UNIVERSITY OF CALIFORNIA,
IRVINE

Disparity in performance on tone-scramble tasks: generalizability and relevance to music

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Science

by

Sebastian Waz

Dissertation Committee:
Professor Virginia Richards, Chair
Professor Charles Chubb
Professor Charles Wright

2022

DEDICATION

In memory of the resigned Magister Ludi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work was supported by the Werner estate and the UCI School of Social Sciences through the Fellowship in Honor of Christian Werner during the Spring 2021 academic quarter (March 2021 through June 2021).

Special thanks is owed to the many undergraduate research assistants who volunteered their time to collect the data reported in Chapter 3, Experiment 1.

Over the past half-decade, my thinking has been shaped and nurtured by many people, and I am grateful for that. I thank many mentors, mentees, collaborators, cronies, friends, and loved ones. These categories are far from mutually exclusive.

VITA

Sebastian Waz

EDUCATION

Doctor of Philosophy in Cognitive Science — **2022 (expected)**
University of California, Irvine — *Irvine, CA*

Master of Science in Statistics — **Sep 2019**
University of California, Irvine — *Irvine, CA*

Bachelor of Science in Cognitive Science and Computing — **Jun 2016**
University of California, Los Angeles — *Los Angeles, CA*

EMPLOYMENT

Research Scientist Intern — **Jun 2022–Sept 2022**
Meta Reality Labs, Audio Team — *Redmond, WA*

Instructor — **Jun 2021–Sept 2021**
University of California, Irvine — *Irvine, CA*
Course: PSYCH 10C (Probability & Statistics)

Teaching Assistant — **Sep 2017–current**
University of California, Irvine — *Irvine, CA*
Courses: STATS 7, STATS 8, PSYCH 114M, SOCSCI 10C)

GIS Analyst — **Jun 2016–Sep 2017**
Easter Island Statue Project — *Santa Monica, CA*

OTHER PROFESSIONAL EXPERIENCE

Task Force Member **Aug 2020–current**
Acoustical Society of America
Technical Committee on Psychological & Physiological Acoustics
Task Force on Remote Testing

Volunteer Developer **Jun 2021–Mar 2022**
`https://www.findingfive.com/`

Colloquium Organizing Committee Member **Sep 2020–May 2021**
UCI Cognitive Sciences Colloquium Committee

Writer **Mar 2020–Jul 2021**
The Loh Down on Science
Broadcast on NPR
LDOS Media Lab, Inc.
`https://lohdownonscience.com`

FUNDING AND AWARDS

Fellowship in Honor of Christian Werner **Spring 2021**
Werner estate and UCI School of Social Sciences
Awarded September 2020

AGS Travel Award **Summer 2020**
Associate Graduate Students of UC Irvine
Awarded July 2019

PROPS Scholarship **Winter 2016–Spring 2016**
Psychology Research Opportunity Program (PROPS)
UCLA Department of Psychology
Awarded November 2015

MENTORSHIP

Cascading Mentorship Huddle Leader **Jul 2021–Jun 2022**
UCI Cognitive Sciences
All Eyes and Ears Huddle
Members: 7

Cascading Mentorship Certification **Jun 2021**
UCI School of Social Sciences

Mentorship Excellence Program Certification **Apr 2018**
UCI Graduate Resource Center

Undergraduate research supervision:

Harmonic Structure and the Discrimination of Major/Minor Modes **May 2019**
26th annual UCI Undergraduate Research Symposium

JOURNAL ARTICLES

Remote testing for psychological and physiological acoustics **2022**
Journal of the Acoustical Society of America

Evidence for strictly monocular processing in visual motion opponency and Glass pattern perception. **2021**
Vision Research

Initial report of the ASA P&P Task Force on Remote Testing **2020**
Proceedings of Meetings on Acoustics

CONFERENCE POSTERS

How do listeners use context frequencies in tone-scramble tasks? Evidence from a web-based experiment **December 2020**
179th Meeting of the Acoustical Society of America: Acoustics Virtually Everywhere

Evidence of a single neural mechanism underlying scale-sensitivity **September 2019**
Society for Music Perception and Cognition

Laterally connected neural field provides precise centroid estimates **September 2018**
2nd Computational Cognitive Neuroscience Conference

Boundary Extension: Insights from Signal Detection Theory **May 2015**
24th Annual Psychology Undergraduate Research Conference at UCLA

ABSTRACT OF THE DISSERTATION

Disparity in performance on tone-scramble tasks: generalizability and relevance to music

By

Sebastian Waz

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2022

Professor Virginia Richards, Chair

Music has a remarkable power to arouse the feelings of those who listen to it. What features of music imbue it with such emotional resonance? A prevailing notion in music theory is that musical scale has a central role in giving meaning to music. Indeed, many studies find that, according to listeners' average ratings, music of the major scale sounds happy, and music of the minor scale sounds sad. However, recent discoveries involving "tone-scramble" stimuli complicate our understanding of these results and suggest that sensitivity to scale is not universal. This thesis details a series of experiments designed to investigate (1) the generalizability of the findings of laboratory-based tone-scramble experiments and (2) the musical nature of a latent cognitive resource that is theorized to underlie performance in tone-scramble tasks. Chapter 1 reports that the same bimodal distribution in performance repeatedly observed in laboratory-based tone-scramble experiments is observed in a large, linguistically diverse web-based sample. Chapter 2 considers whether low-performing listeners are limited by some of the non-musical qualities inherent to tone-scrambles; data are provided to show that changes in presentation rate, frequency height, and timbre that yield tone-scrambles akin to actual music do not provide low-performing listeners any advantage over ordinary tone-scrambles. In Chapter 3, the latent cognitive resource theorized to underlie performance on tone-scramble tasks is shown to operate on musical scale and not individual frequencies, drawing a clear relationship between performance in tone-scramble

tasks and sensitivity to musical scale.

Chapter 1

Tone-scramble findings generalize to a broad population of listeners and do not depend on native language.

1.1 Introduction

A prevailing notion in music theory holds that the emotional connotation of music is determined in large part by *scale*, the set of musical notes used in a piece of music ordered by fundamental frequency (e.g., Rameau, 1722; Schoenberg, 1922; Tymoczko, 2011). Indeed, many studies find that, on average, listeners judge music of the *major* scale to be happy and music of the *minor* scale to be sad (e.g., Hevner, 1935; Crowder, 1984; Crowder, 1985a; Kastner and Crowder, 1990; Gerardi and Gerken, 1995; Gagnon and Peretz, 2003; Temperley and Tan, 2013; Bonetti and Costa, 2019); however, recent work involving "tone-scramble" stimuli complicates our understanding of these results. This work may, in turn, change our understanding of the role of musical scale in eliciting an emotional response from the listener.

Experiments involving tone-scrambles suggest that sensitivity to scale is not universal. Instead, it follows a bimodal distribution, with most listeners lacking sensitivity to scale (Chubb et al., 2013; Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020; Adler et al., 2020; Ho et al., 2022). A tone-scramble is a randomly ordered sequence of pure-tones whose pitches are drawn from a musical scale.

The most extensively studied tone-scramble task is the "3v**4**-task". In this task, the tone-scrambles contain thirty-two 65-ms pure-tones, including eight each of the notes G_5, D_6, and G_6; additionally, minor (major) tone-scrambles contain eight pure-tones of the note $B\flat_5$ ($B\natural_5$) which is three (four) semitones above the lowest note, G_5, hence the "3" ("**4**", whose bold font signals that this is a major interval) in the name "3v**4**-task". Tone-scrambles in the 3v**4**-task are presented one at a time, and the listener strives to classify the tone-scramble presented on each trial as major or minor with trial-by-trial feedback. Across listeners, the proportion-correct in the 3v**4**-task follows a bimodal distribution. The majority of listeners ($\approx 70\%$) hear little difference between the major and minor tone-scrambles and form a mode near 55% correct. The remaining listeners are highly sensitive to this difference, forming another mode near 100% correct.

In the context of related literature on listeners' sensitivity to musical scale, the results achieved using tone-scrambles are not a pure anomaly. Others have found that across various levels of musical training, listeners struggle to discriminate melodies differing only in scale (Halpern, 1984; Halpern et al., 1998; Leaver and Halpern, 2004). The results of tone-scramble experiments are consistent with most of the previous research showing that, on average, listeners hear music in the major (minor) scale as "happy" ("sad"). In many of these studies, the mean effect is modest and concordant with a bimodal distribution in sensitivity to scale (Hevner, 1935; Crowder, 1984; Crowder, 1985b; Kastner and Crowder, 1990; Gerardi and Gerken, 1995; Gagnon and Peretz, 2003). In such studies, effects that are "statistically significant" may be driven by a small number of strongly sensitive listeners in

a sample consisting mostly of listeners with little or no sensitivity. Thus, it is possible that all of these findings may be theoretically unified.

Although the results of tone-scramble experiments may be theoretically reconciled with the aforementioned behavioral results regarding scale, the finding that most listeners lack sensitivity to scale in the context of tone-scrambles remains counter-intuitive. Researchers have collected abundant psychological and physiological evidence of the emotional influence that music can have on listeners (Eerola and Vuoskoski, 2013; Juslin, 2013; Koelsch, 2014), and as cited above, music theorists have ascribed a primary role in evoking these emotional responses to scale. The results put forth by music psychologists corroborate this account, even if the results cited above may obscure an underlying disparity in sensitivity to scale. The general prominence of the major and minor scales in Western music should not be overlooked, either. Altogether, these facts provide a context in which the tone-scramble findings are quite surprising.

1.1.1 Is the bimodal distribution specific to the samples studied in prior experiments?

Given the counter-intuitive nature of the findings of tone-scramble experiments, we must consider whether the bimodality in the distribution of proportion-correct on the 3v**4**-task is an effect that is specific to the samples studied in prior experiments. Chubb et al. (2013), Dean and Chubb (2017), Mednicoff et al. (2018), and Ho and Chubb (2020) recruited listeners from the UCI School of Social Sciences Subject Pool, a population of listeners consisting primarily of undergraduate students enrolled at the University of California, Irvine, in a social-science degree program. Although this kind of sampling is common practice in the psychological sciences, such a sampling mechanism leaves open the possibility that conclusions thereby drawn do not generalize to the population of listeners at large. The population sampled in

prior tone-scramble experiments likely differs from the population of interest (e.g., listeners globally) along variables such as age, ethnic/cultural background, and socio-economic status. Such variables are known to be associated with musical ability. For example, Deutsch et al. (2006) found that Mandarin-speaking students from the Central Conservatory of Music in Beijing had a much greater proportion of listeners with absolute pitch than non-Asian students from the Eastman School of Music in Rochester, New York, who spoke a non-tonal language. Hove et al. (2010) reported similar differences with respect to a relative-pitch interval-identification task.

Is it possible that the large proportion of low-performing listeners observed in the 3v**4**-task – like the proportion of music-school students found to have absolute pitch – reflects some demographic characteristics of the population sampled thus far? The data currently available provide limited insight into this issue. Prior studies have broached this topic primarily with respect to musical background, finding a modest positive association between listeners' years-of-musical-training and their performance on tone-scramble tasks (Chubb et al., 2013; Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020). This association is not adequate to explain the disparity between low-performing and high-performing listeners in the 3v**4**-task. However, if the listeners in prior tone-scramble experiments exhibited a greater (lesser) degree of musical training than the population of listeners at large, the positive association between years-of-musical-training and performance would imply that prior tone-scramble experiments underestimate (overestimate) the proportion of low-performing listeners in the population at large. With regard to other demographic covariates, no prior tone-scramble studies have reported any formal analysis. It thus remains an open question whether any such covariates predict listeners' performance in tone-scramble tasks and, in turn, restrict the generalizability of results from prior experiments.

To address potential issues of generalizability, the current study uses the "citizen-science" approach to data collection (Hilton and Mehr, 2021), recruiting listeners from a more di-

verse population than would be available, for example, via institutional subject pool. We developed a tone-scramble experiment that ran within listeners' web browsers and garnered interest from lay Internet users by offering entertainment (the experiment had a video-game component) and personalized feedback (a determination of each listener's "super-listener" status) in the course of participation. This approach yielded a large ($N = 59,897$), demographically diverse sample of participants.

The present study represents a form of *remote testing*, using a web-based experimental paradigm to overcome the geographic constraints inherent to laboratory-based ("lab-based") testing. Remote testing has received considerable interest from the auditory research community (Peng et al., 2022), especially in the area of music cognition (e.g., Peretz and Vuvan, 2017; Mehr et al., 2018). At this time, remote testing typically entails some loss of data quality and methodological control. Listeners in web-based experiments may misrepresent themselves (Kan and Drummey, 2018), and certain groups may be underrepresented due to self-selection bias, limiting the generalizability of the research findings (Bethlehem, 2010; Turner et al., 2020). Further, most web-based experiments do not provide listeners with calibrated hardware and thus cannot completely control the stimulation parameters.

Despite these obstacles, we anticipated that the current web-based experiment would yield results consistent with prior lab-based experiments. The findings of tone-scramble experiments are expected to generalize over a reasonably wide range of stimulus levels, ruling out a need for carefully calibrated hardware. The current web-based tone-scramble experiment was conducted in a manner similar to web-based experiments of recently published auditory studies (e.g., Cooke and García Lecumberri, 2021; Kothinti et al., 2021; Viswanathan et al., 2021) which have found general agreement between lab-based and web-based findings, albeit with increased variability in the performance measures in some cases (Merchant et al., 2021). Moreover, researchers have made important progress in developing methods to assess listeners' compliance with instructions (e.g., screening tests for headphone usage; Woods et al.,

2017; Milne et al., 2021), and one such method was used here (Woods et al., 2017).

1.1.2 Does the mixing proportion of low-performing listeners to high-performing listeners depend on native language?

The diversity and size of the sample made accessible by the citizen-science approach of the current study is useful not only with respect to generalizability: such a sample allows us to explore potential relationships between performance in the tone-scramble task and demographic covariates. Such exploration may yield insight into the origin of the bimodal distribution.

Language exposure is especially interesting in this regard. In particular, the literature on tonal-language development raises the possibility that listeners who natively speak a tonal language might perform better on tone-scramble tasks than listeners who do not natively speak a tonal language. A tonal language is one in which pitch is used *lexically*, assigning different meanings to words that consist of the same syllables but different pitch contours (Pike, 1948; van der Hulst et al., 2010). It is well known that a listener's sensitivity to variations in speech sounds is shaped by the listener's exposure to speech during early development (Kuhl, 2004). Moreover, listeners who natively speak tonal languages have been found to be more sensitive to small differences in the lexical pitch contours of their language than listeners who natively speak a non-tonal language (Bidelman and Lee, 2015). Might these language-specific differences in sensitivity to speech sounds be related to differences in sensitivity to the musical variations of tone-scrambles? The same question is worth asking for native speakers of pitch-accented languages (i.e., languages for which pitch contours are used lexically, unlike non-tonal languages, but in a limited fashion; van der Hulst et al., 2010).

Although it has been proposed that musical training can influence language skills (Patel,

2014; Kraus and Chandrasekaran, 2010; Kraus et al., 2014), the data currently available are inconclusive as to whether tonal-language exposure influences musical skills. As a demonstrative example, Pfordresher and Brown (2009) found that native tonal language speakers had an advantage in discriminating musical intervals but not individual notes. Giuliano et al. (2011) found the opposite: native tonal language listeners had an advantage discriminating individual notes but not musical intervals. Pfordresher and Brown (2009) and Giuliano et al. (2011) are two of a number of studies that report conflicting results on the matter of whether native tonal language is associated with greater musical sensitivity. These include studies that suggest that tonal language *enhances* sensitivity to non-speech pitch contours/melodies (Bidelman et al., 2013; Bradley, 2016) and studies that suggest that tonal language *interferes* with sensitivity to non-speech pitch contours/melodies (Bent et al., 2006; Chang et al., 2016). It may be the case that these conflicting results are due to small sample size. In each of these studies, the sample consisted of fewer than 40 listeners. Consequently, these studies may yield spurious and/or non-representative findings with respect to the population of listeners at large. Using a much larger sample of more than 270,000 participants, Liu et al. (2021) found that speaking a tonal language was associated with enhanced performance in a melodic discrimination task but found no such association with performance on other musical tasks (a mistuning perception task and a beat alignment task); pitch-accented languages showed a relatively small advantage over non-tonal languages in all three of these tasks.

There is little reason to believe *a priori* that native language will have any such association with performance on tone-scramble tasks. Adler et al. (2020) have shown that 6-month-old infants generate the same bimodal distribution in 3v**4**-task performance as adults. Research on linguistic development suggests that these infants had yet to reach the end of their critical period for changes in phoneme perception: 6-month-old infants are still capable of discriminating sounds by phonetic category without prior specific language experience, and this ability declines substantially by 12 months of age (Werker and Tees, 1984; Best and McRoberts, 2003; Kuhl et al., 2003; Kuhl, 2004). Typically during this period, listen-

ers' sensitivity to phonetic categories changes from broad to narrow, tuning to the phonetic categories of their exposed language. The fact that 6-month-old listeners have the same distribution in performance as adults on the 3v**4**-task suggests that such perceptual adaptations due to early speech experience are unrelated to the disparity between low-performing and high-performing listeners. By 6 months of age, most infants studied by Adler et al. (2020) were already low-performing on the 3v**4**-task.

1.1.3 Can low-performing listeners hear a difference between major and minor tone-scrambles while failing to label them individually?

