Title
Inadequate pitch-difference sensitivity prevents half of all listeners from discriminating major vs minor tone sequences.

Journal
The Journal of the Acoustical Society of America, 151(5)

ISSN
0001-4966

Authors
Ho, Joselyn
Mann, Daniel S
Hickok, Gregory
et al.

Publication Date
2022-05-01

DOI
10.1121/10.0010161

Peer reviewed

JASA ARTICLE

Inadequate pitch-difference sensitivity prevents half of all listeners from discriminating major vs minor tone sequences

Joselyn Ho, Daniel S. Mann, Gregory Hickok, and Charles Chubb[a)]
Department of Cognitive Sciences, University of California Irvine, Irvine, California 92617, USA

ABSTRACT:
Substantial evidence suggests that sensitivity to the difference between the major vs minor musical scales may be bimodally distributed. Much of this evidence comes from experiments using the "3-task." On each trial in the 3-task, the listener hears a rapid, random sequence of tones containing equal numbers of notes of either a *G* major or *G* minor triad and strives (with feedback) to judge which type of "tone-scramble" it was. This study asks whether the bimodal distribution in 3-task performance is due to variation (across listeners) in sensitivity to differences in pitch. On each trial in a "pitch-difference task," the listener hears two tones and judges whether the second tone is higher or lower than the first. When the first tone is roved (rather than fixed throughout the task), performance varies dramatically across listeners with median threshold approximately equal to a quarter-tone. Strikingly, nearly all listeners with thresholds higher than a quarter-tone performed near chance in the 3-task. Across listeners with thresholds below a quarter-tone, 3-task performance was uniformly distributed from chance to ceiling; thus, the large, lower mode of the distribution in 3-task performance is produced mainly by listeners with roved pitch-difference thresholds greater than a quarter-tone. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0010161

(Received 27 November 2021; revised 17 March 2022; accepted 24 March 2022; published online 11 May 2022)
[Editor: James F. Lynch] Pages: 3152–3163

I. INTRODUCTION

As emphasized by theories of Western music composition, the qualities that music can achieve by variations in musical scale are central to its meaning (Rameau, 1971; Schoenberg, 1978; Tymoczko, 2011). For example, on average, listeners tend to hear music in the major scale as sounding "happy" and in the minor scale as sounding "sad" (e.g., Blechner, 1977; Bonetti and Costa, 2019; Crowder, 1985a; Crowder, 1985b; Cunningham and Sterling, 1988; Bella *et al.*, 2001; Gerardi and Gerken, 1995; Heinlein, 1928; Hevner, 1935; Kastner and Crowder, 1990; Leaver and Halpern, 2004; Peretz *et al.*, 1998; Temperley and Tan, 2013).

However, there are many indications that listeners are less sensitive to the difference between music in the major and minor scales than one might expect given their central role in music theory and composition. First, many listeners find it surprisingly difficult to discriminate major vs minor melodies (Halpern, 1984; Halpern *et al.*, 1998; Leaver and Halpern, 2004). Other studies suggest that the distribution of sensitivity to major vs minor triadic chords is bimodally distributed across listeners (Blechner, 1977; Crowder, 1985a) with some listeners highly sensitive to the difference and others showing little or no sensitivity. Consistent with this finding, studies that have been careful to isolate effects of scale on judgments of musical affect from other aspects of musical structure (e.g., tempo) have typically found mean effects that are statistically significant but modest in size (in line with the idea that the mean is elevated above chance by a minority of highly sensitive listeners).

As reviewed in Sec. I A, additional evidence for a bimodal distribution in sensitivity to the difference between the major and minor scales comes from experiments using rapid, random sequences of tones ("tone-scrambles"; Chubb *et al.*, 2013; Dean and Chubb, 2017; Mednicoff *et al.*, 2018; Ho and Chubb, 2020). Nonetheless, tone-scrambles differ from actual music in several important ways: they are faster than nearly all music (923 BPM), higher in pitch than most music, and composed of pure tones. This raises the possibility that the bimodal distribution in sensitivity to major vs minor tone-scrambles may not generalize to actual music.

Moreover, although previous findings based on more musical stimuli are largely consistent with the proposal that sensitivity to scale variations is bimodally distributed, at least several studies seem to suggest that, on the contrary, sensitivity to the difference between music in the major and minor scales is nearly universal (Bonetti and Costa, 2019; Temperley and Tan, 2013). These two studies differ from other studies in important ways, however. Most studies investigating the qualities evoked by major and minor music present listeners on a given trial with a single musical segment and ask them to judge the emotional quality produced by the segment. For example, in the study by Hevner (1935), a given listener heard a given musical segment only once. Some listeners heard the major version of the segment; other listeners heard the minor version. This is not true in the studies by Bonetti and Costa (2019) and Temperley and Tan (2013). In the study by Temperley and Tan (2013), the listener was presented with two versions of the same melody, each with the same tonic but with the key signature altered to change the scale between the two presentations; the listener then judged which sounded "sadder."

[a)]Electronic mail: cfchubb@uci.edu

 0001-4966/2022/151(5)/3152/12/$30.00

https://doi.org/10.1121/10.0010161

In the study by Bonetti and Costa (2019), the listener heard major and minor versions of the same segment multiple times across multiple trials in several different conditions. We shall return to this issue in Sec. IV (General Discussion).

In Sec. IV, we shall argue that sensitivity to the difference between actual music in the major and minor scales is bimodally distributed, yet, our immediate goal is more modest. Here, we ask, what is the source of the bimodal distribution in sensitivity to major vs minor tone-scrambles?

In particular, we investigate whether sensitivity to the difference between major and minor tone-scrambles depends on basic sensitivity to differences in pitch. We will probe this question by testing listeners (1) in a task that requires them to classify tone-scrambles as major vs minor and (2) in tasks that require them to judge the direction of the pitch-difference between two successive tones. Our results reveal that performance in the major-minor task depends critically on performance in the pitch-difference tasks: only listeners whose pitch-difference sensitivity exceeds a specific baseline are able to hear the difference between major and minor musical stimuli.

Many studies have observed positive correlations between musical training and cognitive target-skills unrelated to music, leading some researchers to propose that music training can heighten these target-skills. For example, it has been proposed that musical training can heighten language skills (Patel, 2011, 2014; Kraus and Chandrasekaran, 2010; Kraus *et al.*, 2014). Other work emphasizes the need to consider the alternative possibility that (1) people imbued with high levels of certain musically relevant processing resources (which are immune to musical training) may be more likely to seek out musical training than other people, and (2) these same processing resources also contribute to target-skills unrelated to music (e.g., language skills; Swaminathan and Schellenberg, 2020; Kragness *et al.*, 2021). The latter scenario is likely to lead to a positive correlation between music training and the target-skills; however, under this scenario, the target-skills are immune to musical training.

The results reported here contribute to this discussion by singling out pitch-difference sensitivity as a processing resource that is important for musical skill and may be important for nonmusical skills as well (e.g., skills related to speech-processing). Moreover, as we shall show, this resource varies dramatically across subjects. Whether or not pitch-difference sensitivity is immune to training remains an open question.

A. Sensitivity to major and minor tone-scrambles

The chromatic scale contains 12 notes, each separated by a semitone (100 cents) from its neighbors. Each of the major and minor diatonic scales contains seven notes drawn from the chromatic scale. If we call the notes of the chromatic scale $c_1, c_2, \ldots, c_{12}$, then the major scale includes notes c_1, c_3, c_5, c_6, c_8, c_{10}, and c_{12}. These notes are called "degrees" 1,2,…,7 of the major scale. There are several variants of the minor scale. The seven degrees of the "natural" minor scale are c_1, c_3, c_4, c_6, c_8, c_9, and c_{11}. The "descending melodic minor" scale is identical to the natural minor scale; however, the "ascending melodic minor" scale has c_{10} and c_{12} as degrees 6 and 7 instead of notes c_9 and c_{11}. The "harmonic minor" scale is identical to the natural minor scale except that it has c_{12} as degree 7 instead of c_{11}.

Thus, at the core of the difference between the major and minor scales is the triad composed of the scale degrees 1, 3 and 5. Degrees 1 and 5 are crucial for establishing the context within which variations in other scale degrees influence scale-defined qualities. Among all seven scale degrees, degree 3 is unique in the following respect: it alone differs in the major scale and in all common variants of the minor scale; major-scale degree 3 is c_5 and minor-scale degree 3 is c_4.

