questions by comparing the performance of trained musicians in various auditory tasks to that of nonmusicians. For example, Spiegel and Watson (Ref. 33) documented that trained musicians performed better (on average) than a matched sample of nonmusicians on several different tasks requiring the listener to discriminate the pitches of notes. Similarly, McDermott *et al*. (Ref. 35) found that trained musicians were better than nonmusicians at several different sorts of tasks requiring discrimination of note pitch, note loudness, note brightness as well as discrimination of two-note intervals in these different note properties. If, however, the sensitivity required to perform well in the tone-scramble classification task is (i) unlearnable for some people, and (ii) important for musical performance, then listeners who are highly sensitive to the difference between major and minor tone-scrambles will be more likely to proceed to higher levels of training than insensitive listeners. Under this scenario, the strategy of directly comparing the sensitivities of nonmusicians vs trained musicians might well yield a strong positive correlation between musical training and sensitivity in the tone-scramble task even though no causal relationship exists in fact.
[26]K. Hevner, "Experimental studies of the elements of expression in music," Am. J. Psychol. **48**, 246–268 (1936).
[27]F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music* (MIT Press, Cambridge, MA, 1983).
[28]R. Jackendoff and F. Lerdahl, "The capacity for music: What is it, and what's special about it?," Cognition **100**, 33–72 (2006).
[29]F. Lerdahl, "Genesis and architecture of the GTTM project," Music Percept. **26**(3), 187–194 (2009).
[30]C. L. Krumhansl, "Music: A link between cognition and emotion," Curr. Dir. Psychol. Sci. **11**(2), 45–50 (2002).
[31]C. W. Tyler and M. B. Clarke, "The autostereogram," Proc. SPIE **1258**, 182–196 (1990).
[32]If the prior density, f_{prior}, were nonuniform, then we would have $R = [\Lambda(C)f_{\text{prior}}(C)]/[\Lambda(_{n-1}S)f_{\text{prior}}(_{n-1}S)]$.
[33]M. F. Spiegel and C. S. Watson, "Performance on frequency-discrimination tasks by musicians and nonmusicians," J. Acoust. Soc. Am. **76**(6), 1690–1695 (1984).
[34]W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**(1), 97–109 (1970).
[35]J. H. McDermott, M. V. Keebler, C. Micheyl, and A. J. Oxenlam, "Musical intervals and relative pitch: Frequency resolution, not interval resolution, is special," J. Acoust. Soc. Am. **128**(4), 1943–1951 (2010).
[36]Although the histogram shown in Fig. 1 appears strongly bimodal, one might wonder whether this impression would be confirmed by a formal statistical test. We performed such a test. Specifically, we used a likelihood ratio test to compare the fits provided to these data by two nested models. The restricted (unimodal) model assumed that the number of correct responses for each listener is a binomial random variable with parameters p and $n = 45$, where p is a random variable drawn from a beta density $f_{z,w}(p)$. (By varying z and w, $f_{z,w}$ can flexibly capture a wide range of different unimodal density functions defined on the interval [0, 1].) The fuller (bimodal) model assumed that p is drawn from a mixture of two beta densities. The fit provided by the fuller model must be better than that provided by the restricted model; the likelihood ratio test assesses whether the fit provided by the fuller model is significantly better than would be expected by chance under the null hypothesis that the restricted model is true. This test rejects the null hypothesis with vanishingly small p-value for the data shown in Fig. 1. However, it also rejects the null hypothesis with very small p-values for all of the other data sets we report, even those whose histograms appear much more unimodal than Fig. 1 (e.g., those in the right two panels of Fig. 5). We conclude that none of our data sets are actually unimodal. The mixture parameter in the fuller model reflects the strength of the data's bimodality, with small values indicating little need for a second mode. We do not bother to report these values, however, because they merely confirm what is already clear from the histograms.