It may be the case that many listeners who perform poorly on the standard 3v**4**-task are capable of hearing a difference between the two types of stimuli in that task but are not adequately familiar with the concept of musical scale to identify the two types of stimuli individually. The kinds of music that listeners are most likely to encounter in a casual listening scenario tend to confound scale variations (like major and minor) with variations in timbre and rhythm (for example). Under such conditions, listeners do not need to attend to any single feature category to glean the emotional connotation intended by the music's composer. Consequently, listeners may not learn about the concept of scale despite frequent exposure to music.

To test this possibility, the current study tested listeners on a same/different-task wherein, on each trial, listeners heard a two-interval stimulus consisting of two tone-scrambles either of the same type (e.g., both containing B_5 tones) or of different types (e.g., one containing B_5 tones, one containing $B\flat_5$ tones). An ideal observer whose decision statistic is formed by taking the difference of two noisy input variables may produce more reliably accurate responses than an ideal observer whose decision statistic is formed from just one of those

noisy input variables, assuming the noise of the two input variables is correlated. This is summarized by a well-known identity of probability theory that states that

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\ \mathrm{Cov}(X, Y), \tag{1.1}$$

yielding

$$\mathrm{Var}(X - Y) < \mathrm{Var}(X) \tag{1.2}$$

when

$$\mathrm{Var}(Y) < 2\ \mathrm{Cov}(X, Y). \tag{1.3}$$

By the same token, this kind of comparison strategy may produce more variable responses (and thus lower expected accuracy) when the noise between the two variables is uncorrelated. It is not immediately clear which of these two possibilities might apply to the 3v**4**-task. Can low-performing listeners minimize the influence of noise on their decisions by making comparisons between two tone-scrambles and reporting the presence of a type difference instead of trying to label tone-scrambles by type individually?

While this is an open question, we have good reason to believe that the same/different-task will offer little leverage to low-performing listeners. In their Experiment 2, Chubb et al. (2013) fit a probit model to each listener's data. In this model, the listener responded major on a given trial, k, if

$$\mu(k) + X_k > 0 \tag{1.4}$$

for

$$\mu(k) = W_A A_k + W_C C_k + W_D D_k \tag{1.5}$$

$$+ W_{AC} A_k C_k + W_{AD} A_k D_k \tag{1.6}$$

$$+ W_{CD} C_k D_k + W_{ACD} A_k C_k D_k \tag{1.7}$$

$$+ \text{Bias}, \tag{1.8}$$

where X_k was a standard normal variable, A_k took the value 1 (-1) if the tone-scramble on trial k was major (minor), $C_k = A_{k-1}$, and D_k took the value 1 (-1) if the listener's response to stimulus $k-1$ was "major" ("minor"). Chubb et al. found that the estimates of W_C (among the 69 listeners for which the stable estimates of the model's parameters were available) had an average of 0.2115 which was statistically greater than 0 ($t_{68} = 5.58$; $p < 10^{-5}$). In other words, listeners tended to respond on each trial by mimicking the correct response to the previous trial. This result is consistent with the following possibility: low-performing listeners already *do* compare tone-scrambles in a one-back fashion and, failing to reliably hear a difference between the stimuli of the current trial and previous trial, respond with the correct answer of the previous trial. Given this finding, we anticipated that the same/different-task would offer little advantage to low-performers of the standard 3v**4**-task.

1.2 Methods

1.2.1 Participants

Listeners participated in the experiment by visiting the citizen-science website `https://themusiclab.org` and following a link to the "Are You a Super-Listener?" game. Listeners discovered the website and game by word-of-mouth (e.g., through postings on the Internet

forum Reddit) and volunteered to participate. Listeners were not recruited directly. All participants gave informed consent under an ethics protocol (IRB Protocol: #2000033433) approved by the Committee on the Use of Human Subjects, Harvard University's Institutional Review Board.

The analysis data set was generated by 59,897 presumed-unique listeners. These data represent a subset of the entire data collected. Each website visitor was assigned a unique user-ID number, and any data generated by the visitor was associated with this user-ID. Between February 22, 2021, and April 4, 2022, data were collected from a total of 149,169 visitors with unique user-IDs. Due to limitations on tracking a visitor across multiple devices, browsers, and visits, this count may overestimate the number of unique listeners. For user-IDs with multiple recorded playthroughs, we used only the data from the earliest playthrough. We excluded listeners who reported having played the "Are You a Super-Listener?" game previously but for whom only a single playthrough was associated with their user-ID. We then excluded listeners who did not complete the game in its entirety, as indicated by the number of trials recorded. The remaining data came from 59,897 unique user-IDs, and these user-IDs are presumed to represent unique listeners who were participating in the experiment for the first time.

Headphone usage was not a requirement for participation, and listeners were asked to self-report whether they were using headphones. Listeners who indicated that they were using headphones completed a screening test (Woods et al., 2017) to verify their headphone usage. A total of 41,937 listeners (70.0%) reported wearing headphones and passed the screening test. Listeners who reported that they were using headphones and subsequently failed the screening test proceeded with the experiment as usual.

When asked "Do you have a hearing impairment?", 2.8% of listeners responded "Yes", 84.7% responded "No," and 12.0% responded "I don't know". A majority of listeners (65.3%) reported having taken music lessons. When asked to rate their own skill at making music

Table 1.1: The number of pips of each note in the two types of tone-scramble used in each of the eight conditions for which stimuli were single tone-scrambles. Dots stand for zero. The labels in the top row correspond to thirteen frequencies satisfying $f_k = 783.99 \times 2^{\frac{k}{12}}$ Hz, for $k = 0, 1, \cdots, 12$. Throughout the paper, we will embolden the numbers that refer to major intervals. Note that the name of each condition refers to the distance in semitones between its "target" pips and the pip labeled "0" (G_5).

Task	Type	0	1	**2**	3	**4**	5	6	7	8	**9**	10	**11**	12
1v**2**	1	3	3	·	·	·	·	·	3	·	·	·	·	3
	2	3	·	3	·	·	·	·	3	·	·	·	·	3
2v3	1	3	·	3	·	·	·	·	3	·	·	·	·	3
	2	3	·	·	3	·	·	·	3	·	·	·	·	3
3v**4**	1	3	·	·	3	·	·	·	3	·	·	·	·	3
	2	3	·	·	·	3	·	·	3	·	·	·	·	3
4v5	1	3	·	·	·	3	·	·	3	·	·	·	·	3
	2	3	·	·	·	·	3	·	3	·	·	·	·	3
5v6	1	3	·	·	·	·	3	·	3	·	·	·	·	3
	2	3	·	·	·	·	·	3	3	·	·	·	·	3
8v**9**	1	3	·	·	·	·	·	·	3	3	·	·	·	3
	1	3	·	·	·	·	·	·	3	·	3	·	·	3
9v10	1	3	·	·	·	·	·	·	3	·	3	·	·	3
	1	3	·	·	·	·	·	·	3	·	·	3	·	3
10v**11**	1	3	·	·	·	·	·	·	3	·	·	3	·	3
	1	3	·	·	·	·	·	·	3	·	·	·	3	3

using an instrument or by singing, 18.6% of listeners responded "I'm an expert" or "I have a lot of skill"; most listeners reported having "some skill" (35.7%), being a novice (23.2%), or having "no skill at all" (21.5%).

A total of 210 native languages were represented in the sample, with the most numerous being English (55.7%), Spanish (12.5%), German (3.7%), and French (2.5%). When asked "What is your gender?", 69.6% of listeners responded "Male", 27.8% responded "Female", and 2.2% responded "Other".

1.2.2 Stimuli

A total of nine different conditions were tested in the current study. In eight of these conditions (i.e., all but the same/different-task described below), the stimulus presented on each trial was a tone-scramble comprising 12 consecutive "pips," each pip being a 65-ms pure tone windowed by a raised cosine function with 22.5 ms rise and decay times. Tone-scrambles were thus 780 ms in duration. All pips had equal amplitude. Within each condition, two types of tone-scramble were presented, and the two types differed in their note histogram. Table 1.1 provides the note histograms of the two types of stimuli used in the eight conditions that presented a single tone-scramble per trial. The notes in a given tone-scramble were presented in random sequence.

In the remaining ninth condition (the same/different-task), each stimulus was a pair of tone-scrambles, each generated in the manner described above. The two tone-scrambles were presented in sequence and separated by a 300 ms silent gap. The two types of stimuli in this condition were "same"-stimuli which contained two tone-scrambles with the same note histogram and "different"-stimuli which contained two tone-scrambles with different note histograms. All tone-scrambles in the same/different-task had note histograms matching tone-scrambles in the 3v**4**-task. Thus, in "different"-stimuli, one tone-scramble had 3-pips while the other had **4**-pips; in "same"-stimuli, both tone-scrambles had one of these types of pips. The order of pips in each of the two tone-scrambles in this condition was independently randomized (i.e., the sequence of pips between the two tone-scrambles was different with overwhelming probability).

The stimuli were generated at trial time and played back using the Web Audio API (`https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API`) which is supported by most contemporary web browsers. The audio sampling rate depended on listeners' personal hardware and was not recorded, but given the present standards for audio playback on

consumer devices, it may be presumed that for most listeners, the sampling rate was either 44.1 kHz or 48 kHz. Volume was adjusted manually by each listener to a comfortable level prior to the experiment. Listeners read prompts and entered responses via a video-game interface built with jsPsych (De Leeuw, 2015) and p5.js (`https://p5js.org`).

1.2.3 Design

Each listener was tested in three conditions: the same/different-task, the 3v**4**-task, and one wild-card condition. The wild-card condition presented to any given listener was chosen completely at random from the 1v**2**-, **2**v3-, **4**v5-, 5v6-, 8v**9**-, **9**v10-, and 10v**11**-tasks. Listeners completed three blocks, one for each of the three conditions assigned to them. For each listener, the conditions were ordered such that the wild-card task was always last, and the order of the first two conditions was a random ordering of the same/different-task and the 3v**4**-task. On each trial, listeners were presented a single stimulus (a single tone-scramble, except in the same/different-task where each stimulus comprised two tone-scrambles) and strove to classify it as Type-1 or Type-2. Each block contained 20 trials (10 Type-1 stimuli and 10 Type-2 stimuli) except for the block of the same/different-task which contained 16 trials (four "same"-stimuli with 3-pips, four "same"-stimuli with **4**-pips, four "different"-stimuli with the 3-pips appearing in the first tone-scramble, and four "different"-stimuli with the **4**-pips appearing in the first tone-scramble). The stimuli in each block were ordered completely at random.

1.2.4 Procedure

Intake procedures

Upon accessing the "Are You a Super-Listener?" game via `https://themusiclab.org`, listeners were presented with a study information page listing the university affiliation, IRB protocol number and contact information, and a basic description of the experiment. The study information page also recommended that listeners wear headphones if available. On the following page, a brief survey collected demographic information about the listener, including age, sex, native language, and musical background. In order, listeners were asked to (1) report if they had played the game before, (2) rate their general enjoyment of music using a continuous slider, (3) rate their music listening skills relative to other people using a continuous slider, (4) adjust their volume to a comfortable level as a musical example (the *Super Mario Bros.* theme) was played back, (5) report their gender from the options "Male", "Female", and "Other", (6) report their age in years from a drop-down menu listing integers from 3 to 118, (7) report their country of residence from a drop-down menu, (8) report their native language from a drop-down menu, (9) report whether they spoke another language fluently (and which language, if yes), (10) report whether they had a hearing impairment from the options "Yes", "No", or "I don't know", (11) report whether they can tell if they are out-of-tune while singing, (12) report whether they can tap in time to a musical beat, (13) report whether or not they are wearing headphones (and complete the headphone-screening task by Woods et al., 2017, if yes), and finally if they had reported playing the game previously, (14) whether they achieved "super-listener" status (i.e., whether they achieved a percent-correct of 75% or greater, marginalizing over all three blocks) on their previous attempt.

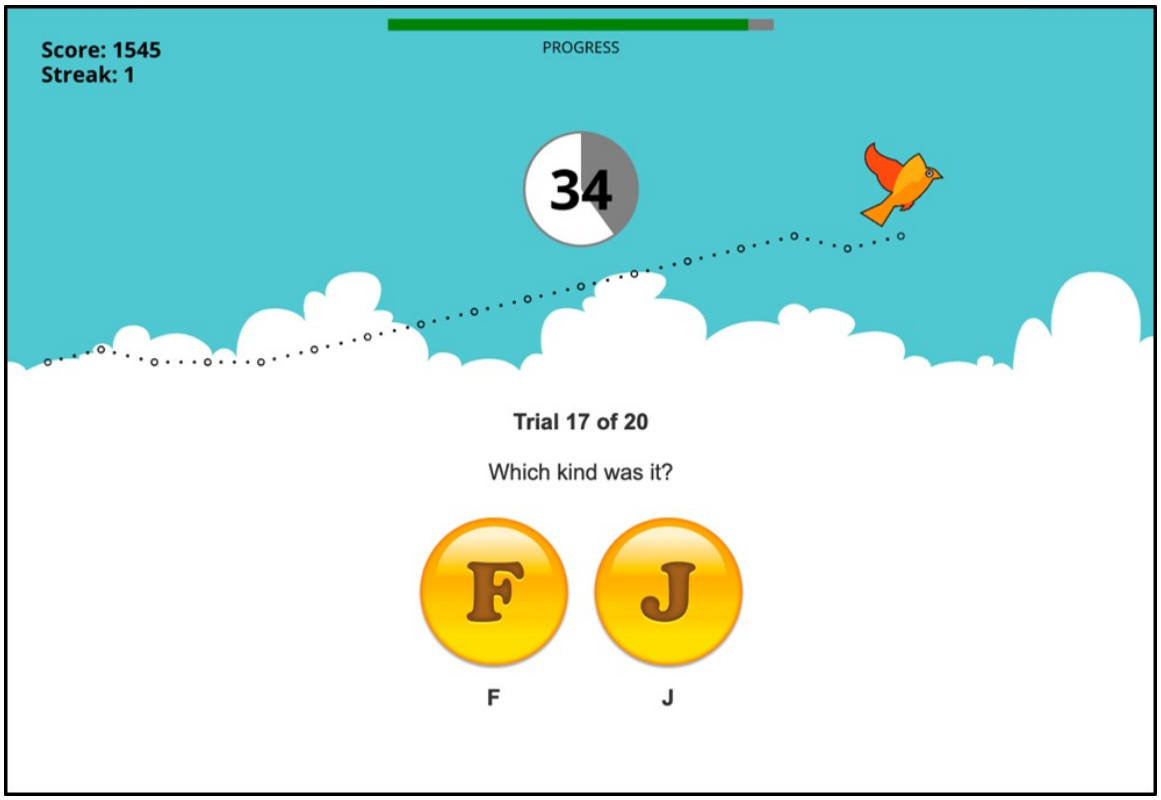

Figure 1.1: A screenshot of the "Are You a Super-Listener?" game showing the game's main visual elements, taken during trial 17 of the third block.

The Super-Listener Game

After completing the introductory survey, listeners began the experiment proper. The experiment was presented to listeners as a video game. A user-interface in the lower half of the screen led each listener through a behavioral experiment, and the visual elements of the video game were presented in the upper half of the screen, updating with the listener's progress through the game. On each trial, a listener who was playing with a keyboard (touch-screen device) used the "F" and "J" keys (on-screen buttons labeled "1" and "2") to label stimuli as Type-1 and Type-2. The mapping of the "F" and "J" keys (the "1" and "2" buttons) to the Type-1 and Type-2 stimuli was randomized for each listener on each block. During the same/different-task, the images used to represent the "F" and "J" keys (the "1" and "2" buttons) were replaced by images bearing the symbols "=" and "≠" (matching the randomly chosen mapping of stimuli to keys/buttons for that listener). After each response, feedback was provided immediately via on-screen messages ("CORRECT" and "INCORRECT") in the lower half of the screen and via animated, colorized video-game elements in the upper half of the screen. The feedback remained on-screen for 700 ms. This was followed by a 150 ms post-trial gap after which the stimulus of the next trial began to play automatically.

At the start of each block, listeners were provided text instructions for that block and a brief training sequence. During the training sequence, listeners were first provided with a labeled example of the stimulus type mapped to the "F" key ("1" button) followed by a labeled example of the stimulus type mapped to the "J" key ("2" button). This was done twice. After each example, the listener was required to press the key (button) corresponding to the given stimulus type to proceed. After four examples, the listener was given a brief practice sequence of four trials, presenting two Type-1 stimuli and two Type-2 stimuli, unlabeled, in random order. Listeners were required to provide responses and received feedback as they would during the test trials of each block. At the end of the four practice trials, the listener chose to proceed with the game or to repeat the training sequence (four examples and four

practice trials). During the training sequence, the visual elements in the upper half of the screen remained in an idle state that did not update.

Once a listener completed the block's training sequence and chose to proceed with the game, the sequence of test trials for that block was initiated, and the game's visual elements in the upper half of the screen began to update with each response from the listener. The game's main visual elements are shown in the upper half of Figure 1.1, which shows what the screen looked like mid-trial for a listener playing with a keyboard. During the game, a bird avatar moved along an invisible grid based on the correctness of the listener's response on each trial. The bird started each block at the bottom-leftmost position of the grid. After each trial, the bird moved rightward one unit on the grid. If the listener's response was correct, the bird also moved up one unit. If the listener's response was incorrect, the bird also moved down one unit unless the bird was at level 0 in which case it stayed at level 0. As the listener completed the last trial of each block, the bird reached the rightmost edge of the grid.