Therefore, one might expect the qualitative difference between the major and minor scales to be vividly expressed by the major and minor stimuli used in the "3-task" (Adler *et al.*, 2020; Chubb *et al.*, 2013; Dean and Chubb, 2017; Ho and Chubb, 2020; Mednicoff *et al.*, 2018). In this task, stimuli are rapid (923 BPM), randomly ordered sequences of pure tones. The major and minor stimuli contain eight each of the notes G_5, D_6 (degree 5 of the G major and minor scales), and G_6. The purpose of these 24 context notes is to establish G firmly as tonic on every trial. In addition, major stimuli contain eight $B\natural_5$'s (degree 3 of the G major scale), whereas minor stimuli contain eight $B\flat_5$'s (degree 3 of the G minor scale). On each trial, the listener hears a single stimulus and strives (with feedback) to classify it as major or minor. An example of a major (minor), 3-task tone-scramble is provided in Mm. 1 (Mm. 2).

Mm. 1. Example of a major tone-scramble from the 3-task. This is a file of type "wav" (204 KB).

Mm. 2. Example of a minor tone-scramble from the 3-task. This is a file of type "wav" (204 KB).

Surprisingly, as shown in Fig. 1, the 3-task yields a dramatic, bimodal distribution in performance: approximately 70% of listeners perform near chance while the remaining 30% perform near ceiling.

B. What is the source of the bimodal distribution in 3-task performance?

Previous research has ruled out several possible explanations of the bimodal distribution shown in Fig. 1.

First, the difference between high- and low-performers is not due to musical training. A scatterplot relating years of musical training to sensitivity in the 3-task (as gauged by d') is shown in Fig. 2. Although years of musical training is positively correlated with 3-task-d', this correlation is driven mainly by a large group of listeners with no training who perform poorly in the 3-task. Strikingly, we also observe a large number of listeners with many years of musical

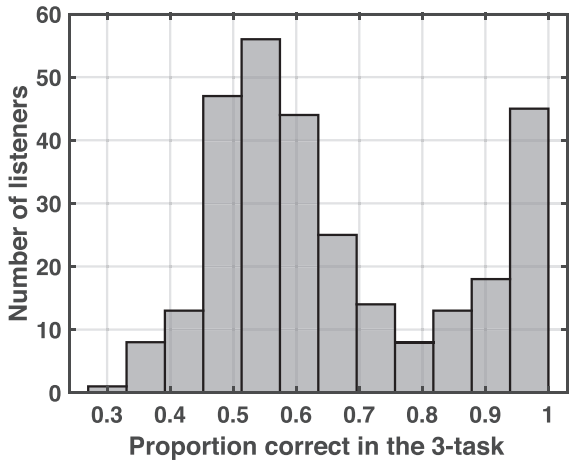

FIG. 1. The histogram of the proportions correct achieved in the 3-task by listeners pooled from the experiments of Chubb *et al.* (2013), Dean and Chubb (2017), and Mednicoff *et al.* (2018). All proportions are based on 50 trials (with trial-by-trial feedback), and in each case, these test trials were preceded by at least 40 practice trials (also with feedback).

training whose d' values are near zero, and other listeners with little or no training whose d' values are close to ceiling.

These findings are in accord with the proposal that the positive correlations observed between years of musical training and various musical abilities may reflect preexisting differences that make listeners who are high in ability more likely to pursue music lessons than listeners who are low in ability (Kragness *et al.*, 2021). In addition, the finding of Adler *et al.* (2020) that 6-month-old infants show the same bimodal distribution in 3-task performance as adults suggests that the sensitivity underlying performance in this task may be largely determined very early in life.

Second, one might speculate that high- and low-performers do not differ in their sensitivity to the difference between major and minor music but only in their ability to extract these qualities from the very rapid sequences of notes (923 BPM) presented in tone-scrambles. If so, then the lower mode of the bimodal distribution shown in Fig. 1

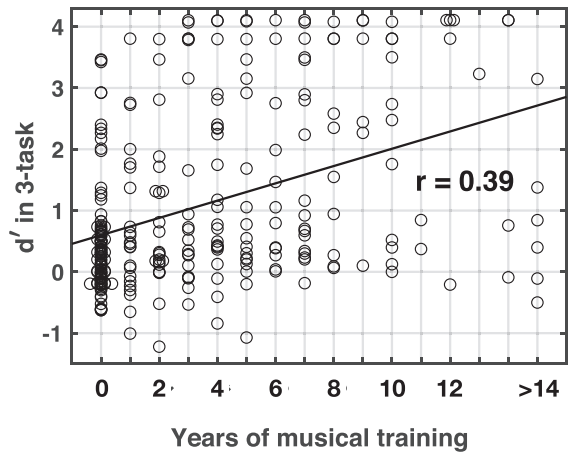

FIG. 2. Scatterplot of years of musical training vs d' in the 3-task (results pooled from Chubb *et al.* (2013), Dean and Chubb (2017) and Mednicoff *et al.* (2018)).

should vanish if the stimuli are slowed down. This does not happen. On the contrary, the task becomes even more challenging for low-performers when stimuli comprise just four notes (one each of the notes G_5, D_6, G_6 and either a single B_5 or single $B\flat_5$) presented in random order, each for 520 ms (Mednicoff *et al.*, 2018).

C. The current project

This study investigates the possibility that individual differences in basic pitch-processing ability play a role in producing the bimodal distribution in 3-task performance (Fig. 1). In particular, we focus on the relationship between performance in the 3-task and performance in "roved pitch-difference" (RPD) tasks. In a RPD task, the listener hears two pure tones on each trial; the first tone is chosen randomly from a large range of frequencies, and the task is to judge whether the second tone is higher or lower than the first.

Building on previous studies focused on listeners with cortical lesions (Johnsrude *et al.*, 2000; Tramo *et al.*, 2002), Semal and Demany (2006) showed that there exist listeners with otherwise normal hearing for whom RPD tasks are highly challenging for the following, unexpected reason: although they can tell when the two tones in a given trial are different, these listeners are markedly impaired at discerning the direction of the difference. In the main experiment of Semal and Demany (2006), the listener heard two pairs of pure tones on each trial. In one pair, the tones were identical; in the other pair, the tones differed in frequency. In the "detection" task, the listener judged which tone-pair contained the change (without reporting the direction of the change). In the "identification" task, the listener judged the direction of the change (without reporting which pair contained the change). Semal and Demany (2006) demonstrated that for some listeners (whose hearing was otherwise normal), the threshold frequency difference for the identification task was substantially higher than the threshold difference for the detection task. Mathias *et al.* (2010) replicated the experiment of Semal and Demany (2006) and showed, in addition, that the difficulties experienced by such "direction-challenged" listeners are (1) are greatly decreased if the first tone is fixed across trials (i.e., if the rove is removed), and (2) most dramatic when the first tone is roved across a very wide range of frequencies.

In the current study, the main purpose of experiment 1 was to determine whether performance in the 3-task is related to performance in a RPD task.

Some terminology related to pitch-difference tasks will be useful. We will be consistent in using the symbols ϕ_1 and ϕ_2 for the frequencies of the first and second tones, respectively, presented on a trial in a pitch-difference task. We will use the term "pitch-difference magnitude" to refer to the absolute difference, d, in cents between ϕ_2 and ϕ_1, i.e., $d = 1200|\log_2(\phi_2/\phi_1)|$. We will use the term "roved pitch-difference threshold" (rPDT) to refer to the pitch-difference magnitude required for a given listener to achieve 80%

https://doi.org/10.1121/10.0010161

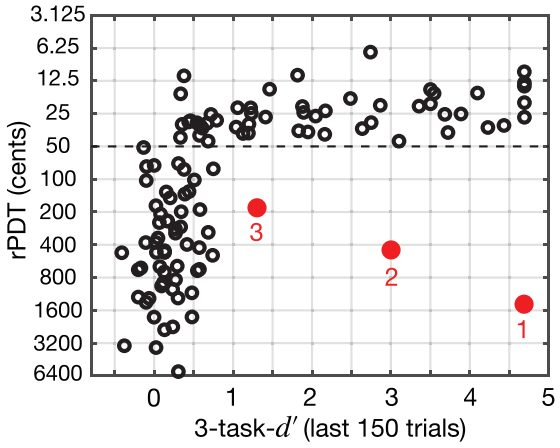

FIG. 3. (Color online) The scatterplot of rPDT as a function of 3-task-d'. The rPDTs are plotted (on a log scale) with the values decreasing from bottom to top to reflect increasingly good performance. The dashed line is at 50 cents (a quarter-tone). Out of the 59 listeners whose rPDTs were higher than 50 cents, only pattern-breakers #1, #2, and #3 (the filled red circles) achieved d' values greater than 0.75 (which corresponds to the proportion correct ≤ 0.65) in the 3-task.

correct responses. (We will use the lower case "r" as a prefix to "PDT" to signal that the threshold comes from a RPD task; we drop the "r" only for the thresholds from the fixed pitch-difference task in experiment 2.) It is important to bear in mind that *low* rPDTs indicate *high* performance in a RPD task. We will emphasize this in Figs. 3, 5, 7, and 8 by associating decreasing rPDTs with vertically increasing y axis locations. Thus, in all of our plots, higher values on the y axis correspond to higher performance.