As seen in Figure 1.1, a dotted path was drawn behind the bird to show its progression over the course of the block, and a progress bar at the top-center of the screen showed the listener's progress through the entire experiment. Counters in the top-left corner of the screen showed the listener's current streak (i.e., the number of correct responses the listener had provided since their last incorrect response, reset to 0 at the beginning of each block) and the listener's score. A listener's score was set to 0 at the beginning of the game (i.e., before the first block of the experiment), and the listener's score increased (decreased) after each correct (incorrect) response by the following amount:

$$50 - \text{floor}\big(40 \times \min\{1, r\}\big) \tag{1.9}$$

where r is the response time of the listener (i.e., the time between the end of the stimulus and input of the listener's response) in seconds. In other words, on each correct (incorrect)

trial, the listener's score increased (decreased) by 50 points if the listener responded at the end of the stimulus instantaneously, 10 points if the listener responded more than a second after the end of the stimulus, or some integer value between 50 and 10, decreasing linearly with response times between 0 and 1 second, respectively. Once each stimulus ended and while the game was still awaiting a response from the listener, an animated timer appeared on-screen, showing the current value of the score increment/decrement over the course of the first second of the post-stimulus period. All listeners began the game with a score of 0 which was updated cumulatively over all three blocks of the experiment.

Post-game survey and summary

Upon completing the third block of trials, the listener was presented a second set of survey questions. In *randomized* order, these questions asked listeners to report (1) whether they think they have perfect pitch, (2) whether they have ever taken music lessons (as well as their reason for taking music lessons, their degree of enjoyment of music lessons, and their perceived ability relative to peers, if yes), (3) how often their parent sang to them, (4) their degree of familiarity with traditional music from around the world, (5) the amount of time that they spend making music on an average day, (6) a rating of their skill at making music using an instrument or by singing from the options "I'm an expert", "I have a lot of skill", "I have some skill", "I'm a novice", and "I have no skill at all", (7) whether they have ever experienced "chills" or "goosebumps" in response to music, (8) the amount of time that they spend listening to music or watching videos that include music on an average day, (9) any areas of interest that they believe they have more talent, ability, or training than the average person, (10) their ability to imagine sounds, (11) their ability to imagine a visual scene, (12) whether they currently have any illnesses, disabilities, or health conditions (and what they are, if any), (13) their highest level of education completed, (14) their race, (15) whether they are Hispanic or Latino, and (16) their current household income.

On the final screen, the listener was provided a summary of their performance. This summary included the listener's final score, the listener's percentile relative to other players' scores, the listener's position on a density plot representing the distribution of percent-correct on the 3v**4**-task observed in prior lab-based experiments, and the listener's percent-correct achieved in each block represented as a bar plot with the percent-correct of the average player in each condition superimposed. If the listener was correct on 75% of trials overall or more, this screen told the listener that they are "A SUPER-LISTENER"; otherwise it told the listener that they are "NOT A SUPER-LISTENER".

1.3 Results

1.3.1 Binomial-mixture model of proportion-correct on the 3v4-task

Previous lab-based studies using tone-scrambles have found that the 3v**4**-task partitions adult listeners into two distinct groups: one group that performs the task with near-perfect accuracy ($\approx$ 30% of listeners) and one group that performs the task near chance ($\approx$ 70% of listeners) (Chubb et al., 2013; Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020; Adler et al., 2020; Ho et al., 2022). This finding is represented in Figure 1.2 where the distribution of proportion-correct in the 3v**4**-task pooled across four prior lab-based studies (Chubb et al., 2013, Dean and Chubb, 2017, Mednicoff et al., 2018, Ho et al., 2022) is shown in red. These proportions-correct are based on the last 50 trials that each listener performed which were, in all cases, preceded by at least 40 practice trials.

Does this finding generalize to a large web-based sample? Figure 1.2 shows in blue the corresponding distribution of proportion-correct observed in the web-based sample. Visual inspection suffices to show that both the lab-based and web-based distributions are bimodal,

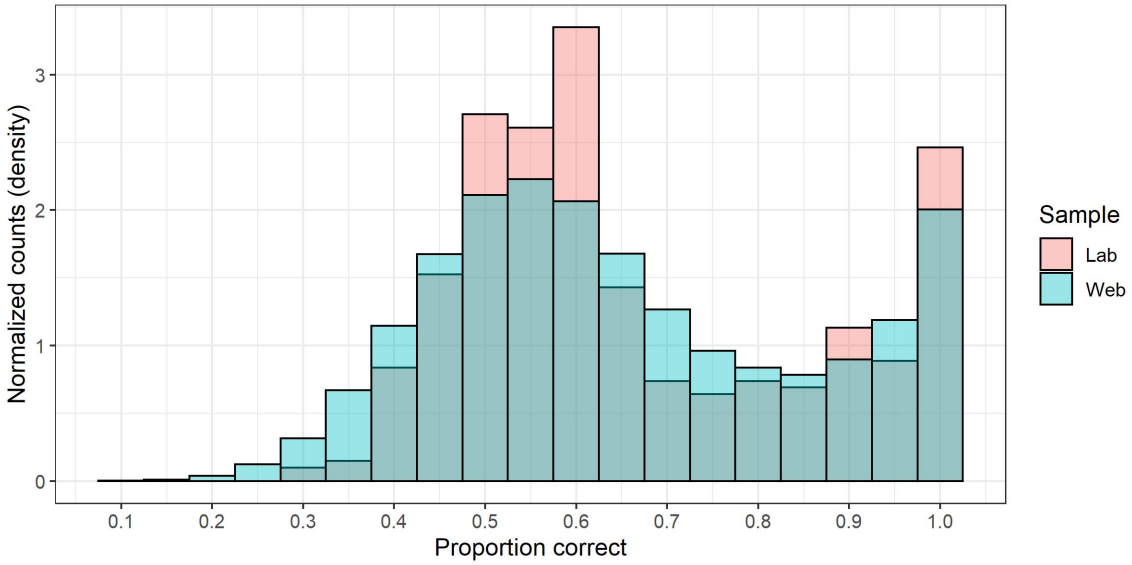

Figure 1.2: (Red) The distribution of proportion-correct in the 3v**4**-task, pooled over Chubb et al. (2013), Dean and Chubb (2017), Mednicoff et al. (2018), Ho et al. (2022) and based on the last 50 trials that each listener performed (in all cases, preceded by at least 40 practice trials). (Blue) The distribution of proportion-correct in the 3v**4**-task observed in the web-based sample of the current study, based on all 20 trials for each listener. For ease of comparison, the lab-based and web-based distributions have been normalized to have an area of 1 and are shown superimposed.

with the majority of listeners belonging to the mode near chance (i.e., a proportion-correct of 0.5).

We also note, however, that the lab-based distribution is more sharply sculpted (with higher peaks at the two prominent modes and a deeper valley between them) than the web-based distribution. This is to be expected for the following reason: the estimates of proportion-correct from the web-based study are noisier than those from the lab-based studies because they are based on only 20 (as opposed to 50) trials. This added noise will soften the contours of the web-based distribution in comparison to the lab-based distribution.

A binomial-mixture model may be used to estimate the relative mass attributed to the upper and lower modes of these distributions and the mean correct-response rate of each group. This model has the form

$$P(Y_s = y) = \alpha \binom{n}{y} (1-p)^{n-y} p^y + (1-\alpha) \binom{n}{y} (1-q)^{n-y} q^y \tag{1.10}$$

where s indexes the listener, Y_s represents the count of correct responses achieved by subject s, and $n = 20$, the maximum number of correct responses possible. According to this model, each listener's count of correct responses is drawn from one of two binomial distributions with success probabilities p and q, respectively, with probability α that the count is drawn from the former distribution (note that all parameters of this model are fixed across all listeners, so individual effects are not captured by this model). Without constraint, these parameters are not identifiable (an equivalent model can be produced by setting α equal to $1 - \alpha$ and swapping p and q). Thus, we imposed the constraints

$$0 < p \leq \frac{7}{10} \tag{1.11}$$

and

$$\frac{7}{10} \leq q < 1. \tag{1.12}$$

Under these constraints, we may interpret p (q) as the probability that a low-performing (high-performing) listener will produce a correct response on any one trial, and we may interpret α as the proportion of low-performing listeners out of all listeners.

We fit the binomial-mixture model to the pooled lab-based sample and the web-based sample separately. The binomial-mixture model has a total of 3 free parameters (α, p, and q). We estimated these parameters with Bayesian methods, assuming the following priors: $\alpha \sim$ Uniform(0, 1), $p \sim$ Uniform(0, 0.7), and $q \sim$ Uniform(0.7, 1). For each fit, point estimates were calculated by taking the median of 200,000 MCMC samples, thinned to every 100^{th} sample. These were drawn after first taking 200,000 burn-in samples. We found that these hyperparameters were sufficient to mitigate issues of autocorrelation among posterior samples. Note that for the lab-based sample, the maximum count of correct responses was $n = 50$ whereas for the web-based sample $n = 20$.

Figure 1.3 shows the distribution of posterior samples returned by MCMC for the binomial-mixture-model parameters when fit to the pooled lab-based data (red) and the current web-based sample (blue). The median posterior estimate for the proportion of listeners belonging to the low-performing group in the lab-based sample (α_{lab}) was 0.724 with 95% credible interval [0.677, 0.768]. The corresponding posterior estimate for the web-based sample was 0.743 with 95% credible interval [0.739, 0.747]. Listeners of the low-performing group in lab-based studies were estimated to have an expected proportion-correct (p_{lab}) of 0.561 with 95% credible interval [0.553, 0.570]; the low-performing group of the present web-based sample was estimated to have a correct-response rate of 0.556 with 95% credible interval [0.555, 0.557]. Listeners of the high-performing group in lab-based studies were estimated

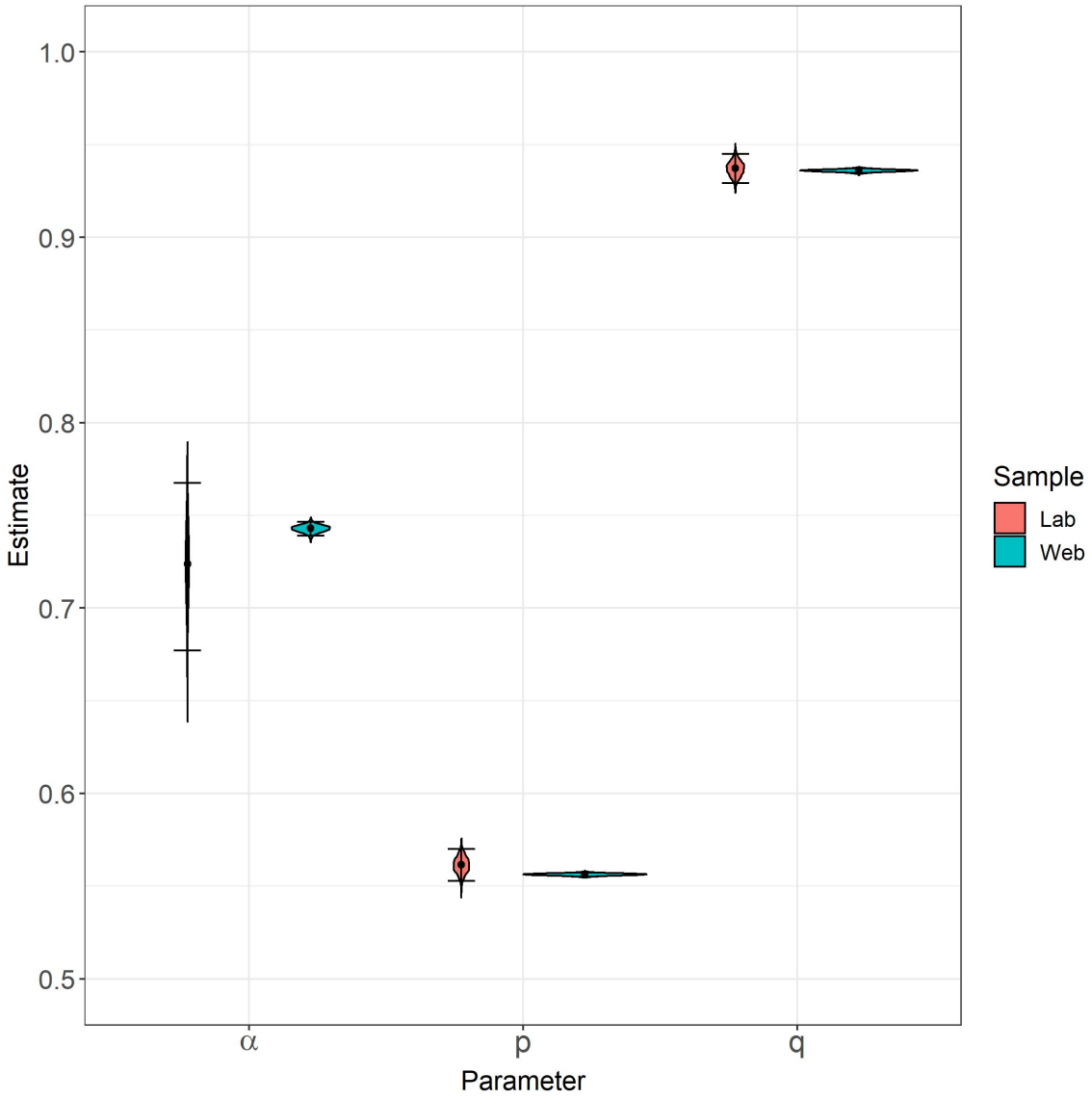

Figure 1.3: Violin plot of the approximate posterior distributions of α, p, and q of the binomial-mixture model for the pooled lab-based sample (red) and the web-based sample (blue). Points represent the median posterior estimate for each parameter, and error bars represent 95% Bayesian credible intervals.

to have a correct-response rate (q_{lab}) of 0.937 with 95% credible interval [0.929, 0.945]; the high-performing group of the present web-based sample was estimated to have an expected proportion-correct of 0.936 with 95% credible interval [0.935, 0.937]. Due to the marked difference in sample size ($N_{\text{lab}} = 406$ whereas $N_{\text{web}} = 59,897$), the posterior distribution for parameters had greater spread when the binomial-mixture model was fit to the lab-based sample, and consequently, the 95% Bayesian credible intervals for the lab-based sample were much wider. The median posterior estimates for the web-based sample fell within the lab-based credible intervals for all three parameters, indicating a strong correspondence between the lab-based and web-based results.

1.3.2 Comparing the mixture of low- and high-performing listeners between native languages

To assure that estimates of the proportion of low-performing listeners within each language were precise, we restricted our analysis to native languages for which our sample contained more than 40 listeners. Among the 210 languages represented in the sample, 60 languages met this criterion for inclusion. Among these languages, the minimum, first quartile, median, third quartile, and maximum of the number of listeners with a given native language were, respectively, 42, 79.5, 151.5, 329.2, and 33,377.

For each of the included languages, we fit the binomial-mixture model described in Section 1.3.1 to the distribution of trials-correct within each language. This yielded an estimated mixing proportion α_ℓ for each language, ℓ. As above, point estimates were calculated by taking the median of 200,000 MCMC samples, thinned to every 100^{th} sample and drawn after first taking 200,000 burn-in samples. The median posterior estimate for each α_ℓ and the 95% credible interval for each estimate is visualized in Figure 1.4. Across all languages, the binomial-mixture model infers that at least half of all listeners, if not the majority,

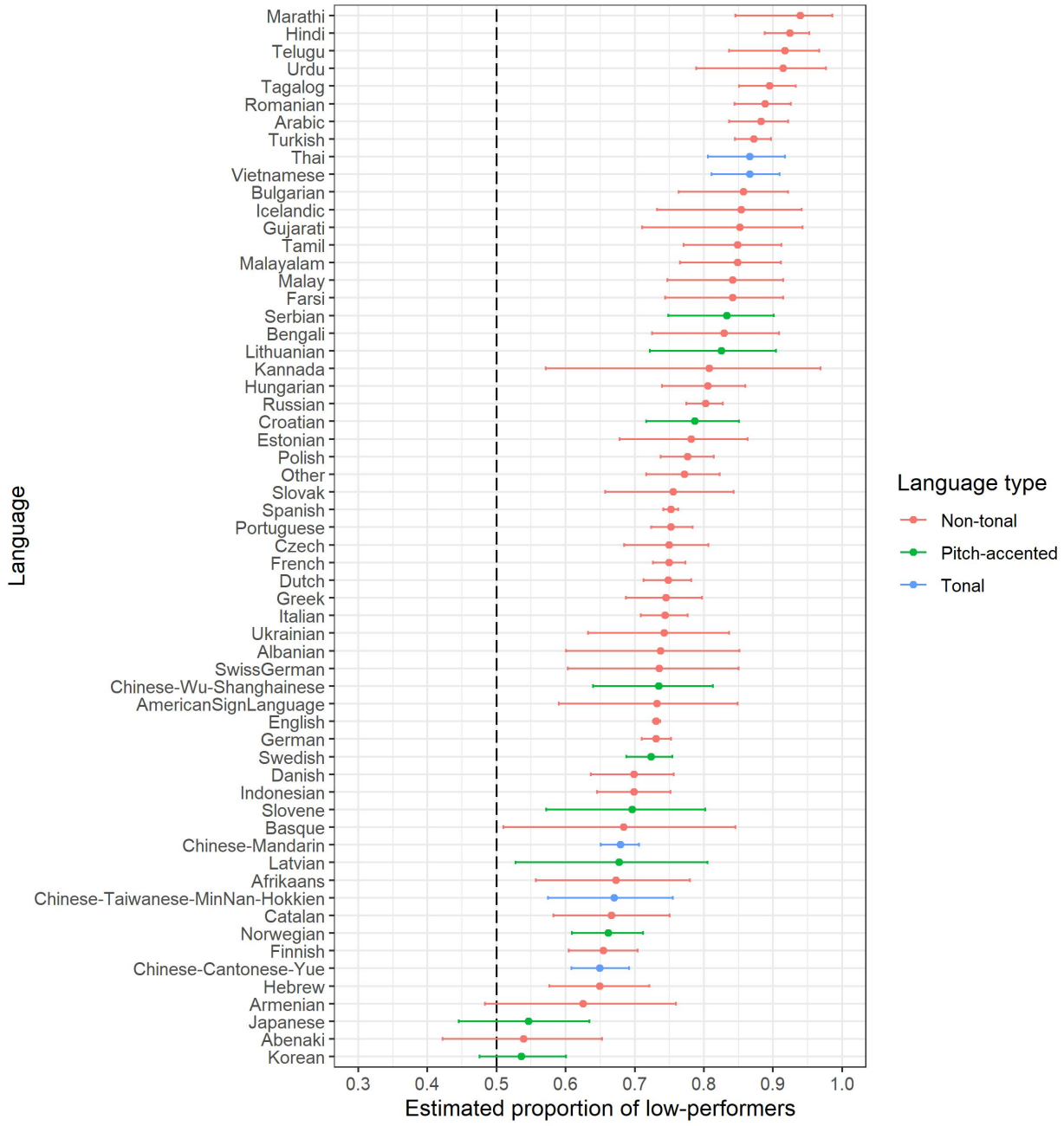

Figure 1.4: Plot comparing the median posterior estimate of α for the binomial-mixture model described in Section 1.3.1 fit to each language with ≥ 40 listeners. Error bars represent 95% Bayesian credible intervals, and colors represent language type (i.e., the degree of tonal significance within a particular language). The dashed line represents $\alpha = 0.5$, i.e., an equal number of high- and low-performing listeners within a language.