A central finding of experiment 1 is that in the RPD task, many listeners (approximately half of those tested) had rPDTs $>$ 50 cents (i.e., a quarter-tone, half the distance in log-frequency between successive notes of the chromatic scale). Strikingly, nearly all of these listeners performed near chance in the 3-task. By contrast, performance in the 3-task is approximately uniformly distributed from chance to ceiling across the listeners with rPDTs $<$ 50 cents. The results of experiment 1 led us to wonder what features of the RPD task used in experiment 1 were important for producing this pattern. In experiment 2, to probe this issue,

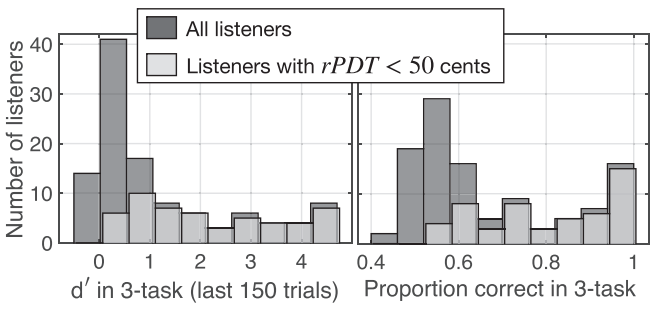

FIG. 4. The histograms of d' values in the 3-task in Experiment 1 (across the last 150 of 200 trials). The dark gray bars show the histogram for all listeners. The light gray bars show the histogram for only those listeners who achieved rPDTs lower than 50 cents.

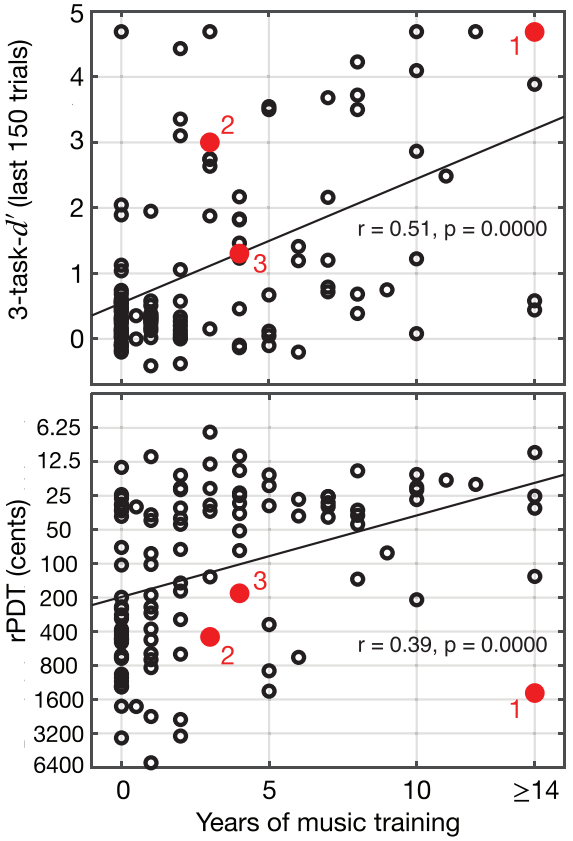

FIG. 5. (Color online) The relationship between years of musical training and 3-task-d' (top) and rPDT (bottom). The large red dots indicate pattern-breakers #1, #2 and #3 from Fig. 3. Pattern-breaker #1 had 15 yr of musical training.

we tested a new group of 99 listeners in several different pitch-difference tasks (as well as the 3-task).

II. EXPERIMENT 1

A. Methods

All of the methods were approved by the University of California Irvine (UCI) Institutional Review Board.

1. Participants

112 undergraduate students were recruited from the Social Science Human Subjects Pool at UCI. Participation in the experiment was awarded with extra credit applied to one of their courses. One of these listeners was later dropped from the study because this listener responded incorrectly on 75% of the trials in the RPD task; we took this to indicate that this listener was making no effort in the RPD task. (This was more than double the number of errors made by any other listener.)

2. Procedure

Each listener was tested in four tasks in random order: a 3-task, pitch-difference task, pitch memory task, and the Scale-Violated Melody-Comparison Task from the Montreal Battery of Evaluation of Amusia (MBEA; Peretz *et al.*, 2003). The results for only the 3-task and pitch-

difference task are reported here. The results from all four tasks are described in Mann (2014). Prior to testing, each listener completed a brief survey to report (among other information) their years of musical experience.

The experiment took place in a quiet laboratory on a Dell computer (Dell Technologies) running Windows (Microsoft Corporation) with a standard Realtek (Sony Corporation) audio/soundcard using MATLAB (The MathWorks, Inc.). The stimuli were presented at the rate of 50 000 samples/s, and listeners wore JBL Elite 300 noise-canceling headphones (JBL, Inc.) with the volume adjusted to their comfort level.

3. 3-task

a. Stimuli. The stimuli were tone-scrambles. Each tone-scramble contained eight copies each of the following notes from the standard equal-tempered chromatic scale: G_5 (783.99 Hz), D_6 (1174.66 Hz), and G_6 (1567.98 Hz). In addition, major (minor) stimuli contained eight copies of B_5 (987.77 Hz) [$B\flat_5$ (932.33 Hz)]. Each individual tone was 65 ms in duration and windowed by a raised cosine function with a 22.5-ms rise time. Thus, each stimulus lasted 2.08 s. An example of a major (minor), 3-task tone-scramble is provided in Mm. 1 (Mm. 2).

b. Task. Before beginning the task, the listener heard eight example stimuli that alternated between major and minor. Each major stimulus was labeled as "major (happy)" and each minor stimulus was labeled as "minor (sad)." Then, on each trial, the listener heard a single stimulus and strove to classify it as major (happy) or minor (sad) by pressing either the "1" key on the keyboard for major or the "2" key for minor. The feedback ("correct" or "incorrect") was printed to the screen after each trial, and the proportion correct was given at the end of the task. The next trial began 0.22 s after the listener entered the response to the previous trial. The listener completed 4 blocks of 50 trials (200 trials total).

4. RPD task

a. Stimuli and task. The stimuli were pairs of pure tones. Each tone had a duration of 500 ms and was windowed by a raised cosine function with a 22.5-ms rise time. The interstimulus interval was 1 s. Let ϕ_1 (ϕ_2) be the frequency of the first (second) tone on a given trial. The task was to judge whether ϕ_2 was higher or lower than ϕ_1. At the start of the task, the listener heard two examples each of a "higher" trial and a "lower" trial. After each trial during the task, feedback ("correct" or "incorrect") was printed to the screen. After hearing the stimulus on a given trial the listener entered "1" to indicate that $\phi_2 < \phi_1$ or "2" to indicate that $\phi_2 > \phi_1$. The next trial began 2 s after the previous response. Each listener completed 2 blocks of 50 trials (100 trials total).

b. How the frequencies of the two tones were determined on each trial. The frequency difference between the two tones in a given trial was determined by one of two randomly interleaved staircases. In a given staircase, the task-difficulty was controlled by a parameter, θ, whose value was adjusted by the staircase. In one staircase, θ was set initially to "1" and in the other staircase, θ was set initially to "0.4"; otherwise, the two staircases followed the same rules. For each staircase, after each trial, if the previous three responses in that staircase were correct, then θ was decreased to 0.75θ; otherwise (if the number of trials was fewer than three or any of the previous three responses was incorrect), θ was increased to 1.25θ.

To derive ϕ_1 and ϕ_2 on a given trial, we first selected $\hat{\phi}_1$ randomly from the linear frequency interval from 300 to 2000 Hz. We then set $\hat{\phi}_2$ equal to $\hat{\phi}_1 X$, where X is a random variable that takes one of the values $1+\theta/2$ or $1-\theta/2$ with equal probability. Then, we proceed as follows:

(1) If $300 \geq \hat{\phi}_2 \geq 2000$, we set $\phi_1 = \hat{\phi}_1$ and $\phi_2 = \hat{\phi}_2$.
(2) If $\hat{\phi}_2 < 300$, $\phi_1 = \min\{\hat{\phi}_1 + 300 - \hat{\phi}_2, 2000\}$ and $\phi_2 = 300$.
(3) If $\hat{\phi}_2 > 2000$, $\phi_1 = \max\{\hat{\phi}_1 - (2000 - \hat{\phi}_2), 300\}$ and $\phi_2 = 2000$.

Thus, the maximum possible difference between ϕ_2 and ϕ_1 occurred when one of these was 300 Hz and the other was 2000 Hz; this difference is $\log_2((2000-300)/300) = 2.5025$ octaves.