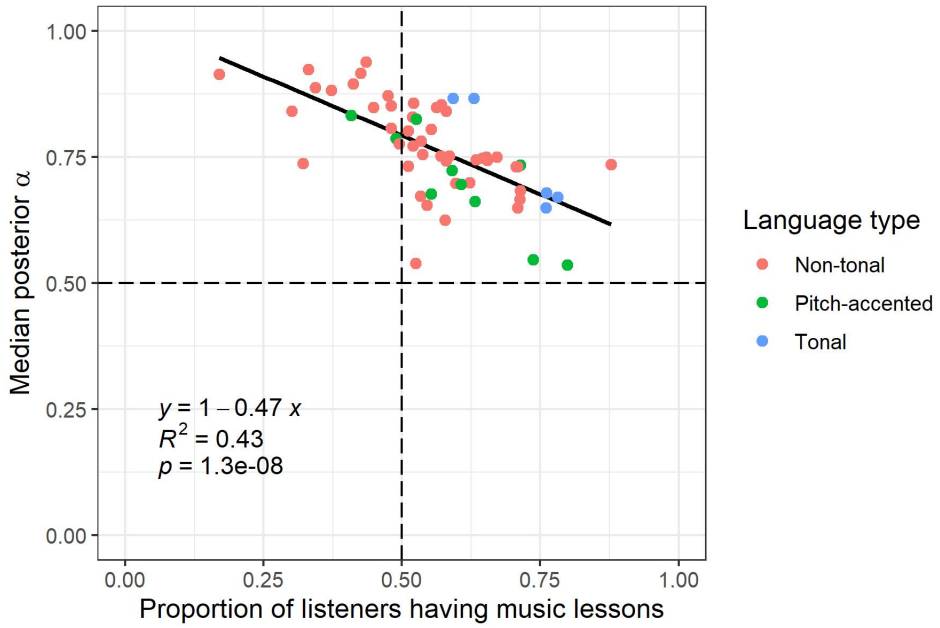

Figure 1.5: Scatterplot visualizing the relationship between α (y-axis) and the proportion of listeners having music lessons (x-axis) in a given language. Each point represents one language, and the colors of points indicate the language type (i.e., the degree of tonal significance within a particular language). The regression line best fitting these data is superimposed, and the details of this regression line are provided by the inset text.

are low-performing. There is a considerable degree of variability in α_ℓ, with the extreme languages, Korean and Marathi, having an estimated α of 0.536 and 0.939, respectively (indicating that the sample of Marathi listeners had a higher proportion of low-performing listeners). It is not clear that this variability is related to language type (i.e., the degree of tonal significance within a particular language).

Might differences in α_ℓ be due to confounding variables? Figure 1.5 shows the relationship between α_ℓ and the proportion of listeners who received music lessons within each language, and Figure 1.6 shows the relationship between α_ℓ and the proportion of listeners who self-reported a high degree of musical skill (defined as responding "I'm an expert" or "I have a lot of skill" when asked to rate their skill at making music using an instrument or by singing). There is a strong negative relationship between α_ℓ and both the proportion of listeners who received music lessons and the proportion of listeners with self-reported high musical skill.

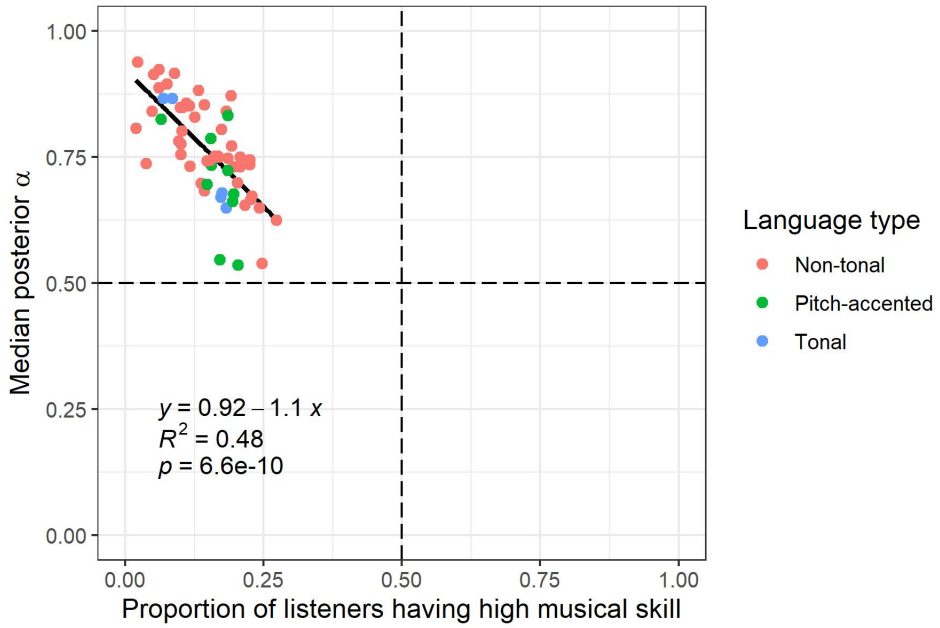

Figure 1.6: Scatterplot visualizing the relationship between α (y-axis) and the proportion of listeners having high self-reported musical skill (x-axis) in a given language. Each point represents one language, and the colors of points indicate the language type (i.e., the degree of tonal significance within a particular language). The regression line best fitting these data is superimposed, and the details of this regression line are provided by the inset text.

This suggests that any effect that native language may appear to have on α (e.g., as might be inferred by Figure 1.4) may in fact be more a matter of the specific subset of listeners sampled from that native language rather than an effect of native language itself.

To adjust for these effects, the α_ℓ estimates were modeled using a simple multiple linear regression model of the form

$$\alpha_\ell = \beta_0 + \beta_1 p_{\ell,1} + \beta_2 p_{\ell,2} + \varepsilon_\ell \tag{1.13}$$

with

$$\varepsilon_\ell \sim N(0, \sigma^2) \tag{1.14}$$

where ℓ indexes the native language, $p_{\ell,1}$ is the proportion of listeners of native language

ℓ who received music lessons, $p_{\ell,2}$ is the proportion of listeners of native language ℓ who had high self-reported musical skill, and β_0, β_1, β_2, and σ are the four free parameters of the model. Fitting this model using least-squares, β_0 was estimated to be 1.02 with 95% confidence interval [0.94, 1.09], representing the predicted α_ℓ for a language ℓ with no listeners of high musical skill and no listeners with music education; β_1 was estimated to be -0.26 with 95% confidence interval [-0.42, -0.10], representing the estimated change in α_ℓ for each unit change in the proportion of listeners with music lessons; β_2 was estimated to be -0.74 with 95% confidence interval [-1.09, -0.38], representing the estimated change in α_ℓ for each unit change in the proportion of listeners with high self-reported musical skill.

To adjust for music lessons and self-reported musical skill, the predicted α values given by the fitted multiple linear regression model were subtracted from the raw estimated α values (i.e., the points represented in Figure 1.4) and the credible interval around each estimate was translated accordingly. The resulting residual α values and their shifted credible intervals are shown in Figure 1.7. For 43/60 languages, the translated credible interval contains zero. In other words, for the majority of languages analyzed, the proportion of listeners with music lessons and the proportion of listeners with self-reported musical skill are sufficient to fit the estimated α values within the margin of their 95% credible intervals. While a number of languages still have residual α values that deviate substantially from zero, this result suggests that, overall, the variation across languages observed in Figure 1.4 is not due to native language itself.

1.3.3 Comparing the facilitation strengths of a single-resource model between native languages

Previous studies using tone-scramble stimuli have used a "single-resource" model to describe the pattern in performance across listeners and to assess the relative difficulty of different

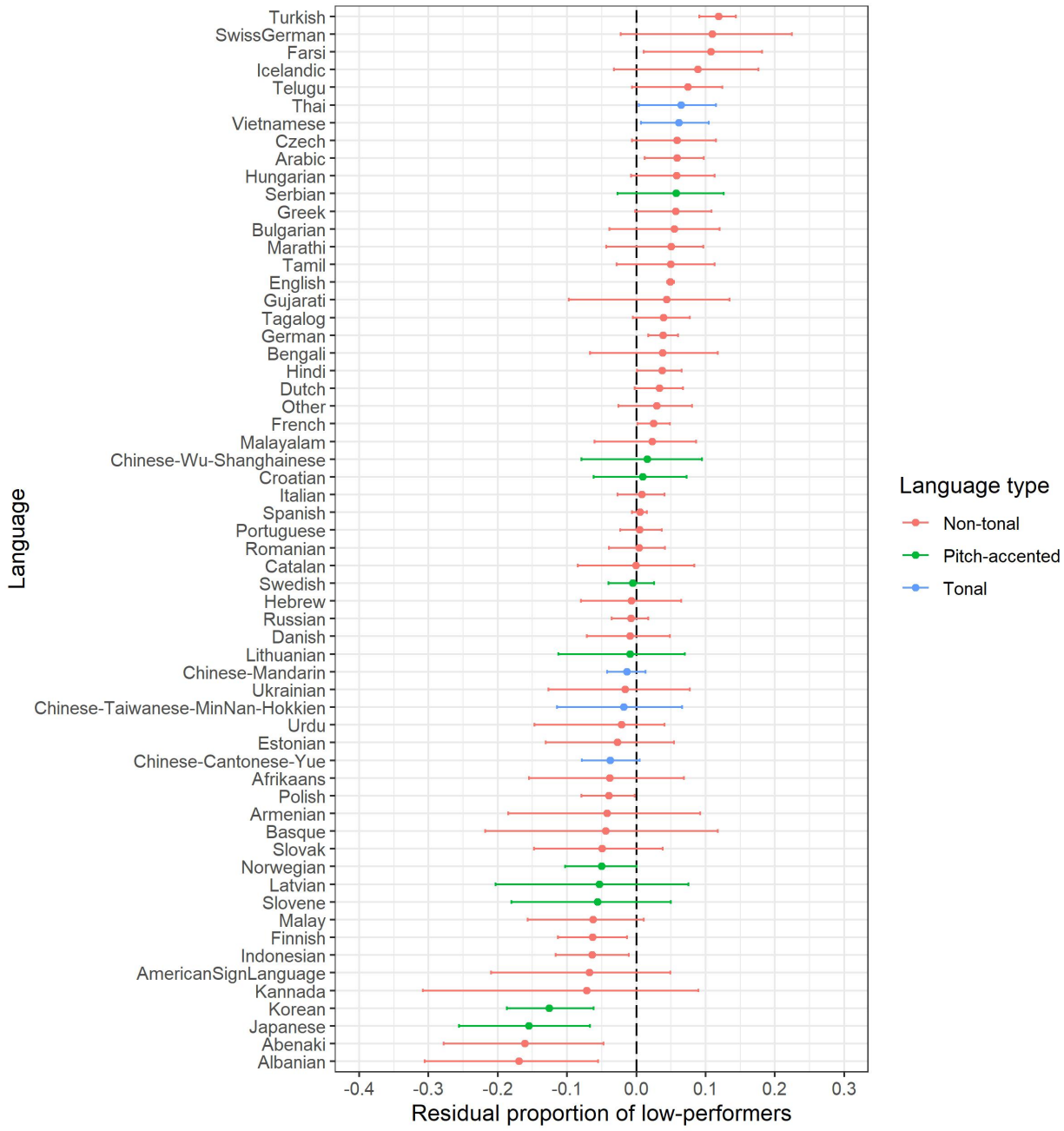

Figure 1.7: Plot of the residual α values calculated by taking the α estimates returned by fitting the binomial-mixture model (as plotted in Figure 1.4) and subtracting the predicted α given by the fitted multiple linear regression model in Equation 1.13. Error bars represent the 95% Bayesian credible intervals, translated according to the subtraction of predicted α values. Color represents language type (i.e., the degree of tonal significance within a particular language). The dashed line represents residual $\alpha = 0$, i.e., an α value that is perfectly predicted by the multiple linear regression model.

conditions. Unlike the analysis presented thus far, the dependent variable representing performance in the single-resource model is d' of signal detection theory. For each listener in each condition (excluding the same/different-task), we used all 20 trials to calculate d'. To calculate d' under our experimental paradigm, we needed to label trials as "signal" and "noise" trials arbitrarily. We chose to treat trials with Type-2 (Type-1) stimuli as signal (noise) trials. This choice did not influence our d' measures. Each listener provided two data points: one d' measure for the 3v**4**-task, and one d' for whichever wild-card task that they completed. If a listener achieved a perfect hit rate (i.e., 10 hits out of 10 signal trials) in a given condition, the proportion of hits was set to $\frac{10-0.5}{10} = 0.95$ (as recommended by Macmillan and Kaplan, 1985). Analogous adjustments were made for correct rejection rates. Given these adjustments, the maximum observable d' was 3.29.

This maximum d' was substantially lower than the maximum observable d' in prior experiments where, for example, a total of 100 trials (50 signal trials and 50 noise trials) would lead to a maximum observable d' of 4.65 under the adjustments of Macmillan and Kaplan (1985). Importantly, listeners in prior lab-based studies achieved such d' values, so the maximum *observable* d' of 3.29 was likely to underestimate the maximum *achievable* d' for many listeners. It was found that replacing all of the d' estimates at the maximum observable value in the present experiment (3.29) with the maximum observable value given 100 trials (4.65) did not meaningfully change the parameter estimates for the single-resource model and did not change the results of the analysis presented below.

The single-resource model that was fitted to the d' values across listeners had the following form:

$$d'_{s,t} = R_s F_t + \varepsilon_{s,t} \tag{1.15}$$

subject to the constraint

$$\sum_{t=1}^{T} F_t = T, \tag{1.16}$$

where s indexes the listener, t indexes the condition (in order, the 1v**2**-task, **2**v3-task, 3v**4**-task, **4**v5-task, 5v6-task, 8v**9**-task, **9**v10-task, and 10v**11**-task; thus, $T = 8$), and $\varepsilon_{s,t} \sim N(0, \sigma^2)$ independently and identically distributed (iid). According to this model, a single latent cognitive resource R governs performance in all conditions. Different listeners possess different levels of R (hence R_s), which facilitates performance in condition t with relative strength F_t. Note that F_t is fixed across listeners, so aside from measurement error, variation between listeners is attributed only to the amount of variation in the amount of R possessed by listeners.

In the present study, unlike any prior tone-scramble study, several of the conditions were tested between subjects, presenting a complication in fitting the single-resource model. To explain the complication, consider the matrix containing all of the d' estimates for native English speakers. In this matrix, each row corresponds to one listener (yielding 59,897 rows for the entire sample), and each column corresponds to one condition (yielding 8 columns). Excluding the same/different-task, every listener was tested in just two conditions: the 3v**4**-task and one randomly selected wild-card task. Consequently, each row of the matrix contains six missing values and only two known values. Thus, even though the single-resource model provides predictions for every combination of s and t, our likelihood function is not simply expressed as a product over every combination. To resolve this issue, we used a fitting procedure described in Section A based on conditional maximization which allowed us to compute maximum-likelihood estimates and frequentist 95% confidence intervals for each parameter.

In practice, we found that a relatively large sample size was required for the conditional

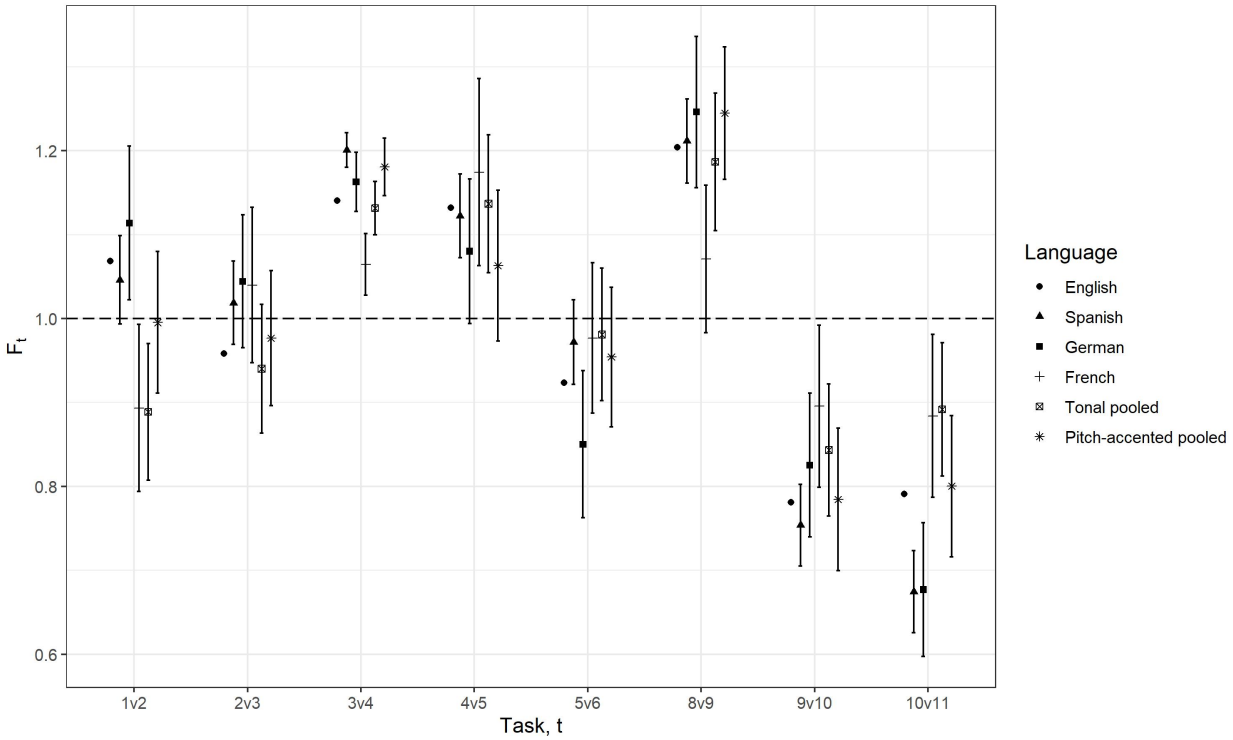

Figure 1.8: Estimated F_t values for each task condition t for each of 6 native-languages: English, Spanish, German, French, tonal languages (pooled), and pitch-accented languages (pooled). The dashed line represents $F_t = 1$, which is the mean value of F_t across all tasks t. Error bars represent 95% frequentist confidence intervals, applying the Bonferroni correction within each language. In other words, for each language, the confidence intervals, taken together, represent a simultaneous confidence region for which the probability of falsely rejecting at least one null hypothesis for one of the parameters is 0.05. Error bars for English are omitted due to issues of computational tractability (addressed in Section A); however, the error bars may be assumed to be very narrow given that the sample size for English is an order of magnitude larger than any other subset.

maximization procedure to converge and produce useful estimates. Thus, we fit the bilinear model to each of the six following subsets of listeners: native English speakers (N = 33,377), native Spanish speakers (N = 7,506), native German speakers (N = 2,218), native French speakers (N = 1,508), pooled tonal-language speakers (N = 2,265), and pooled pitch-accented-language speakers (N = 1,995). All fits converged within 25 iterations or fewer of conditional maximization. Convergence was defined as a cumulative absolute change in parameter estimates smaller than 10^{-6} from one iteration to the next. Figure 1.8 compares the resultant estimates of F based on each of these estimates. Under the imposed constraint that $\sum_t F_t = 8$, we may interpret any condition t for which $F_t > 1$ as a condition that is easier than the average condition, and similarly, any condition t for which $F_t < 1$ as harder than the average condition. Overall, we find that the different subsets perform very similarly across the variety of conditions tested. Conditions that are easier on average for one language tend to be easier on average for all languages. Notably, the 3v**4**-task and 8v**9**-task tended to be easier across all language subsets while the 10v**11**-task tended to be a more difficult task across all language subsets. A similar result was obtained by Dean and Chubb (2017).

1.3.4 Performance on the same/different-task among listeners with low performance on the 3v4-task

Using the parameter estimates of the binomial-mixture model presented in Section 1.3.1, we may select an informed criterion for categorizing listeners as low-performing versus high-performing. Figure 1.9 visualizes the *predicted* distribution of counts based on the median posterior for the parameters of the binomial-mixture model fit to the web-based sample in Section 1.3.1. The visualization is such that a single point on either curve represents the joint probability that an observed count has that value *and* belongs to that group. Thus, given the value of a particular count, the most likely group that that count belongs to is the

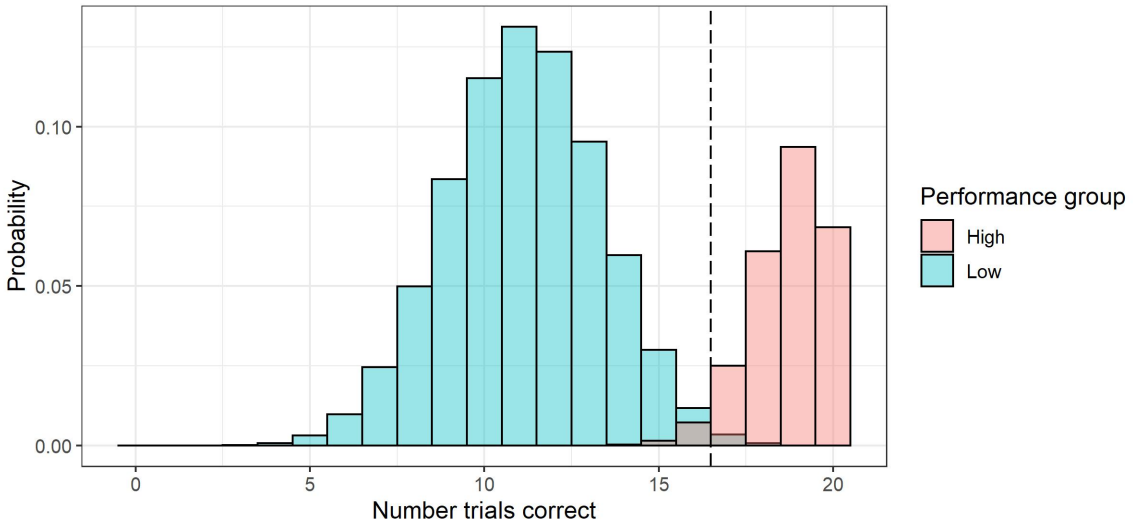

Figure 1.9: Predicted distribution of counts for the web-based sample based on the median posterior estimates of α, p, and q reported in Section 1.3.1. Note that the probability mass functions of the high-performing (red) and low-performing (blue) groups are weighted by α and $1 - \alpha$ respectively such that areas under the two curves summed together equal 1. In other words, a single point on either curve represents the joint probability that an observed count has that value *and* belongs to that group.

group which has a higher probability for that count in Figure 1.9. Accordingly, the informed criterion may be placed at the point where these two probability mass functions cross. By this reasoning, all listeners with 16 or fewer responses correct in the 3v**4**-task were labeled as "low-performing" (45,299 listeners), and all listeners with 17 or more correct responses were labeled as "high-performing" (14,598 listeners).

In Figure 1.10, low-performing listeners are sorted into bins according to their particular combination of number-correct on the 3v**4**-task and number-correct on the same/different-task. The resultant figure is a heat map representing the bivariate distribution of listeners' performance on the two tasks. If it is the case that low-performing listeners can hear a difference between major and minor tone-scrambles but cannot identify them as major or minor individually, we should observe that the greatest mass of listeners falls above the dashed line representing equal accuracy on the two tasks. This is not the case. Low-performing listeners tended to perform similarly on the same/different-task as they do on

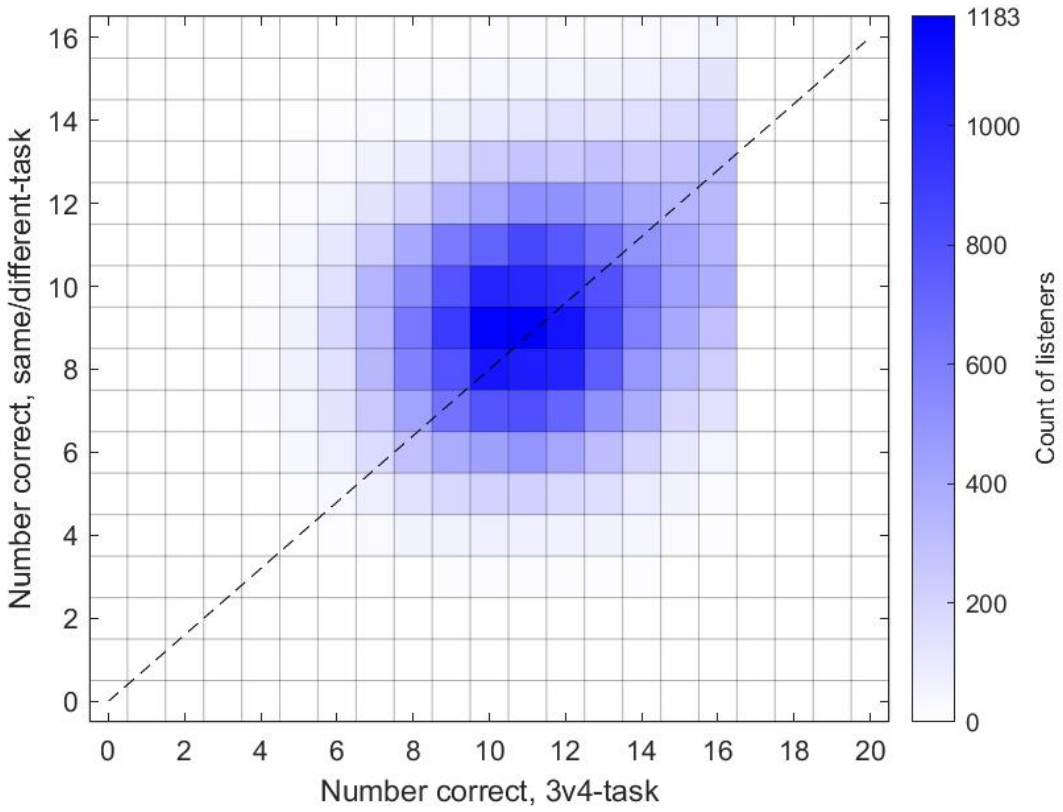

Figure 1.10: Heat map representing the bivariate density of low-performing listeners' number-correct on the 3v**4**-task (x-axis) and number-correct on the same/different-task (y-axis). The color of each bin represents the number of listeners who achieved that particular combination of numbers-correct on the two tasks. A dashed line indicates where the proportions-correct on the two tasks are equal. Note that the dashed line is not identical to the $x = y$ line due to the different total number of trials between the two tasks. Since low-performing listeners are defined as those listeners who provided a correct response on 16 or fewer trials of the 3v**4**-task, no listeners are contained in bins where $x >= 17$.

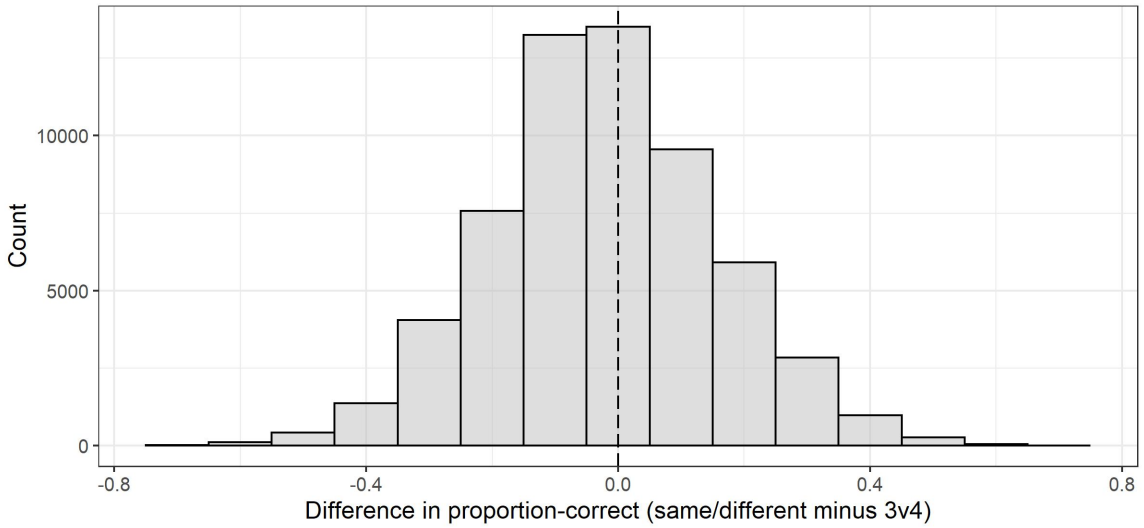

Figure 1.11: Histogram of the difference in proportion-correct across low-performing listeners, subtracting the proportion-correct on the 3v**4**-task from the proportion-correct on the same/different-task. The dashed line indicates a difference of zero.

the 3v**4**-task.

Figure 1.11 shows the distribution of differences in proportion-correct comparing the same/different-task to the 3v**4**-task. As might be inferred from Figure 1.10, the mode of this distribution is near 0. A two-tailed one-sample t-test of the null hypothesis that the mean difference is exactly equal to 0 yields statistically significant results ($t_{59,896} = -21.482$ and $p < 2.2 \times 10^{-16}$) providing evidence that the mean difference is not equal to zero; however, the mean difference is -0.016 and is not practically different than zero. This result suggests that for low-performing listeners, the same/different-task is very slightly more difficult than the 3v**4**-task.

1.4 Discussion

1.4.1 Lab-based findings using tone-scrambles generalize to a large web-based sample

The results reported in Section 1.3.1 show a strong correspondence between the current web-based sample and the results obtained with prior lab-based samples recruited via institutional subject pool. In Figure 1.2, we see that the distributions of proportion-correct in the 3v**4**-task observed in the lab and in the current web-based sample were nearly identical. The lower mode appeared to be a bit wider for the web-based sample, but this may be attributed to the lower number of trials in the web-based sample (20 trials) than in the lab-based sample (50 trials). Fitting the binomial-mixture model, we found that the estimated mixture α of low-performing listeners in the web-based sample was 0.743, well within the 95% credible interval for α of the lab-based sample [0.677, 0.768] as seen in Figure 1.3. The estimated correct-answer rate within each of the low- and high-performing groups for the web-based sample, 0.556 and 0.936, were also within the 95% credible interval for the respective correct-answer rates in the lab-based sample, [0.553, 0.570] and [0.929, 0.945]. Although not all auditory phenomena are appropriate to study remotely, this study contributes to a growing number of examples of auditory phenomena for which web-based experiments yield similar results to lab-based studies (Cooke and García Lecumberri, 2021; Kothinti et al., 2021; Viswanathan et al., 2021).

Given the greater diversity of the web-based sample studied here compared to that of earlier studies, we conclude that the effects that have been reported in the tone-scramble literature thus far are not specific to the samples being studied. This conclusion is unsurprising given that past work has found only a modest association between demographic factors like years-of-musical-training and performance in tone-scramble tasks (Dean and Chubb, 2017;

Mednicoff et al., 2018; Ho and Chubb, 2020).

These results not only allow us to generalize the conclusions drawn in earlier tone-scramble studies but also provide evidence to support a number of methodological improvements to the general design used in tone-scramble studies. For one, the tone-scrambles of the present experiment were shorter in duration (0.78 s) than half that of tone-scrambles used in prior studies (2.08 s) by virtue of having three pips per note (Table 1.1) instead of eight pips per note. The strength of agreement between the current results and prior lab-based studies provides evidence that these shorter tone-scrambles can be used to study disparity in performance on tone-scramble tasks without a reduction in effect size. Using shorter tone-scrambles, researchers can test listeners in more trials and more conditions in a single sitting than would be possible using the longer tone-scrambles of prior studies. Moreover, the fact that these data were collected using a web-based paradigm shows that tone-scramble effects can be reliably measured using uncalibrated hardware and an automated procedure in the absence of a research assistant. This makes a larger and more diverse pool of listeners available to participate in these studies, providing us greater leverage in uncovering the reason why some listeners perform well in tone-scramble tasks and others do not. It also suggests that tone-scramble tasks may be a good candidate for use in a quick diagnostic procedure that is performed by an automated program on listeners' personal devices. For example, such procedures are of great interest to researchers in the area of self-fitting hearing aids (Vyas et al., 2022; Boothroyd and Mackersie, 2017). Although we have yet to identify a clinically relevant variable that is associated with performance in tone-scramble tasks, it is clear that one's status as a low-performing or high-performing listener can be assessed quickly and reliably without the presence of a researcher/audiologist or specialized equipment. Thus, the study and treatment of any clinical outcome that is found to be associated with performance in tone-scramble tasks would benefit greatly from the use of tone-scrambles as a diagnostic.

1.4.2 Mixing proportion of low- and high-performing listeners varies with native language, but this may be explained by differences in musical experience

In each of the native languages studied (including tonal and pitch-accented languages), low-performing listeners formed at least half – if not the majority – of listeners represented by the sample. This is seen in Figure 1.4, where the α estimates across languages all fell to the right of the line $x = 0.5$. Taken alone, this result suggests that native language does not have a critical a role in determining whether a listener is low-performing or high-performing. If native language were a critical determinant of performance, one would expect some languages to have a greater proportion of high-performing listeners than low-performing listeners. This agrees with previous findings that the bimodal distribution in performance on tone-scramble tasks is observed among 6-month-old infants (Adler et al., 2020), prior to the critical period in linguistic development where listeners lose sensitivity to phonetic categories that are not relevant to their native language (Werker and Tees, 1984; Best and McRoberts, 2003; Kuhl et al., 2003; Kuhl, 2004). The nature of performance in tone-scramble tasks is thus distinct from that of other abilities like absolute pitch (Deutsch et al., 2006) and melodic discrimination (Liu et al., 2021) where native language is implicated.