B. Results

The sensitivity in the 3-task, as reflected by d', was computed from the last 3 blocks of 50 trials. The first block of trials was treated as practice. If a listener was tested on n major (minor) stimuli and responded correctly on all of them, then the probability of a correct response was adjusted to n - 0.5/n (as suggested by Macmillan and Kaplan, 1985).

To estimate the rPDT for a given listener, we need to fit a psychometric function to the data for that listener. However, the situation is complicated by the vast spread of performance in the RPD task. We manage this large spread in Fig. 3 by logarithmically compressing the y axis; this serves to distribute the rPDTs > 50 cents approximately uniformly from $\log_2$(50 cents) to $\log_2$(6400 cents). As this might suggest, the data from any given listener tend to be well described by a psychometric function of log(cents).

Accordingly, we proceed as follows: for a given listener, we fit (using a maximum-likelihood criterion) the following Weibull function to the data from that listener in the RPD task:

$$\Psi(D) = 0.5 + 0.48\left[1 - \exp\left(-\left(\frac{D}{A}\right)^B\right)\right], \quad (1)$$

where $D = \log_2(d)$, for the pitch-difference magnitude on a given trial, d [i.e., $d = 1200|\log_2(\phi_2/\phi_1)|$]. A and B are the Weibull function threshold and steepness parameters, respectively. Note that

(1) $\Psi(0) = 0.5$, reflecting the fact that chance performance is 0.5 in this task; and

https://doi.org/10.1121/10.0010161

(2) $\Psi(D) \rightarrow 0.98$ as D grows large. This limit on probability correct is intended to cover the possibility of "finger errors," i.e., incorrect responses that occur even when the listener knows the correct answer.

We use $d = 2^A$ as our estimate of rPDT for a given subject. For $\log_2(d) = A$, the probability of responding correctly is 0.8034; therefore, the rPDTs reported here are predicted to support performance around 80% correct.[1]

The scatterplot of 3-task-d' vs rPDT is shown in Fig. 3. In Fig. 3, rPDTs are plotted on a log scale decreasing from bottom to top to reflect increasing levels of performance. There are two things to note about Fig. 3:

(1) All except 3 of the 59 listeners whose rPDTs were higher than 50 cents performed near chance in the 3-task. The three listeners who depart from this pattern are marked by the red circles; we will refer to them as pattern-breakers #1, #2 and #3.
(2) Across the 52 listeners whose rPDTs are lower than 50 cents, the distribution of 3-task-d' values is approximately uniform from near 0 to ceiling.

The dark bars of the left panel in Fig. 4 plot the histogram of d' in the 3-task across all 111 listeners; the dark bars of the right panel in Fig. 4 plot the corresponding histogram of proportion correct. As seen in previous studies (Chubb *et al.*, 2013; Dean and Chubb, 2017; Ho and Chubb, 2020; Mednicoff *et al.*, 2018), the histogram of 3-task-d' has a large mode near zero and strong positive skew, and the histogram of proportion correct is bimodal with one mode near 0.5 (chance performance) and another mode at 1.0 (ceiling). The lighter bars in the left panel of Fig. 4 show the distribution of d' in the 3-task when the listeners with rPDTs above 50 cents are excluded, and the lighter bars in the right panel of Fig. 4 plot the corresponding histogram of proportion correct. The large peak near chance performance ($d' = 0$ and proportion correct = 0.5) is greatly reduced in each panel of Fig. 4, resulting in a roughly uniform distribution of 3-task-d' and a distribution of proportion correct with a single prominent mode at ceiling performance.

1. Relationship of musical training to 3-task-d' and rPDT

The top panel of Fig. 5 plots self-reported years of musical training against 3-task-d', and the bottom panel of Fig. 5 plots years of musical training against rPDT. As found in previous studies, years of musical training is correlated with 3-task-d'. However, the distribution of years of musical training is highly non-normal with a strong positive skew: large numbers of listeners have three or fewer years of musical training, and very few have ten or more years of musical training. Thus, the correlation coefficient is likely to be misleading. When we look at the upper panel of Fig. 5, we note a large group of listeners with three or fewer years of musical training whose 3-task-d' values are near zero. The least squares linear prediction line must come close to the mean of this group (to minimize the distances from itself to the points in this group). On the other hand, the mean value of 3-task-d' for the relatively few listeners with ten or more years of musical training is up around 2.5. Even though these listeners are few in number, their mean 3-task-d' value exerts very strong influence on the prediction line. Concerning the latter group of listeners, though, we note that their 3-task-d' values are highly variable around their mean: although three of these listeners performed at ceiling, three others performed near chance. We also note that the sample contained a single listener with no musical training who performed perfectly in the 3-task. This pattern echoes the results of previous studies using the 3-task in suggesting that musical training is neither necessary nor sufficient for high performance in the 3-task. Similar comments apply to rPDTs: log (rPDT) is also correlated with years of musical training. Nonetheless, the sample contains many listeners with little or no musical training with rPDTs lower than 50 cents and several listeners with many years of musical training but high rPDTs, suggesting that musical training is neither necessary nor sufficient to have a low rPDT.

C. Discussion

Figure 3 shows that 56 of the 59 listeners with rPDTs greater than 50 cents (a quarter-tone) perform near chance in the 3-task. As shown in Fig. 4, it is this group of listeners that produces the large mode in the distribution of performance (as reflected either by d' or proportion correct) in the 3-task.

Of the three "pattern-breakers" (the filled red circles numbered 1, 2, and 3) in Fig. 3, pattern-breaker #1 stands out as a drastic counterexample to the proposal that a listener must have a rPDT less than 50 cents to do well in the 3-task. Pattern-breaker #1 performed nearly perfectly in the 3-task, responding incorrectly on only a single trial out of the last 150 trials; however, this listener's rPDT was more than an octave (1404 cents).

One might wonder whether something is amiss with the rPDT estimate for pattern-breaker #1 or whether pattern-breaker #1 failed to make sufficient effort in the RPD task for some reason; there is no indication that either of these issues is a problem. As we show in detail in Sec. I B of the supplementary materials,[2] close scrutiny of the RPD task data for pattern-breaker #1 suggests that, first, this listener was making a reasonable effort in the task and, second, the rPDT estimate reflects the performance of this listener.

We must conclude, then, that pattern-breaker #1 is a genuine counterexample to the proposal that it is necessary to have a rPDT less than 50 cents to perform well in the 3-task. It should also be noted (as shown in Fig. 5) that pattern-breaker #1 had 15 yr of musical training. Only one other listener with rPDT greater than 50 cents had at least 14 yr of musical training. (This listener performed poorly in the 3-task.) Perhaps the particular sort of training that pattern-breaker #1 received was instrumental in enabling this listener to perform well in the 3-task.

III. EXPERIMENT 2

As shown in Fig. 3, nearly all listeners with rPDTs greater than a quarter-tone perform very near chance in the 3-task. By contrast, for listeners with rPDTs lower than a quarter-tone, sensitivity in the 3-task (as reflected by d') is uniformly distributed from chance to ceiling. Thus,

(1) having a rPDT less than a quarter-tone is an important precondition to perform well in the 3-task; however, it is not sufficient to ensure high performance: there exist many listeners with rPDTs below a quarter-tone who, nonetheless, perform poorly in the 3-task; and
(2) the listeners with rPDTs greater than a quarter-tone produce the lower mode in the bimodal distribution in 3-task performance; when they are removed from the sample of listeners, the distribution of 3-task-d' becomes uniform, and the distribution of proportion correct becomes unimodal with the mode at ceiling.

Experiment 2 explores how variations in the pitch-difference task influence this pattern. A new group of listeners is tested in the 3-task as well as in four pitch-difference tasks. Previous research suggests that fixing the first tone across trials in a pitch-difference task makes the task much easier for nearly all listeners (Mathias *et al.*, 2010). A possible reason for the improved performance observed in such "fixed pitch-difference" tasks is that fixing ϕ_1 enables the listener to create a durable, internal representation of ϕ_1 across trials with which to compare ϕ_2. Such a strategy is not available in RPD tasks. This suggests that low-performers in RPD tasks may have difficulty preserving a temporary memory of ϕ_1 for comparison with ϕ_2. If so, then perhaps RPD task performance will improve for RPD-challenged listeners if the delay between tone-1 and tone-2 is decreased. To investigate this question, we include two RPD tasks, one with an inter-tone interval (ITI) 1.0 s (as used in experiment 1) and the other with an ITI 0.5 s.

Finally, we include a task [the "same-higher-lower" (SHL) task] in which ϕ_2 can be either higher, lower, or equal to ϕ_1, and the task is to classify the stimulus accordingly. This task is included to provide insight into why listeners with high rPDTs make errors. Specifically, we ask, is it true that most listeners with high rPDTs are "direction-challenged" [whose existence was documented by Semal and Demany (2006) and Mathias *et al.* (2010)]? Or do there also exist listeners who are unable to hear any difference between two tones even though they differ by a large interval?

A. Methods

All of the methods were approved by the UCI Institutional Review Board.

1. Participants

A new set of 151 undergraduate students were recruited from the Social Science Human Subjects Pool at the UCI. All listeners had self-reported normal hearing and received course credit for participating in the study. The data for this study were collected during the months between April 2020 and December 2020 during the Covid-19 pandemic; for this reason, the experiment took place online.[3] The data were excluded from analysis if listeners scored below five on the headphone check (see Sec. III A 2) or if their data suggested that they failed to make adequate effort (i.e., they continuously pressed the same button response for at least half of the trials in any task). This led to the exclusion of 52 listeners; as a result, data for 99 listeners were analyzed for this study. Within this group of 99 listeners, 57 reported having at least 1 yr of formal musical training. Within this subgroup, the mean number of years was 4.39 (standard deviation, 4.32).

2. Procedure

The participants were tested online. They were instructed to find a quiet room and wear headphones or earbuds for the entirety of the experiment. Listeners were free to adjust the volume to their comfort level. The sampling rate of the stimulus presentation was adjusted according to the sampling rate of the participant's device. If the sampling rate was outside the range of 44 100 to 48 000 samples/s (which would be unusual for a typical computer), then the participant was instructed to switch devices. The sampling rate was 44 100 samples/s for 37 participants and 48 000 samples/s for 62 participants. The specific sound card of each participant's device was unknown.

The headphone/earbud wear was screened at the start of the experiment via a three-alternative-forced choice task used by Woods *et al.* (2017). This task consists of six trials in which listeners judge which of three 200-Hz pure tones is quietest. Unknown to the listener, one tone in each trial is presented 180° out of phase across the stereo channels. This phase cancellation causes the task to be difficult over loudspeakers but easy over headphones. Woods *et al.* (2017) determined that listeners who score at least five correct trials can be assumed to be wearing headphones.

Following this test, the listeners completed a brief survey to report their native language and number of years of musical training. They were then tested in the 3-task and four pitch-difference tasks. The task order was randomly generated for each listener.

3. 3-task

The 3-task in the current experiment was similar to the 3-task in experiment 1. However, the tone-scrambles used the same set of notes but contained three (rather than eight) copies each of the possible notes. Thus, a single stimulus lasted 780 ms (instead of 2.08 s, which was the stimulus duration in experiment 1). Pilot studies suggested that the 12-tone stimuli used here would yield performance only slightly worse than the 32-tone stimuli used in experiment 1. Therefore, to shorten the duration of the task, we used the briefer stimuli. Before beginning the task, the listener heard two examples each of "type 1 (minor/sad)" and "type 2

https://doi.org/10.1121/10.0010161

(major/happy)" stimuli. Then, on each trial, the listener heard a single stimulus and strove to classify it as type 1 (minor/sad) or type 2 (major/happy) by clicking buttons on the screen. Type 1 stimuli corresponded to a button depicting a sad face emoji on the left side of the screen; type 2 stimuli corresponded to a button depicting a happy face emoji on the right side of the screen. The feedback (correct or incorrect) was printed to the screen after each trial, and the proportion correct was given at the end of the task. The listeners completed 3 blocks of 50 trials.

4. Pitch-difference tasks

a. Stimuli and task. The stimuli were pairs of pure tones. Each tone had a duration of 500 ms and was windowed by a raised cosine function with a 22.5-ms rise time.

The interstimulus interval and frequency of the first tone for each condition are listed in Table I. The interstimulus interval was 1-s in the fixed, gap-1.0, and SHL conditions; the interstimulus interval was 500 ms in the gap-0.5 condition.

In the fixed condition, the first tone in each pair was fixed at 440 Hz. In the gap-0.5, gap-1.0, and SHL conditions, the frequency of the first tone in each pair was selected uniformly from the log-frequency interval of 200–1600 Hz. In all cases, the maximum frequency difference was 1200 cents (1 octave). Hence, on each trial, the second tone fell between 100 and 3200 Hz.

In each pitch-difference task, the listener heard two tones per trial and responded whether the second tone was higher or lower than the first tone. In the SHL task, the listener could also respond "same."

At the start of the SHL task, the listener heard two examples each of a same trial and one example each of a higher trial and a lower trial. At the start of the other tasks (fixed, gap-0.5, gap-1.0), the listener heard two examples each of a higher trial and a lower trial.

Each of the 4 pitch-difference tasks included 2 blocks of 50 trials. The feedback (correct or incorrect) was printed to the screen after each trial, and the proportion correct was given at the end of each task.

b. How the frequencies of the two tones were determined on each trial. The pitch-difference magnitude in a given trial in the fixed, gap-0.5, and gap-1.0 conditions was determined by two interleaved three-down-one-up staircases. In a given staircase, task-difficulty was controlled by a parameter, θ, whose value was adjusted by the staircase. In staircase 1, θ was set initially to 100 cents; in staircase 2, θ was set initially to 700 cents. After each trial, if the previous three responses in staircase 1 (staircase 2) were correct, then θ was decreased to 0.9θ (0.7θ); otherwise, θ was increased to $\theta/0.9$ ($\theta/0.7$). The frequency differences in the SHL task were determined only by staircase 2. The second tone was higher in frequency than the first tone for exactly half of the trials in the fixed, gap-0.5, and gap-1.0 conditions. In the SHL condition, the second tone was the same as the first tone for exactly half of the trials (regardless of the result of the staircase), and the remaining trials were evenly split between higher and lower trials.

TABLE I. The interstimulus interval (ISI; duration between the two tones in each stimulus, in ms) and frequency of the first tone (in Hz) for each of the four pitch-difference tasks.

Condition	ISI (ms)	Frequency 1 (Hz)
Fixed	1000	440
Gap-0.5	500	Roved
Gap-1.0	1000	Roved
SHL	1000	Roved

B. Results

The histogram of 3-task-d' for all listeners is plotted in Fig. 6 in dark gray bars. As also shown in Fig. 3, the distribution of 3-task-d' values has a large mode near zero (produced by listeners with little or no sensitivity) and strong positive skew (produced by listeners with levels of sensitivity ranging from near zero to ceiling). However, the proportion of low-performing listeners in our sample (about 88% of participants) is higher than has been observed in previous studies (about 70% of participants). Moreover, in contrast to experiment 1, as shown by the light gray bars in Fig. 6, the histogram of 3-task-d' values for the listeners with thresholds below 50 cents has a clear mode at zero. There are several possible reasons for this disparity. First, the 3-task stimuli used in experiment 2 were much briefer (780 ms) than those used in experiment 1 (2080 ms). It seems likely that this made the 3-task used in experiment 2 more difficult. Second, in experiment 1, each listener interacted directly with an experimenter and was tested in a laboratory; by contrast, in experiment 2, each listener performed the experiment online without supervision. It seems likely that the performance of some listeners in experiment 2 may have suffered due to these factors.

We write $\mathrm{PDT}_{\mathrm{fixed}}$, $\mathrm{rPDT}_{1.0}$, and $\mathrm{rPDT}_{0.5}$ for the pitch-difference thresholds in the fixed, gap-0.5, and gap-1.0 tasks, respectively. For each listener, the threshold in a given task was estimated (using the same method as in

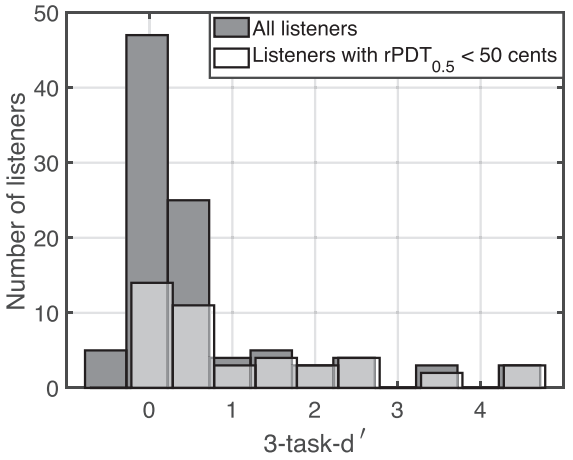

FIG. 6. The histogram of d' values achieved on the 3-task in Experiment 2 by all listeners (gray bars). The white bars (slightly shifted to the right for visualization purposes) represent the distribution of d' values for listeners who achieved $\mathrm{rPDT}_{0.5} < 50$ cents.