Although native language is not a *critical* determinant of tone-scramble performance, the degree of variability in α across native languages in Figure 1.4 is notable, with the estimated proportion of low-performing listeners among speakers of the same native language ranging from 0.536 (Korean) to 0.939 (Marathi). It is not clear that these variations were due to native language because language type (non-tonal, pitch-accented, and tonal) did not appear to predict the estimated α for a particular language. For example, tonal languages Thai and Vietnamese fell into the highest quintile of estimated α while tonal Chinese dialects Mandarin, Taiwanese Hokkien, and Cantonese fell into the lowest quintile. Given that listeners

discovered the game by word-of-mouth and volunteered to participate, it is possible that listeners of a particular native language tended to have shared characteristics that differed from the sample at large and that these characteristics predicted performance in the tone-scramble task. As an example, a listener of native language L may have been a highly skilled musician whose colleagues tended to also be both highly skilled musicians and native speakers of L. If this listener directed their colleagues to the experiment, then the proportion of highly skilled musicians among listeners representing native language L would have increased relative to the baseline. Consequently, the proportion of low-performing listeners in language L was likely to decrease, and this deviation from baseline would not necessarily have been due to native language L or the proportion of highly skilled musicians who are speakers of native language L at large; rather, it would have been due to confounds introduced by the form of convenience sampling used here.

As shown in Figure 1.5 and Figure 1.6, respectively, there was a strong negative association between a language's estimated proportion of low-performing listeners and both (1) the proportion of listeners who had taken music lessons and (2) the proportion of listeners who reported having a high level of musical skill. The negative association indicates that as the proportion of listeners with music lessons or high musical skill increased among speakers of a native language, the proportion of low-performing listeners decreased. This effect is concordant with the modest association (at the level of individual listeners) between performance in tone-scramble tasks and years of musical training observed in prior studies (Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020). As shown in Figure 1.7, most effects of language did not persist after adjusting for music lessons and self-reported musical skill. For 43/60 languages, the translated credible interval contained zero, accounting for any effect that might otherwise have been attributed to native language. Although the residual α values deviated substantially from zero for several languages, the overall trend suggests that the variation across languages observed in Figure 1.4 was not due to native language itself.

It is worth noting that the data on self-reported musical skill were collected after all trials of the experiment, during the second set of survey questions as described in Section 1.2.4. This opens up the possibility that listeners adjusted their self-reported responses based on their experience over the course of the game. Since the proportion of highly musically skilled listeners in a language was calculated as the proportion of listeners who responded "I'm an expert" or "I have a lot of skill" when asked to make a rating of their musical skill, the influence of doing the experiment on a listener's predisposed responses would only change the results if it changed a listener's actual response from one of "I'm an expert", "I have a lot of skill" to one of "I have some skill", "I'm a novice", and "I have no skill at all" or vice versa. Presumably, such influences apply mostly to listeners whose self-assessment is somewhere between "I have some skill" to "I have a lot of skill" (52% of listeners chose one of these two categories). Such an effect does not apply to data on music lessons which are not based on subjective rating and which also strongly predict α.

1.4.3 Single-resource model yields similar facilitation strengths across conditions regardless of language

The notion that performance in tone-scramble tasks is not influenced by native language is further supported by the analysis based on the single-resource model. In prior tone-scramble experiments, the single-resource model has provided an accurate description of the data, explaining between 74% and 84% of the variation in d' values (Dean and Chubb, 2017; Mednicoff et al., 2018; Ho et al., 2022). Importantly, the facilitation strengths F_t of the single-resource model represent the relative ability that the latent resource affords in each condition t.

When partitioning the present sample on the basis of native language, we found that F looked very similar across the language groups. For each language group, tasks involving

stimulus types that clearly differed in terms of majorness/minorness tended to be easier than the average task. These included the 3v**4**-task (wherein stimuli contain either the major third or the minor third) and the 8v**9**-task (wherein stimuli contained either the major sixth or minor sixth). Notably, the 10v**11**-task was harder than the average task across languages, consistent with results obtained by Dean and Chubb (2017). The lack of an effect of native language on the estimates of the single-resource model is unsurprising in light of the lack of relationship between α and native language, as reported above.

1.4.4 For listeners with low performance on the 3v4-task, same/different-task offers no advantage

We found that low-performing listeners in the 3v**4**-task did not perform any better on the same/different-task. Thus, we conclude that these listeners in general do not reliably perceive a difference between the two types of stimuli in the 3v**4**-task. It is not the case that low-performing listeners hear a difference but fail to apply the correct labels to individual stimuli. Notably, the mean difference among the low-performing listeners of the 3v**4**-task (a group of 45,299 listeners, using the criteria specified in Section 1.3.4) comparing the proportion-correct in the same/different-task to the 3v**4**-task was not practically different than zero, suggesting that the two tasks were of equal difficulty. This is consistent with the results reported by Chubb et al. (2013) and discussed in Section 1.1.3. Chubb et al. found that listeners, on average, tended to respond on each trial by mimicking the correct response of the previous trial. For low-performing listeners, the finding of Chubb et al. suggests that they compare the stimulus on each trial to the one that came before it (not unlike the comparison required in the same/different-task) and, failing to hear a difference, respond with the same answer as the correct answer to the previous trial.

Statistically speaking, the mean difference was found to be negative, indicating that the

same/different-task was slightly more difficult. Such an effect might be attributed to some limitation(s) of listeners' memory when required to compare two auditory stimuli; however, this effect was too small to warrant any further investigation.

1.4.5 Future work

It is remarkable that tone-scramble tasks reveal such a profound disparity in ability among listeners of a very broad listenership. The source of this disparity remains to be discovered. Although the present work leads us to conclude that native language is not directly related to tone-scramble performance, there are many other covariates to be considered. Indeed, the present data leave us with a large number of covariates available for each listener yet to be analyzed. These include data on self-reported pitch discrimination, self-reported rhythmic fluency, degree of early-age musical exposure, susceptibility to musical frisson, self-reported fidelity of visual and auditory mental imagery, kinds of non-musical expertise, and concurrent health conditions. Might any of these covariates be strongly linked to listeners' performance in tone-scramble tasks? It remains to be investigated. Data mining may be used to identify individual covariates or combinations thereof that reliably predict whether a listener falls into the low- or high-performing group. Of course, any such exploratory analysis will need to take care to use some form of validation to assess out-of-sample generalization. In general, data sets as large as the present one are especially amenable to such validation techniques. Any results from such an exploratory analysis would be helpful in forming hypotheses about the origin of tone-scramble performance which could in turn be studied using both lab-based and web-based testing.

The citizen-science approach used here offers many unique opportunities previously unavailable in the psychological sciences, especially with regard to the generalizability of research findings (Hilton and Mehr, 2021). The large sample sizes that are available using the citizen-

science model are not only useful for the purposes of studying a more diverse (and, ideally, more globally representative) sample of research participants; they are also useful for simultaneously studying a wide variety of conditions that would be too numerous to practically test in a single sitting (as might be done in a more common lab-based experimental paradigm). In this study, we were able to test eight varieties of tone-scramble task other than the basic 3v**4**-task, some of which replicate prior variants (Dean and Chubb, 2017; Ho et al., 2022) and some of which were novel. Given the amount of data expected, we were able to test all of the tone-scramble tasks belonging to a single category (i.e., the category of tone-scramble tasks in which the target notes of the two stimulus types differed by one semitone) without having to make decisions about which variants were most likely to produce interesting results. Indeed, the tone-scrambles studied here are still a smaller category than all of those available, and studying a broad set of tone-scramble tasks will help us gain a better understanding of the origin of tone-scramble performance and the nature of the theorized latent resource R that underlies it. By the same token, the citizen-science approach might allow us to study how tone-scramble performance relates to other hearing abilities by having listeners perform some selection of non-tone-scramble tasks chosen from a large set. The roved pitch-difference task studied by Ho et al. (2022) is one promising candidate for such a study as Ho et al. found that a roved pitch-difference threshold of a quarter-tone (50 cents) was required to perform the tone-scramble task. Other candidates include preference tasks or discrimination tasks where the relevant stimulus space is too large to be studied within-subjects (e.g., using recordings or melodic compositions selected from a musical corpus).

The main limitation to the citizen-science approach is participant motivation. In the present study, we tried to elicit excitement from listeners by formatting the experiment as a video game and by offering listeners an informed rating of their hearing ability (advertised here as "super-listener" status). Since no compensation was offered for this study, these factors likely influenced the amount of participation we received. The degree of participation in any citizen-science study is likely to depend on a complex interaction of many sociological

factors, so any of the possible follow-ups outlined above will also need to be informed by their relevance to the broad audience of potential participants.

Chapter 2

Piano timbre, lower frequency, and reduced presentation rate do not improve performance in tone-scramble tasks.

2.1 Introduction

The most outstanding feature of music is, perhaps, its emotional resonance. Music has a remarkable power to arouse the feelings of those who listen to it, as is psychologically and physiologically evident (Eerola and Vuoskoski, 2013; Juslin, 2013; Koelsch, 2014). What features of music imbue it with such emotional resonance?

According to theories of composition, scale has a central role in giving meaning to music (Rameau, 1722; Schoenberg, 1922; Tymoczko, 2011). Indeed, many studies find that, on average, music of the major scale sounds happy to listeners, and music of the minor scale

sounds sad (e.g., Hevner, 1935; Crowder, 1984; Crowder, 1985a; Kastner and Crowder, 1990; Gerardi and Gerken, 1995; Gagnon and Peretz, 2003; Temperley and Tan, 2013; Bonetti and Costa, 2019). However, we would be remiss to ascribe the emotional character of music to scale alone. Music is a complexly structured auditory stimulus, and its affective information is unlikely to be conveyed entirely by scale. In a single piece of music, rhythm, timbre, and harmony may interact in eliciting an emotional response. To understand the role of scale in this elicitation, we must acknowledge that the affective quality of scale may be mediated by other features such as instrumentation, tempo, or phrasing.

In fact, there is mounting evidence that the majority of listeners are less sensitive to scale *per se* than might be predicted given the central role of scale in music theory. Studies have found that listeners of all degrees of musical training struggle to discriminate melodies differing only in scale (Halpern, 1984; Halpern et al., 1998; Leaver and Halpern, 2004). Moreover, the finding that major (minor) stimuli are judged to be happy (sad) *on average* does not imply that this pattern holds for all listeners. Studies by Blechner (1977) and Crowder (1985a) suggest that sensitivity to scale in major vs minor triadic chords may be bimodally distributed, with some listeners exhibiting high sensitivity and others exhibiting little sensitivity, if any. Indeed, in studies that have carefully manipulated scale while fixing other features of music, the mean effect is found to be modest and generally consistent with the idea that sensitivity to scale follows a bimodal distribution. Statistically significant effects in such studies may be explained by a small number of strongly sensitive listeners in a sample in which most listeners have little or no sensitivity.

Experiments using randomly ordered tone sequences ("tone-scrambles") support the theory that sensitivity to scale follows a bimodal distribution, with most listeners lacking sensitivity to scale (Chubb et al., 2013; Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020; Adler et al., 2020; Ho et al., 2022). In the typical tone-scramble task (the "3-task"), the tone-scrambles contain thirty-two 65-ms pure-tones, including eight each of the notes G_5,

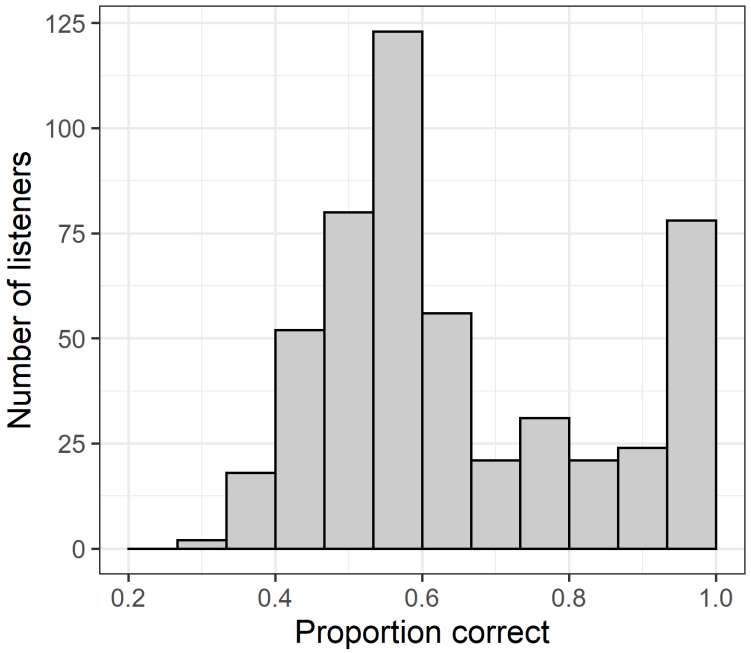

Figure 2.1: Histogram of proportion correct in the 3-task pooled over Chubb et al. (2013), Dean and Chubb (2017), Mednicoff et al. (2018), Ho et al. (2022). In each of these studies, proportion correct is based on 50 trials which were preceded by at least 40 practice trials.

D_6, and G_6; in addition, major (minor) tone-scrambles contain eight pure-tones of the note B_5 ($B\flat_5$). Tone-scrambles are presented one at a time, and the listener strives to classify the tone-scramble presented on each trial as major or minor with trial-by-trial feedback. As seen in Fig. 2.1, proportion correct in the 3-task follows a bimodal distribution. The majority of listeners ($\approx 70\%$) hear little difference between the major and minor tone-scrambles and form a mode near 55% correct. The rest of listeners are highly sensitive to this difference and amass near 100% correct.

2.1.1 The current study

Does the bimodal distribution of sensitivity to scale in major vs minor tone-scrambles generalize to actual music? Given that the findings based on tone-scrambles are consistent with previous research on major and minor music, we believe that the answer is yes. However, as

previously noted, rhythm, timbre, and harmony may interact within a single piece of music, and the affective quality of scale may be mediated by other features such as instrumentation, tempo, or phrasing. Perhaps all listeners have sensitivity to scale in the context of an adequately musical stimulus because such a stimulus has other features that make variations in scale salient. It is important to acknowledge that tone-scrambles differ from actual music in several important ways: they are very rapid (roughly equivalent to a sequence of 32^{nd} notes at 120 BPM), randomly sequenced, relatively high in frequency (spanning G_5 to G_6), and composed of isolated pure-tones. These properties may obscure scale or may fail to provide some structure that most listeners use to hear scale, and it may be that high-performing listeners are specially capable of overcoming these obstacles.

A more musical stimulus might be slower, purposefully sequenced, lower in frequency, and composed of more natural sounds. Already, others have found that slowing the presentation rate (Mednicoff et al., 2018) and adding sequential structure (Ho and Chubb, 2020) do not alone elevate low-performers and close the performance gap visible in Fig. 2.1. Nonetheless, tone-scrambles under these manipulations may still lack musicality because they remain high in frequency with an unnatural timbre. The current study manipulated three factors – presentation rate, timbre, and frequency height – to create tone-scrambles that were more firmly musical. Relative to the tone-scrambles of the 3-task (referred to here as the "Fast, Pure, G_5" condition), we slowed the presentation rate of notes to roughly that of 16^{th} notes at 90 BPM, we replaced pure-tones with piano recordings, and we shifted the stimulus down 19 semitones in log frequency to span C_4 ("middle C" in common music notation) to C_5. These manipulations were tried in six different combinations as described in Sec. 2.2.1 to achieve various levels of musicality.

Do any of these manipulations increase the salience of scale for low-performing listeners of the "Fast, Pure, G_5" condition? One may approach this question in a number of ways. We tested whether listeners achieved significantly higher mean performance in any condition compared

to "Fast, Pure, G_5", as reported in Sec. 2.3.1. However, as might be the case with previous studies of major and minor stimuli cited earlier, effects on mean performance might be driven mostly by a high-performing minority of listeners. To this end, it may be more informative to examine whether the group of high-performing listeners changes from one condition to the next. Thus, as reported in Sec. 2.3.2, we considered whether any condition relative to "Fast, Pure, G_5" produced a larger proportion of high-performing listeners. Finally, Dean and Chubb (2017) found that performance across several tone-scramble tasks was well-described by a single-resource model. According to this model, a single cognitive resource R affords ability in all tone-scramble tasks, and low- and high-performing listeners differ only in the amount of R available to them. If the manipulation of musicality in the current study makes scale salient to listeners who perform poorly in the "Fast, Pure, G_5" condition, then the data must depart from the single-resource model, which predicts that a listener who performs poorly in one condition will perform poorly in all conditions. We assess the fit of this model in Sec. 2.3.3.

The vast majority of studies on the emotional quality of major and minor music use stimuli that are moderate in tempo and frequency and are sounded with natural timbres like the tone-scrambles in the current study. As noted earlier, the results of those studies are consistent with a bimodal distribution of scale sensitivity. There is little experimental evidence to suggest that our manipulations will provide a special advantage to low-performing listeners. Indeed, as we report below, our manipulations overall did little to unlock scale for low-performing listeners in the "Fast, Pure, G_5" condition. This finding suggests that low-performing listeners in tone-scramble tasks are not prevented from hearing scale by the unmusical qualities of tone-scrambles, and it lends further support to the theory that the bimodal distribution of scale sensitivity generalizes to a broader class of musical stimuli.

2.2 Methods

All methods were approved by the UCI Institutional Review Board.

Participants

Thirty-three listeners participated. Listeners were recruited through the UCI School of Social Sciences Subject Pool and were compensated with course credit. The data were collected over the internet between June 26, 2022 and October 5, 2022.