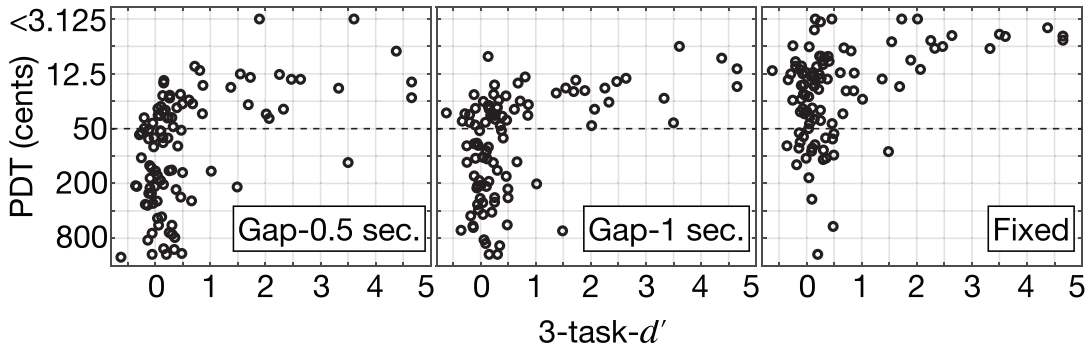

FIG. 7. The scatterplots of $rPDT_{0.5}$, $rPDT_{1.0}$, and PDT_{fixed} as a function of 3-task-d'. The thresholds are plotted (on a log scale) with values decreasing from bottom to top to reflect increasing performance. The dashed line is at a quarter-tone.

experiment 1) from the last 85 trials in that task, and 3-task-d' was estimated from the last 100 trials in the 3-task.

Figure 7 shows scatterplots of 3-task-d' vs PDT_{fixed}, $rPDT_{1.0}$, and $rPDT_{0.5}$. As in Fig. 3, the thresholds are plotted along the y axis of each plot on a log scale decreasing from bottom to top to reflect increasing levels of performance, and the horizontal dotted line is at a quarter-tone. Of the 55 listeners with $rPDT_{0.5} > 50$ cents, only 3 (5.5%) achieved 3-task-d' values > 1; by contrast, of the 44 listeners with $rPDT_{0.5} < 50$ cents, 16 (36%) achieved 3-task-d' values > 1. Similarly, of the 48 listeners with $rPDT_{1.0} > 50$ cents, only 2 (4%) achieved 3-task-d' values > 1; by contrast, of the 51 listeners with $rPDT_{0.5} < 50$ cents, 17 (33%) achieved 3-task-d' values > 1. Thus, the two RPD tasks yielded results similar to those in experiment 1. This confirms our previous observation that having a rPDT below 50 cents is an important precondition for performing well in the 3-task.

The cloud of points in the scatterplot for the fixed task in Fig. 7 appears to be shifted upward along the y axis compared to the clouds for the other two tasks, indicating that thresholds in the fixed task tend to be lower than those in the gap-1.0 and gap-0.5 tasks. This effect is seen more clearly in the left two panels of Fig. 8. Let $PDT_{k,fixed}$, $rPDT_{k,0.5}$, and $rPDT_{k,1.0}$ be the pitch-difference thresholds for a given

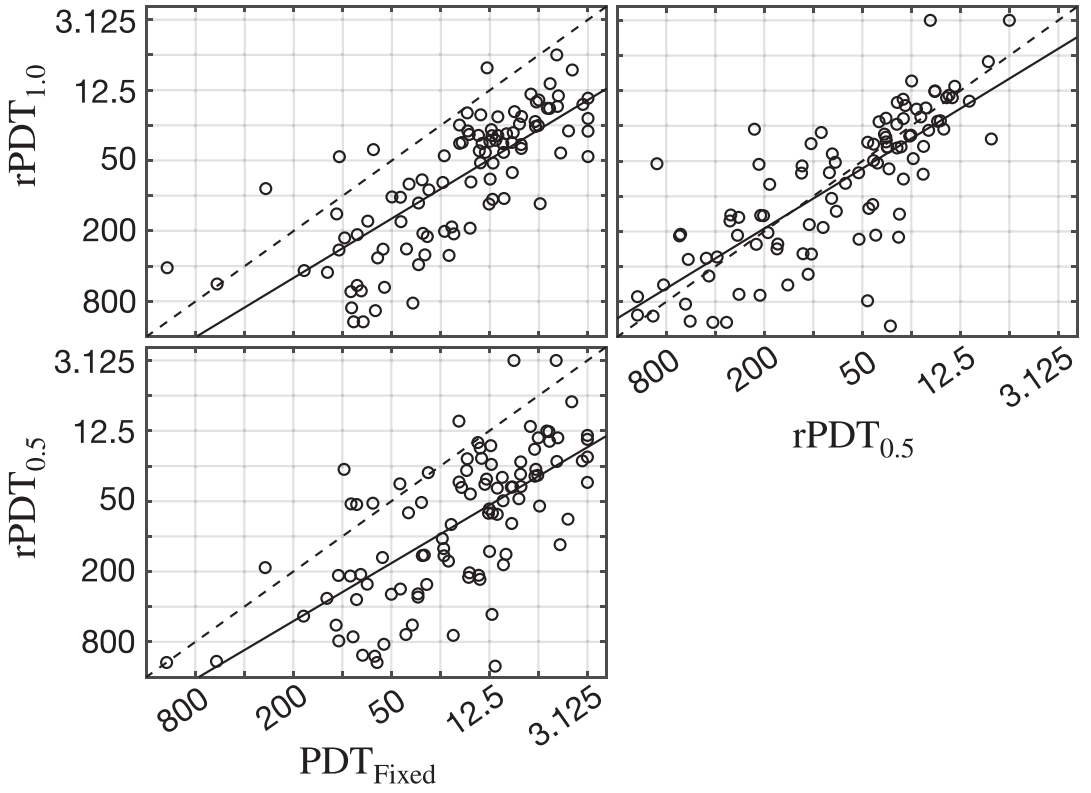

FIG. 8. The scatterplots showing the relationships between thresholds in the three pitch-difference tasks. PDT_{fixed} vs $rPDT_{1.0}$ (upper-left), $rPDT_{0.5}$ vs $rPDT_{1.0}$ (upper-right), PDT_{fixed} vs $rPDT_{0.5}$ (lower-left). In each panel, the dashed line shows $y = x$, and the solid line shows the least-squares regression line. The thresholds decrease along the x and y axes to reflect increasing performance. Note that the regression lines are shifted downward from the line $y = x$ by a factor of around 4 in the upper- and lower-left panels.

https://doi.org/10.1121/10.0010161

listener k in the fixed, gap-0.5 and gap-1.0 tasks, respectively. The top-left (bottom-left) panel of Fig. 8 plots $\mathrm{PDT}_{k,\mathrm{fixed}}$ against $\mathrm{rPDT}_{k,1.0}$ ($\mathrm{rPDT}_{k,0.5}$) for all listeners, k. In each of these two plots, the regression line (solid) is lower than the line $x = y$ (dashed) by roughly a factor of 4. Across all listeners, k, the geometric mean of $R_k = \mathrm{rPDT}_{k,1.0}/\mathrm{PDT}_{k,\mathrm{fixed}}$ was 3.70, and the geometric mean of $R_k = \mathrm{rPDT}_{k,0.5}/\mathrm{PDT}_{k,\mathrm{fixed}}$ was 4.08. Both of these values deviate signficantly from one [in each case, a t-test of the null hypothesis that the mean of $\log_2(R_k)$ was zero yielded $p = 0.0000$]. Therefore, on average, across all listeners, the pitch-difference thresholds in the two roved tasks were roughly four times higher than they were in the fixed task.

The scatterplot of $\mathrm{rPDT}_{k,0.5}$ vs $\mathrm{rPDT}_{k,1.0}$ (Fig. 8, upper-right) suggests that these two conditions yielded a performance that was roughly equal on average, and (confirming this impression) the geometric mean of $R_k = \mathrm{rPDT}_{k,0.5}/\mathrm{rPDT}_{k,1.0}$ was 1.07. This value was not significantly different from one; specifically, a (two-tailed) t-test of the null hypothesis that the mean of $\log_2(R_k)$ was zero yielded $t_{98} = 0.63$; $p = 0.53$.

In each panel of Fig. 8, the slope of the regression line is slightly less than one; however, it would be a mistake to take these slopes seriously. In each panel, the slope of the regression line depends heavily on the sparsely scattered points corresponding to listeners for whom both thresholds are elevated (i.e., the outlying points in the lower-left quadrant of the panel).

1. Results from the SHL task

The main focus of this paper is on the relationship between the performance in the pitch-difference tasks and the 3-task. The SHL task was included in experiment 2 to address the ancillary question of what makes RPD tasks difficult for some listeners. Accordingly, we relegate the detailed analysis of these data to the supplementary materials.[2] We briefly summarize the results here. Our analysis focuses on the errors that the listeners made on trials in which the two tones presented are different ("tones-different trials"). We call an error in this class an "undetected-difference" error if the listener judges that the two tones were the same; otherwise, the error is called a "wrong-direction" error. In addition, we call the number of cents between ϕ_1 and ϕ_2 on a tones-different error-trial the error "magnitude."

The results from the SHL task confirm [as documented by Semal and Demany (2006) and Mathias *et al.* (2010)] that some listeners with high values of $\mathrm{rPDT}_{0.5}$ and $\mathrm{rPDT}_{1.0}$ can hear that two tones are different without being able to judge the direction of the difference. Most of the errors made by these listeners on tones-different trials are wrong-direction errors, and for these listeners, the magnitudes for wrong-direction errors tend to be substantially higher than those for undetected-difference errors. However, for other listeners, most errors on tones-different trials tend to be undetected-difference errors, and the magnitudes tend to be roughly equal on undetected-difference and wrong-direction errors. Thus, the behavior of these listeners suggests that whenever they can hear that the two tones are different on a given trial, they can correctly judge the direction of the difference; however, they have difficulty detecting even fairly large differences. Other listeners with high rPDTs seem to fall between these two extreme classes.

2. Relationship between musical training and each of 3-task-d' and $\mathrm{rPDT}_{0.5}$

The scatterplots of 3-task-d' and $\mathrm{rPDT}_{0.5}$ against the self-reported years of musical training are shown in Fig. 3 of the supplementary materials.[2] As found in experiment 1, each of 3-task-d' and $\mathrm{rPDT}_{0.5}$ is positively correlated with the self-reported years of musical training; yet, the distribution of years of musical training is highly non-normal with strong positive skew, making the correlation coefficient misleading. Similar comments as those in Sec. II B 1 apply here. In each plot, we see listeners with little or no musical training who perform well and other listeners with many years of training who perform poorly, suggesting that musical training is neither necessary nor sufficient for high 3-task-d' or $\mathrm{rPDT}_{0.5}$.

C. Discussion

The current results confirm the previous findings that show that the performance is better in the pitch-difference tasks in which the frequency of the first tone is fixed than it is in tasks in which the first tone is roved over a large interval (Mathias *et al.*, 2010; Semal and Demany, 2006). In particular, we find that, on average, rPDTs in each of the gap-0.5 and gap-1.0 tasks are roughly four times higher than PDTs in the fixed task.

What makes the roved tasks more difficult than the fixed task? The fixed task affords the listener the possibility to develop a stable internal representation of ϕ_1 (the frequency of tone-1) that can be refined across trials. By contrast, in the two roved tasks, the listener must construct a new memory trace for ϕ_1 on each trial. It seems likely that the decrease in performance in the two roved tasks compared to the fixed task is due to a decrease in the quality of the ϕ_1 memory trace that must be used in the roved tasks. Perhaps, in the roved tasks, the ϕ_1 memory trace is unstable across the temporal interval during which ϕ_1 must be remembered for comparison with ϕ_2. If so, then decreasing the duration of the temporal interval during which ϕ_1 must be retained should improve the performance in a RPD task. In this case, the performance should be better in the gap-0.5 task than it is in the gap-1.0 task. We find no evidence of this: across all listeners, k, the geometric mean $\mathrm{rPDT}_{k,0.5}/\mathrm{rPDT}_{k,1.0}$ is 1.07. Thus, on average, the rPDTs in the gap-0.5 and gap-1.0 tasks are roughly equal.

The current results, therefore, argue against the idea that the memory trace used in roved tasks degrades over the brief time during which it must be retained; this suggests that the memory process used to compare ϕ_1 to ϕ_2 in roved tasks is deficient (compared to the process used in fixed

tasks) in some other way. It is possible that in roved tasks, the presentation of tone-2 disrupts the ϕ_1 trace. It is also possible that in roved tasks, the ϕ_1 trace is noisier in its initial construction than the ϕ_1 trace developed across trials in the fixed tasks. The finding that the performance in roved tasks varies gradually depending on the size of the rove argues for the latter possibility (Mathias *et al.*, 2010).

IV. GENERAL DISCUSSION

The current results show that (1) an important precondition for success in the 3-task is having a rPDT < 50 cents (a quarter-tone), and (2) roughly half of all listeners fail to satisfy this precondition. Hence, what seems to prevent roughly half of all listeners from hearing the difference between the major and minor stimuli used in the 3-task is inadequate sensitivity to variations in the pitch across time.

It is, perhaps, surprising that the critical rPDT is a quarter-tone rather than a semitone. After all, the notes that differ between the major and minor stimuli in the 3-task are $B\flat_5$ and $B\natural_5$, and these differ by a semitone. Thus, a listener whose rPDT falls between a quarter-tone and a semitone should, on the one hand, be able to hear the direction of the difference between two notes that differ by a semitone (e.g., $B\flat_5$ and $B\natural_5$) and, yet (because listeners with rPDT > 50 cents perform near chance in the 3-task), be unable to hear the difference between two tone-scrambles whose notes are identical except that one contains $B\natural_5$'s and the other contains $B\flat_5$'s.

It is important to bear in mind, however, that the characteristic qualities (the "happiness" and "sadness") produced by major and minor tone-scrambles that enable high-performing listeners to tell them apart depend on the intervals formed by the $B\flat_5$ and $B\natural_5$ with the context tones, G_5, D_5, and G_6. All of these intervals are much larger than a semitone. We speculate that having a rPDT > 50 cents may compromise accurate registration of these relatively large intervals thereby obscuring the qualitative differences between major and minor tone-scrambles.

Suppose, as we propose, that listeners with rPDT < 50 cents are blocked from hearing any difference between major and minor tone-scrambles by inadequate pitch-difference sensitivity. In this case, should we expect these same listeners to be able to discriminate the qualities of majorness and minorness in real music? We hypothesize that the answer is no for the following reason: the pitch intervals between the notes of the tonic triads used in real major and minor music are identical to those used in major and minor 3-task stimuli. Therefore, if what limits listeners in the 3-task is inadequate sensitivity to these intervals, then plausibly they will be similarly limited in their experience of real music.

If this is true, then altering tone-scrambles to make them more like real music should not help low-performers do better in the 3-task. Mednicoff *et al.* (2018) have shown that low-performers do not benefit from having the stimuli presented more slowly. In the slowest condition used in that study, the listeners heard (in random order) one each of the notes G_5, D_6, G_6, and either one $B\flat_5$ (in minor stimuli) or one $B\natural_5$ (in major stimuli) with each tone presented for 520 ms. In this slow variant of the tone-scramble task, the performance was significantly worse than the performance in the standard tone-scramble task. Moreover, in none of the slowed-down versions of the tone-scramble task investigated by Mednicoff *et al.* (2018) was performance better than in the standard, 32-tone version in which each tone lasts 65 ms.

It might be objected that 3-task tone-scrambles differ from real music not only in being faster but also in being higher in pitch than most real music and using pure tones instead of notes with the rich timbres and complex attacks typical of real music. To see if these factors would influence 3-task performance, we recently tested 86 listeners in 6 different tone-scramble tasks. One of these tasks was the standard, 32-tone, 3-task and another was a "naturalistic" variant of the task. This variant used notes drawn from the middle of the piano keyboard and presented more slowly and with piano timbre. Specifically, the context notes were C_4, G_4, and C_5, and the two signal notes were $E\flat_4$ (in minor tone-scrambles) and $E\natural_4$ (in major tone-scrmbles). These piano-note-scrambles contained three of each note (presented in random order) with each note lasting 172 ms (for a presentation rate of 349 BPM). Thus, these stimuli were constructed to have a pitch-range, timbre, and presentation typical of natural music. Nonetheless, we found that listeners actually performed significantly better in the standard 3-task than in the naturalistic, piano-note-scramble task (a two-tailed paired samples t-test yielded $t_{85} = 2.24$, $p = 0.028$).

Although this evidence argues that low 3-task performers are unable to hear the difference between real music in the major and minor scales, there exist several studies that seem to refute this proposal. In particular, the studies by Temperley and Tan (2013) and Bonetti and Costa (2019) suggest that nearly all adult listeners are sensitive to musical mode. In both studies, groups of musically untrained listeners produce (without feedback) highly reliable responses in assessing the sadness vs happiness of various musical passages. It should be noted, though, that in each of these studies, each listener heard a given musical passage multiple times altered only in mode (i.e., without changing its rhythm, tempo, or other properties). These studies, therefore, suggest that when presented with musical passages that are identical in structure except for mode, nearly all listeners can hear the qualities characteristic of variations in mode. We note, however, that when one listens to actual music, one must experience the variations in mode directly (without comparing the music to an alternate version identical except for mode). We conjecture that for listeners with rPDT > 50 cents, the qualities produced by variations in the musical mode are subtle in comparison to the qualities produced by other aspects of musical variation. Under this hypothesis, the qualities produced by mode can be revealed for these listeners by allowing them to compare musical passages that are identical except for mode; otherwise, however

(if such comparisons are not available), the qualities produced by mode are swamped by the variations in quality due to note-order, tempo, rhythm, and other aspects of musical structure.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) Training Grant No. T32-DC010775 awarded to J.H. by the UC Irvine Center for Hearing Research. The authors also thank Nellie Kwang and Melissa Huynh for their assistance with piloting.