All listeners had self-reported normal hearing. The mean years of musical training was 4.06 (sample standard deviation: 4.56). Twenty-one of these listeners reported having at least one year of musical training.

All listeners were prompted to complete the experiment in a quiet place with no visual distractions while wearing headphones. Each listener completed a screening test (Woods et al., 2017) at least once at the beginning of the experiment to assess headphone usage. Those who failed the initial screening test were prompted again to wear headphones and were required to perform the screening test once more. Thirty of the 33 total listeners passed the screening test on either their first (27) or second (3) attempt.

2.2.1 Stimuli

We presented listeners with six varieties of stimuli. The six varieties differed in presentation rate, timbre, and/or frequency range, and we refer to them as "Fast, Pure, G_5", "Slow, Pure, G_5", "Slow, Piano, G_5", "Fast, Pure, C_4", "Slow, Pure, C_4", and "Slow, Piano, C_4". Each stimulus variety comprised two stimulus types (major and minor).

In the "Fast, Pure, G_5" condition, each stimulus comprised 32 pure-tones which were 65 ms in duration (for a total stimulus duration of 2.08 s) and windowed by a raised cosine function with 22.5 ms rise and decay times. Thus, the presentation was roughly equal to a sequence of 32^{nd} notes at 120 BPM. All tones had equal peak amplitude. The major (minor) stimulus type comprised eight pure-tones per the following notes: G_5, B_5 ($B\flat_5$), D_6, and G_6. The notes in a given tone-scramble were presented in random sequence.

In the "Slow, Pure, G_5" condition, the tones were relatively longer and fewer: each stimulus comprised 12 pure-tones which were $173.\bar{3}$ ms in duration (as to maintain a total stimulus duration of 2.08 s). Thus, the presentation was roughly equal to a sequence of 16^{th} notes at 90 BPM. These were, again, windowed by a raised cosine function with 22.5 ms rise and decay times. The major (minor) stimulus type comprised three pure-tones per the notes G_5, B_5 ($B\flat_5$), D_6, and G_6, presented in random sequence.

In the "Slow, Piano, G_5" condition, piano recordings were used in place of pure-tones. These recordings were of isolated notes and were taken from the Maestro Concert Grand sound bank by Mats Helgesson: `http://sonimusicae.free.fr/matshelgesson-maestro-en.html`. We used only recordings labelled as mezzo-piano. The relative level of these recordings was not altered, and any difference between stereo channels was retained. The end of each recording was cropped off so that each recording had a duration of $173.\bar{3}$ ms, and a 22.5 ms cosine damp was applied to each (to retain the attack of the original recordings, no ramp was applied). Stimuli of this condition were constructed from 12 such recordings played consecutively (maintaining a total stimulus duration of 2.08 s) and in random order, three per the notes G_5, B_5 ($B\flat_5$), D_6, and G_6 for the major (minor) stimulus type.

The stimuli of the "Fast, Pure, C_4", "Slow, Pure, C_4", and "Slow, Piano, G_5" were like the stimuli of the corresponding conditions above except that the notes of the major (minor) stimulus type were C_4, E_4 ($E\flat_4$), G_4, and C_5.

The stimuli were generated at trial time and played back using the Web Audio API (`https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API`) which is supported by most contemporary internet browsers. The audio sampling rate depended on listeners' personal hardware. For all listeners, the sampling rate was either 44.1 kHz or 48 kHz; potential listeners were not allowed to proceed if any other sampling rate was detected. Volume was adjusted manually by each listener to a comfortable level prior to the experiment. Listeners read prompts and entered responses via a web interface built with jsPsych (De Leeuw, 2015).

2.2.2 Procedure

Each listener was tested in all six conditions described in Sec. 2.2.1. Conditions were blocked and ordered using reverse counterbalancing. For a given participant, the order of the first six blocks (one per condition) was determined by a Latin square. The last six blocks followed the reverse order of the first six. On each trial of a block, the listener was presented with a single stimulus and strove to classify it as either major or minor. A single block consisted of a brief sequence of example stimuli followed by 40 trials (20 major and 20 minor); the stimuli in a block were ordered randomly with a constraint that the first 10 and last 30 trials each had an equal number of major and minor stimuli. Only the last 30 trials of each block were analyzed, yielding 60 trials per condition for analysis since each condition was tested twice.

Listeners completed a series of intake procedures via the experiment web page prior to participating in the experiment proper. On the introductory page, listeners were prompted to use either the Google Chrome or Mozilla Firefox web browsers (which were tested to be compatible with the Web Audio API used in the experiment). On this page, listeners also received a link to access the study information sheet as a PDF file. To proceed to the next page, listeners were required to check boxes to affirm that (1) they were using

headphones and located in a quiet space free of distractions and that (2) they agreed to take part in the study. On the following two pages, listeners answered survey questions about their native language and years of musical training, respectively. Listeners were then provided an interface to play back Gaussian noise; listeners were prompted to use this noise as a reference with which to adjust the computer volume to a loud but comfortable level. Upon completing the volume adjustment, listeners were given a headphone screening test as developed by Woods et al. (2017). This screening test consisted of six trials of a three-interval task on which a half-amplitude tone was to be identified among two whole-amplitude tones (one with stereo channels playing in-phase, one with stereo channels playing anti-phase). Listeners who failed to answer at least five of the six trials correctly were prompted again to wear headphones and were required to perform the screening test once more. Listeners who failed the screening test a second time were allowed to proceed with the experiment (3 of 33 listeners failed the screening test twice).

Upon completing the intake procedures, listeners performed a training battery consisting of 6 blocks of 20 trials each (the data collected over the course of this training battery were not used in the analysis) punctuated by three instructional videos. The conditions demonstrated by each block in the training battery followed the same fixed order for all listeners, as described below. The first instructional video described the difference between major and minor chords in terms of their constituent notes and their emotional connotation according to music theory. This video then described the stimuli of the present experiment, their relation to the major and minor chords, and the task of the listener. Each stimulus given as an example in this video belonged to the "Slow, Piano, C_4" condition and was accompanied by a visualization of its note sequence on a piano keyboard. Following this first instructional video, listeners completed one training block of the "Slow, Piano, C_4" condition followed by one training block of the "Slow, Piano, G_5" condition. Each of these training blocks was preceded by (1) a prompt describing the qualities of the stimuli and restating the task of the listener and (2) a short sequence of examples. The second instructional video

appeared after these first two training blocks. The second instructional video introduced pure-tones in contrast to the piano timbre used thus far and provided examples of major and minor chords and the experimental stimuli in pure-tones. The second instructional video was followed by four training blocks, one for each of the (in order) "Slow, Pure, C_4", "Slow, Pure, G_5", "Fast, Pure, C_4", and "Fast, Pure, G_5" conditions, each preceded by a corresponding prompt and example sequence. The final instructional video appeared at the end of the training battery and informed listeners of the number of blocks and trials that they would be tested on.

At the outset of the K^{th} block, an announcement appeared on-screen: "*Entering Block [K] of 12. Remember, your task is to identify whether each stimulus is* minor *(Type 1) or* major *(Type 2). Press space bar to hear examples.*." Upon pressing the space bar, the listener heard two examples each of the minor and major stimuli in alternating order with correct labels provided. The examples were self-paced. At the end of the example sequence, listeners saw the following prompt:"*Ready? Press Y to begin testing. Or, press N to repeat the examples.*" If the listener pressed N on the keyboard, the above sequence was repeated. If the listener pressed Y, the testing sequence was initiated.

Stimuli were presented one at a time. The listener responded to each trial by pressing "1" for minor and "2" for major. At the onset of each trial, a prompt reminded the listener of the response mapping and the current trial number, X ("*Trial [X] of 40. Press 1 for* minor *(Type 1). Press 2 for* major *(Type 2).*"). Feedback ("*CORRECT*" or "*INCORRECT*") was presented after each response. A progress bar appeared at the top of the screen at all times and was incremented after every trial.

2.3 Results

2.3.1 Do any of the conditions provide a reliable improvement in performance?

If presentation rate, timbre, and/or frequency height modulate the salience of scale in tone-scramble stimuli, then some combination of our manipulations should afford listeners greater traction in the task of labeling tone-scrambles as major and minor. To this end, we tested whether performance in each condition was, on average, greater than that of the "Fast, Pure, G_5" condition.

The dependent variable in our analysis was sensitivity (d' of signal detection theory). We used only the last 30 trials of each block of each task to calculate d', yielding 60 trials per condition. To calculate d' under our experimental paradigm, we needed to label trials as "signal" and "noise" trials arbitrarily. We chose to treat major (minor) trials as signal (noise) trials. This choice did not influence our d' measures. Each of the 33 listeners provided six data points: one d' measure for each of the six conditions, for a total of 198 observations. If a listener achieved a perfect hit rate (i.e., 30 hits out of 30 signal trials) in a given condition, the proportion of hits was set to $\frac{30-0.5}{30} = 0.98\bar{3}$ (as recommended by Macmillan and Kaplan, 1985). Analogous adjustments were made for correct rejection rates. Figure 2.2 shows the histograms of the resulting d''s in each condition based on these calculations.

A series of one-sided paired t-tests found that no condition had an appreciable advantage over the "Fast, Pure, G_5" condition. The p-values of the tests assessing the mean difference in d' between the "Fast, Pure, G_5" and the "Slow, Piano, G_5" ($t_{32} = 1.129$, $p = 0.1337$), "Slow, Pure, G_5" ($t_{32} = -1.409$, $p = 0.9158$), "Slow, Piano, C_4" ($t_{32} = 0.916$, $p = 0.1833$), "Slow, Pure, C_4" ($t_{32} = -0.133$, $p = 0.5525$), and "Fast, Pure, C_4" ($t_{32} = 0.069$, $p = 0.4728$) conditions were all large.

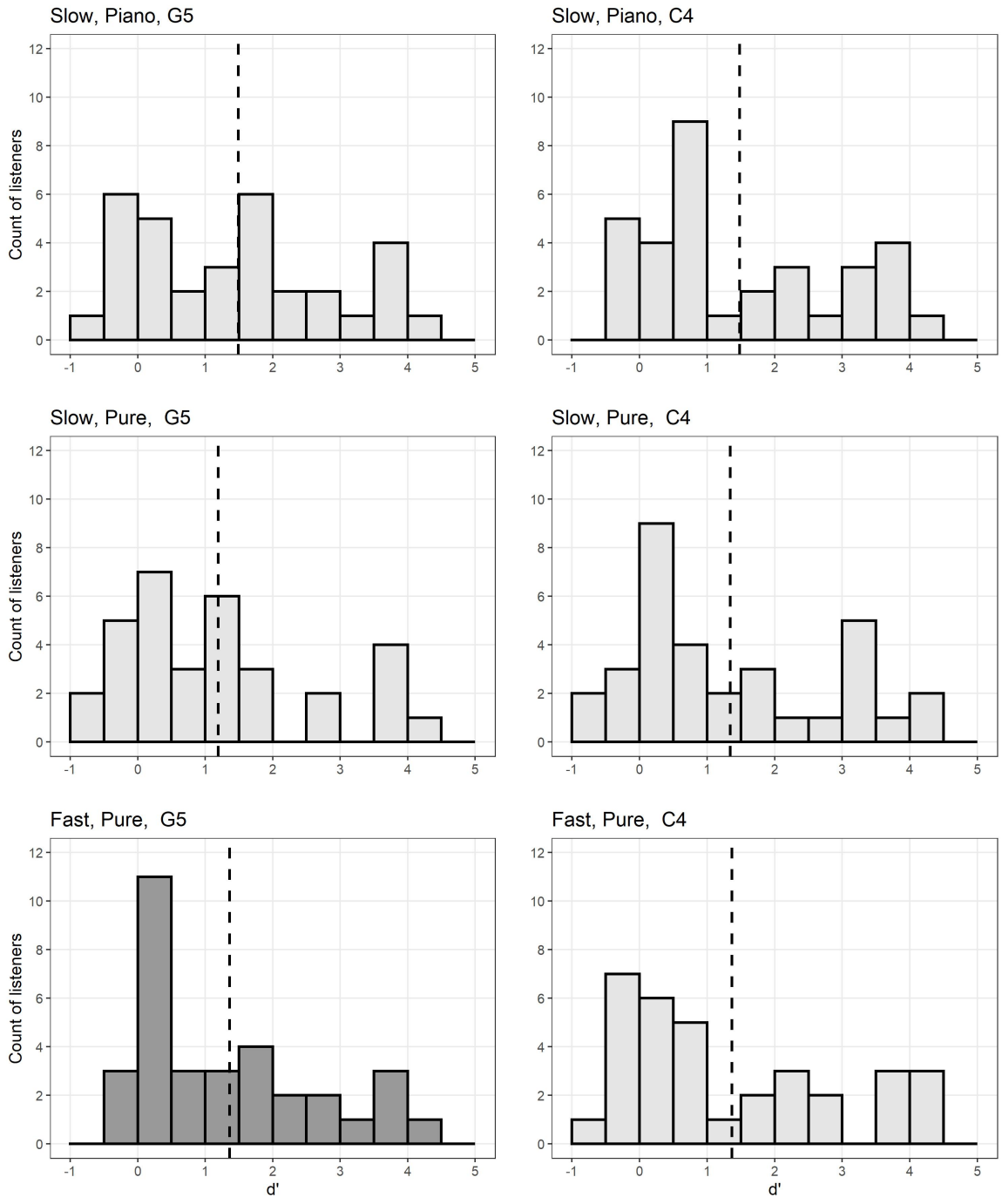

Figure 2.2: Histogram of d' for each condition. The vertical dashed lines each represent the mean d' in a given condition, averaging over all listeners. The "Fast, Pure, G_5" condition is emphasized using darker shading.

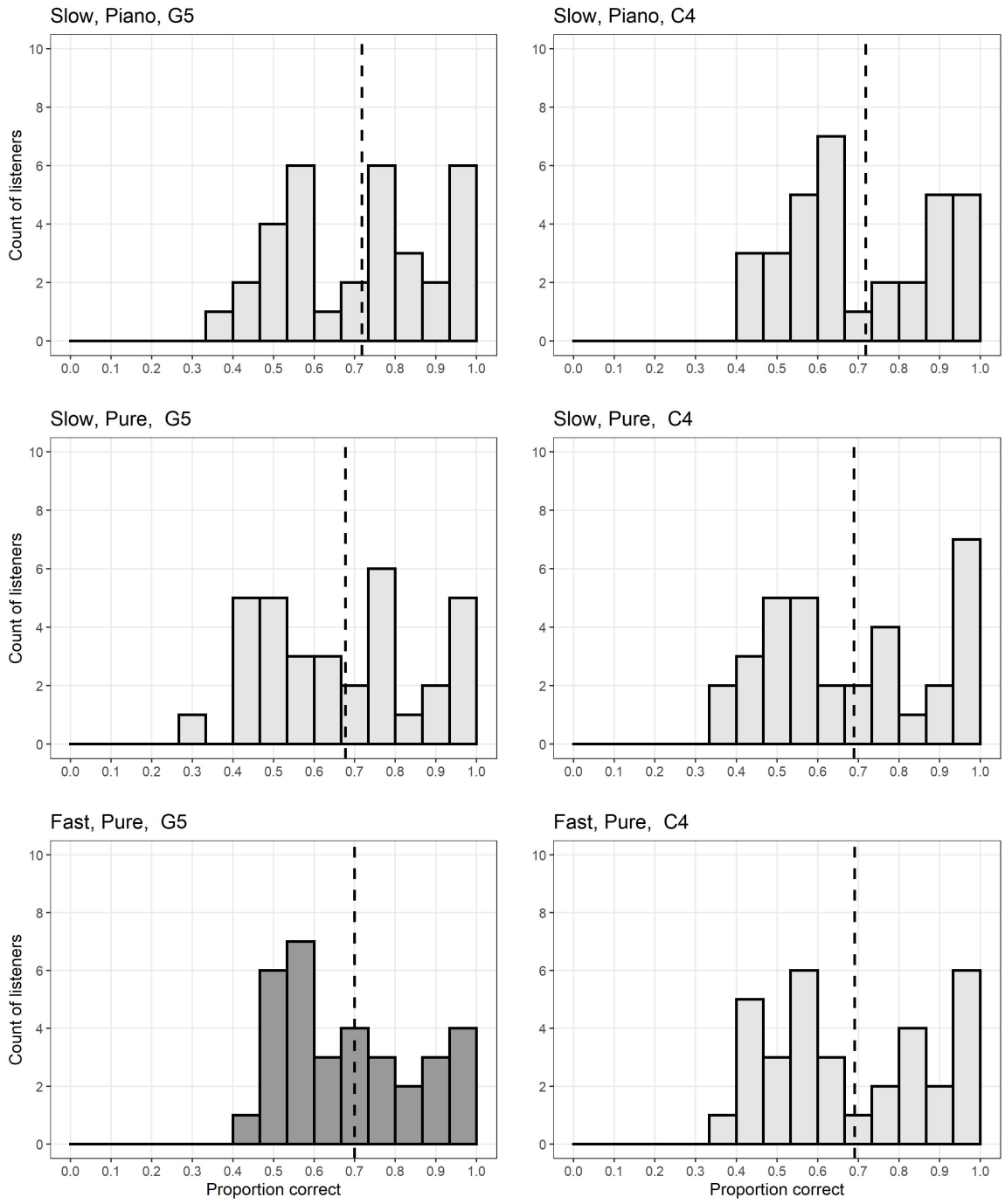

Figure 2.3: Histogram of proportion correct out of 60 trials for each condition. The vertical dashed lines each represent the mean proportion correct in a given condition, averaging over all listeners. The "Fast, Pure, G_5" condition is emphasized using darker shading.