[1]To check the accuracy of the rPDT estimate for each listener, we used a Markov chain Monte Carlo simulation to derive samples from the posterior joint density, characterizing the parameters A and B. If the rPDT of a listener was lower than 50 cents (a quarter-tone), the 100 trials of data obtained from that listener typically sufficed to tightly constrain the estimate of A (i.e., the Bayesian credible interval around A was small). However, if the rPDT of a listener was higher than 50 cents, this was often not true. The data from these low-performing listeners were often very ragged, and the values visited by their staircases tended to range widely; consequently, in such cases, the credible interval around A sometimes spanned several orders of magnitude. [This was true, for example, for the data and corresponding Weibull function fit shown in Fig. 3 of the supplementary materials (footnote 2) for the listener marked by the red dot labeled "1" in Fig. 3.] Nonetheless, the maximum-likelihood Weibull function estimates generally did a reasonable job of capturing the overall trends even in the most aberrant data sets. Thus, although it would be a mistake to take the rPDT estimate for a given, low-performing listener too seriously, in each case, the estimated rPDT appears sensible based on the available data.

[2]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0010161 for details concerning (1) the relation between (self-reported) years of musical training on performance in the 3-task and the Gap-0.5 task in Experiment 2, (2) the performance of Pattern-breaker #1 in the pitch-discrimination task in Experiment 1, and (3) the analysis of the SHL (Same-Higher-Lower) task from Experiment 2.

[3]See https://pitchdiffrove.web.app/ (Last viewed 1/20/2021).

Adler, S. A., Comishen, K. J., Wong-Kee-You, A. M. B., and Chubb, C. (**2020**). "Sensitivity to major versus minor musical modes is bimodally distributed in young infants," J. Acoust. Soc. Am. **147**, 3758–3764.

Bella, S. D., Peretz, I., Rousseau, L., and Gosselin, N. (**2001**). "A developmental study of the affective value of tempo and mode in music," Cognition **80**, B1–B10.

Blechner, M. J. (**1977**). "Musical skill and the categorical perception of harmonic mode," Haskins Lab. Status Rep. Speech Percept. **SR-51/52**, 139–174, available at https://yaleedu-my.sharepoint.com/personal/kraig_eisenman_yale_edu/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Fkraig%5Feisenman%5Fyale%5Fedu%2FDocuments%2FDepartments%2FHaskins%2FHaskinsLabs%2Eorg%2FStatus%20Reports%2FSR051%5F52%281977%29%2FSR051%5F12%2Epdf&parent=%2Fpersonal%2Fkraig%5Feisenman%5Fyale%5Fedu%2FDocuments%2FDepartments%2FHaskins%2FHaskinsLabs%2Eorg%2FStatus%20Reports%2FSR051%5F52%281977%29.

Bonetti, L., and Costa, M. (**2019**). "Musical mode and visual-spatial cross-modal associations in infants and adults," Musicae Sci. **23**(I), 50–68.

Chubb, C., Dickson, C. A., Dean, T., Fagan, C., Mann, D. S., Wright, C. E., Guan, M., Silva, A. E., Gregersen, P. K., and Kowalski, E. (**2013**). "Bimodal distribution of performance in discriminating major/minor modes," J. Acoust. Soc. Am. **134**(4), 3067–3078.

Crowder, R. G. (**1985a**). "Perception of the major/minor distinction: II. Experimental investigations," Psychomusicology **5**(1/2), 3–24.

Crowder, R. G. (**1985b**). "Perception of the major/minor distinction: III. Hedonic, musical, and affective discriminations," Bull. Psychon. Soc. **23**(4), 314–316.

Cunningham, J. G., and Sterling, R. S. (**1988**). "Developmental change in the understanding of affective meaning in music," Motiv. Emot. **12**, 399–413.

Dean, T., and Chubb, C. (**2017**). "Scale-sensitivity: A cognitive resource basic to music perception," J. Acoust. Soc. Am. **142**(3), 1432–1440.

Gerardi, G. M., and Gerken, L. (**1995**). "The development of affective responses to modality and melodic contour," Music Percept. **12**(3), 279–290.

Halpern, A. R. (**1984**). "Perception of structure in novel music," Mem. Cognit. **12**, 163–170.

Halpern, A. R., Bartlett, J. C., and Dowling, W. J. (**1998**). "Perception of mode, rhythm, and contour in unfamiliar melodies: Effects of age and experience," Music Percept. **15**, 335–356.

Heinlein, C. P. (**1928**). "The affective character of the major and minor modes in music," Comp. Psychol. **8**(2), 101–142.

Hevner, K. (**1935**). "The affective character of the major and minor modes in music," Am. J. Psychol. **47**, 103–118.

Ho, J., and Chubb, C. (**2020**). "How rests and cyclic sequences influence performance in tone-scramble tasks," J. Acoust. Soc. Am. **147**, 3859–3870.

Johnsrude, I. S., Penhune, V. B., and Zatorre, R. J. (**2000**). "Functional specificity in the right human auditory cortex for perceiving pitch direction," Brain **123**, 155–163.

Kastner, M. P., and Crowder, R. G. (**1990**). "Perception of the major/minor distinction: IV. Emotional connotations in young children," Music Percept. **8**(2), 189–202.

Kragness, H. E., Swaminathan, S., Cirelli, L. K., and Schellenberg, E. G. (**2021**). "Individual differences in musical ability are stable over time in childhood," Dev. Sci. **24**(4), e13081.

Kraus, N., and Chandrasekaran, B. (**2010**). "Music training for the development of auditory skills," Nat. Rev. Neurosci. **11**(8), 599–605.

Kraus, N., Slater, J., Thompson, E. C., Hornickerl, J., and Strait, D. L. (**2014**). "Music enrichment programs improve the neural encoding of speech in at-risk children," J. Neurosci. **34**(36), 11913–11918.

Leaver, A. M., and Halpern, A. R. (**2004**). "Effects of training and melodic features on mode perception," Music Percept. **22**, 117–143.

Macmillan, N. A., and Kaplan, H. L. (**1985**). "Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates," Psychol. Bull. **98**(1), 185–199.

Mann, D. S. (**2014**). "Processing stimuli over time: Musical modes and audiovisual binding," Ph.D. thesis, University of California, Irvine.

Mathias, S. R., Micheyl, C., and Bailey, P. J. (**2010**). "Stimulus uncertainty and insensitivity to pitch-change direction," J. Acoust. Soc. Am. **127**, 3026–3037.

Mednicoff, S., Mejia, S., Rashid, J., and Chubb, C. (**2018**). "Many listeners cannot discriminate major vs minor tone-scrambles regardless of presentation rate," J. Acoust. Soc. Am. **144**(4), 2242–2255.

Patel, A. D. (**2011**). "Why would musical training benefit the neural encoding of speech? The opera hypothesis," Front. Psychol. **2**, 142.

Patel, A. D. (**2014**). "Can nonlinguistic musical training change the way the brain processes speech? The expanded opera hypothesis," Hear. Res. **308**, 98–108.

Peretz, I., Champod, S., and Hyde, K. (**2003**). "Varieties of musical disorders: The Montreal Battery of Evaluation of Amusia," Ann. N.Y. Acad. Sci. **999**, 58–75.

Peretz, I., Gagnon, L., and Bouchard, B. (**1998**). "Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage," Cognition **68**, 111–141.

Rameau, J. P. (**1971**). *Treatise on Harmony* (Dover, New York).

Schoenberg, A. (**1978**). *Theory of Harmony* (University of California Press, Berkeley, CA).

Semal, C., and Demany, L. (**2006**). "Individual differences in the sensitivity to pitch direction," J. Acoust. Soc. Am. **120**(6), 3907–3915.

Swaminathan, S., and Schellenberg, E. G. (**2020**). "Musical ability, music training, and language ability in childhood," J. Exp. Psychol.: Learn., Mem., Cognit. **46**(12), 2340–2348.

Temperley, D., and Tan, D. (**2013**). "Emotional connotations of diatonic modes," Music Percept. **30**(3), 237–257.

Tramo, M. J., Shah, G. D., and Braida, L. D. (**2002**). "Functional role of auditory cortex in frequency processing and pitch perception," J. Neurophysiol. **87**, 122–139.

Tymoczko, D. (**2011**). *A Geometry of Music—Harmony and Counterpoint in the Extended Common Practice* (Oxford University Press, New York).

Woods, K. J., Siegel, M. H., Traer, J., and McDermott, J. H. (**2017**). "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**(7), 2064–2072.