2.3.2 Do any of the conditions meaningfully change the relative proportion of low-performers to high-performers?

Figure 2.3 shows the histogram of proportion correct for each condition. As with our calculations of d', we used only the last 30 trials of each block of each condition to calculate proportion correct, treating the first 10 trials of each block as practice. This yielded 60 trials per condition. As observed in previous studies with tone-scrambles, the distribution for the "Fast, Pure, G_5" condition is approximately bimodal. The distributions for the other conditions each also appear to be bimodal, with a prominent intermediate mode in some cases.

It is conceivable that the effect of our manipulations on the salience of scale across conditions manifests as changes in the relative mass attributed to the upper and lower modes of these distributions, indicating that the number of listeners belonging to the high-performing group changes from one condition to the next. To examine this possibility, we fit a binomial-mixture model to the counts Y of correct responses. This model has the form

$$P(Y_{s,t} = y) = \alpha_t \binom{n}{y} (1 - p_t)^{n-y} p_t^y + (1 - \alpha_t) \binom{n}{y} (1 - q_t)^{n-y} q_t^y \tag{2.1}$$

where $s = 1, \cdots, 33$ indexes the listener, $t = 1, \cdots, 6$ indexes the condition, and $n = 60$ is the maximum number of correct responses produced by a given listener in a given condition. According to this model, each count of correct responses in a given condition t is drawn from one of two binomial distributions with success probabilities p_t and q_t, respectively, with probability α_t that the count is drawn from the former distribution (note that all parameters of this model are fixed across all listeners, so individual effects are not captured by this model). Without constraint, these parameters are not identifiable (an equivalent

model can be produced by setting α_t equal to $1 - \alpha_t$ and swapping p_t and q_t). Thus, we imposed the constraints

$$0 < p_t \leq \frac{7}{10} \quad \text{for all} \quad t \tag{2.2}$$

and

$$\frac{7}{10} \leq q_t < 1 \quad \text{for all} \quad t. \tag{2.3}$$

Under these constraints, we may interpret p_t (q_t) as the probability that a low-performing (high-performing) listener will produce a correct response on a single trial of condition t, and we may interpret α_t as the proportion of low-performing listeners out of all listeners in condition t.

The model has a total of 18 free parameters. We estimated these parameters with Bayesian methods, assuming the following priors: $\alpha_t \sim$ Uniform(0, 1) i.i.d, $p_t \sim$ Uniform(0, 0.7) i.i.d., and $q_t \sim$ Uniform(0.7, 1) i.i.d. Point estimates were calculated by taking the median of 1,000,000 MCMC samples, thinned to every 100^{th} sample. These were drawn after first taking 1,000,000 burn-in samples.

Figure 2.4 plots the posterior estimates of α_t for all conditions t. The median posterior estimates for the proportion of listeners belonging to the low-performing group for each condition were 0.472 with 95% credible interval [0.311, 0.648] (Slow, Piano, G_5), 0.601 with 95% credible interval [0.401, 0.855] (Slow, Pure, G_5), 0.655 with 95% credible interval [0.475, 0.813] (Fast, Pure, G_5), 0.624 with 95% credible interval [0.451, 0.775] (Slow, Piano, C_4), 0.687 with 95% credible interval [0.513, 0.829] (Slow, Pure, C_4), and 0.615 with 95% credible interval [0.442, 0.770] (Fast, Pure, C_4). Although the median posterior estimate of α for the "Slow, Piano, G_5" condition is noticeably lower than the α of the other conditions in the Fig. 2.4, the credible intervals for these estimates suggest that this difference is not statistically

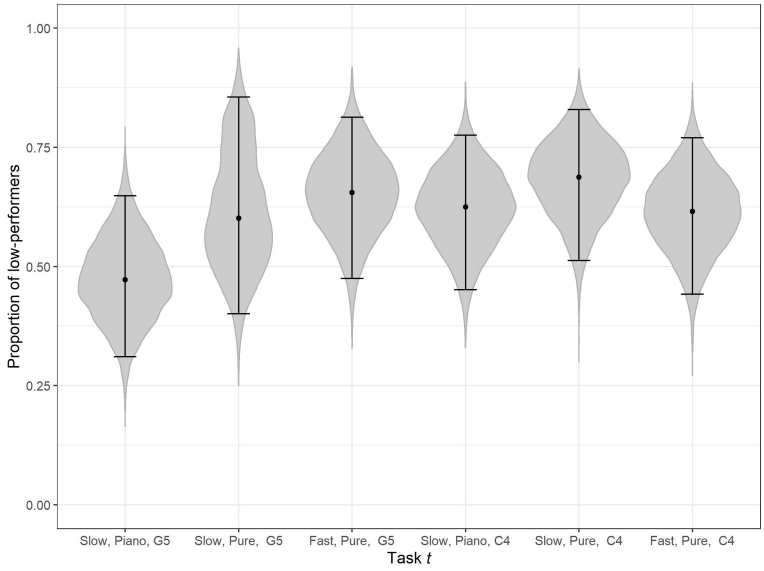

Figure 2.4: Violin plot of the posterior samples of α_t. Each violin represents an estimate of the proportion of listeners belonging to the low-performing group of a given condition condition. Bars indicate 95% Bayesian credible intervals, and points represent median posterior estimates.

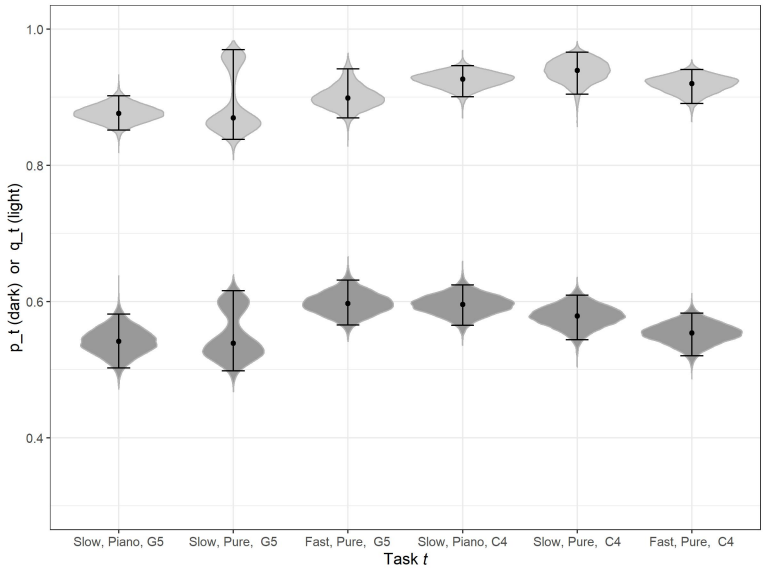

Figure 2.5: Violin plot of the posterior samples of the success probabilities p_t (dark grey) and q_t (light grey). Each dark grey (light grey) violin represents an estimate of the probability that a listener belonging to the low-performing (high-performing) group of a given condition will provide a correct response on a single trial. Bars indicate 95% Bayesian credible intervals, and points represent median posterior estimates.

significant: the median posterior difference in α comparing the "Slow, Piano, G_5" condition to the "Fast, Pure, G_5" condition was -0.181 with 95% credible interval [-0.415, 0.067]. The estimated success probabilities for the low-performing and high-performing groups in each condition are shown in Fig. 2.5.

2.3.3 How well are the data predicted by a single-resource model?

Previous studies (Dean and Chubb, 2017) using tone-scramble stimuli have found that d' data are well-described by a single-resource model having the following form:

$$d' = R_s F_t + \varepsilon_{s,t} \tag{2.4}$$

subject to the constraint

$$\sum_{t=1}^{6} F_t = 6, \tag{2.5}$$

where $s = 1, \cdots, 33$ indexes the listener, $t = 1, \cdots, 6$ indexes the condition, and $\varepsilon_{s,t} \sim N(0, \sigma^2)$ i.i.d. According to this model, a single latent cognitive resource R governs performance in all conditions. Different listeners possess different levels of R (hence R_s), which facilitates performance in condition t with relative strength F_t. Note that F_t is fixed across listeners, so aside from measurement error, variation between listeners is attributed only to the amount of variation in the amount of R possessed by listeners.

We fit this model using Bayesian methods to the d' values calculated in Sec. 2.3.1, assuming the following priors: $R_s \sim$ Uniform(-100, 100) i.i.d., $F_t \sim$ Uniform(-100, 100) i.i.d., and $\sigma^2 \sim$ Uniform(0, 100). This model has a total of 40 free parameters. Given the larger number of parameters in the single-resource model compared to the binomial-mixture model used in Sec. 2.3.2, we used a larger thinning interval (and a proportionately larger number of base

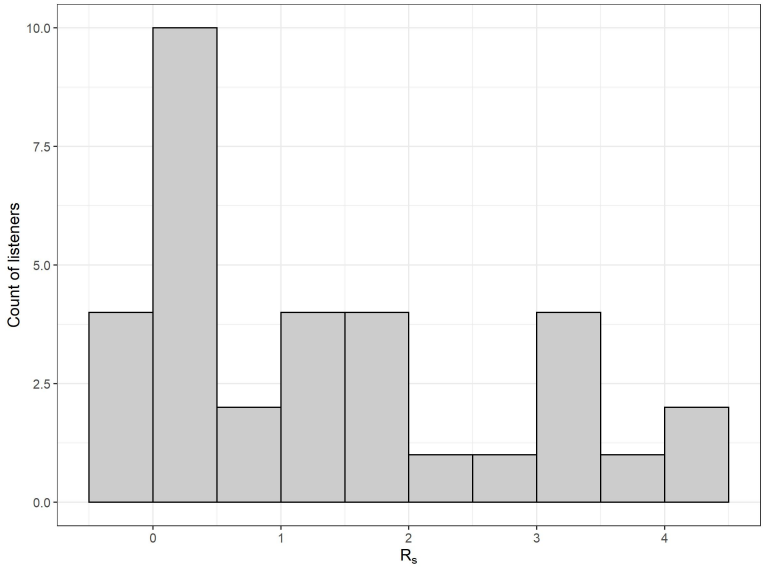

Figure 2.6: Histogram of R_s for the 33 listeners studies. An R of 1 corresponds to a mean d' of 1 (which corresponds to a proportion correct around 0.69).

samples) to draw posterior samples: point estimates were calculated by taking the median of 10,000,000 MCMC samples, thinned to every 1000^{th} sample which were drawn after first taking 10,000,000 burn-in samples. We found that these hyperparameters were sufficient to mitigate issues of autocorrelation among posterior samples.

The histogram of median posterior R_s estimates is provided in Fig. 2.6, and a violin plot of the posterior samples for F_t is provided in Fig. 2.7. As in previous studies, we see that the distribution of R_s has a "spike-and-slab" shape with a prominent mode near 0 and a roughly uniform spread over the positive range. The data were fit exceptionally well by the single-resource model: the model explained 90.3% of the variation observed in d''s (i.e., $r^2 = 0.903$). Figure 2.8 plots the observed d''s vs. corresponding predicted d''s.

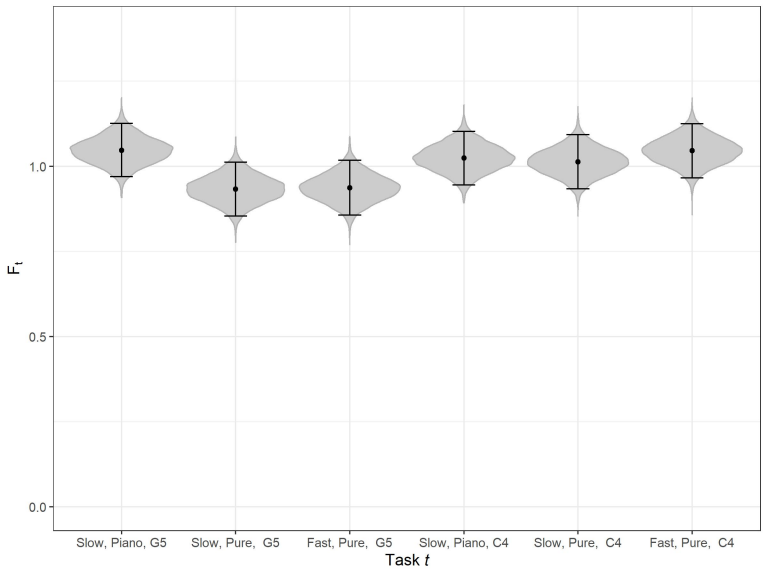

Figure 2.7: Violin plot of the estimated marginal posterior densities of F_t. Points indicate posterior medians, and error bars indicate 95% Bayesian credible intervals.

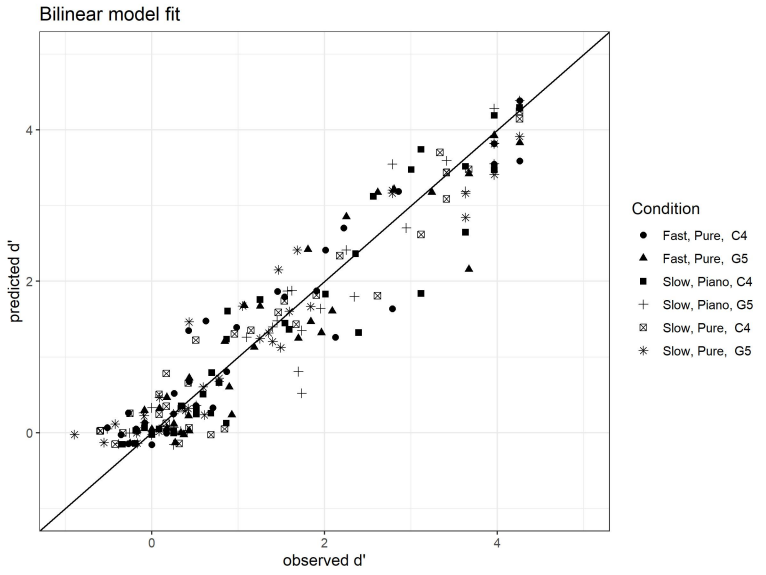

Figure 2.8: Observed d''s plotted against d''s predicted by the single-resource model. The dashed $x = y$ line represents a perfect fit of the data.

2.4 Discussion

2.4.1 Presentation rate, timbre, and frequency height do not increase the salience of scale for low-performing listeners.

It is clear from the results of the current study that overall, our manipulations of presentation rate, timbre, and frequency height do not elevate performance among listeners who perform poorly in the "Fast, Pure, G_5" condition. As evidenced by the paired t-tests in Sec. 2.3.1, all other conditions failed to produce an improvement in performance on average. The parameter estimates of the binomial-mixture model reported in Sec. 2.3.2 lead us to a similar conclusion.

Ultimately, we find that the single-resource model originally used by Dean and Chubb (2017) provides an excellent description of the data. This model supposes that a single resource is used by all listeners and that the relative performance F_t afforded by this resource across conditions t (shown in Fig. 2.7) is fixed across all listeners; variation in performance between listeners is due only to variation in R_s, the amount of that resource available to each listener s. According to this model, a low-performing listener in one condition must be a low-performing listener in all conditions, so the goodness-of-fit of this model to the data rules out any possibility that one or more of our manipulations unlock sensitivity to scale for all listeners.

The current study establishes a stronger connection between research on tone-scrambles and the greater body of research regarding the emotional quality of musical scale (Hevner, 1935; Crowder, 1984; Halpern, 1984; Crowder, 1985a; Kastner and Crowder, 1990; Gerardi and Gerken, 1995; Halpern et al., 1998; Gagnon and Peretz, 2003; Leaver and Halpern, 2004; Temperley and Tan, 2013; Bonetti and Costa, 2019) which has generally used stimuli that are slower, lower in frequency, and more naturally timbred than the tone-scrambles previously

studied. As stated in Sec. 2.1, we believe that the main findings of this literature are consistent with a bimodal distribution in sensitivity to scale like that observed in previous tone-scramble experiments and replicated in the current study. The results of the current study suggest that this distribution in sensitivity generalizes to a larger class of musical stimuli than might be suggested by previous tone-scramble experiments, a class that includes many of the kinds of stimuli previously used to study musical scale.

Nonetheless, we acknowledge that even the most musical stimuli of the present study (i.e., the "Slow, Piano, C_4" tone-scrambles) differ from actual music in a number of ways. For example, the tone-scrambles were played without harmonic or percussive accompaniment. Moreover, the notes in each tone-scramble were ordered at random and all had the same duration. Certainly, such features, when present, influence the emotional character of a piece of music and may covary with pitch of the prevailing note in ways that make scale more salient. Other variations in music occur over a much longer duration than two seconds (the duration of a tone-scramble), such as the variations used to achieve the verse-chorus form. These dimensions of musicality and their influence on scale salience remain to be studied, but we believe that the current study represents a major step in generalizing the results obtained with tone-scramble stimuli.

2.4.2 Why not use a complete factorial design?

The reader may note that the experimental manipulations in the current study permit a factorial design, yet a factorial design was not used. The three factors manipulated in the current study each had two levels: fast and slow (presentation rate), pure and piano (timbre), and G_5 and C_4 (frequency height). A complete $2 \times 2 \times 2$ factorial design based on these manipulations yields 8 conditions, two of which were not studied here. The two omitted conditions – "Fast, Piano, G_5" and "Fast, Piano, C_4" – combined fast presentation rate

with piano timbre. We omitted these two conditions because the appropriate stimuli to use therein would instantiate notes by using only the first 65 ms of each piano recording, and truncating the piano recordings to only their first 65 ms yielded sounds that were dominated by the inharmonic attack of the piano's hammer. To our judgment, tone-scrambles based on such truncation sounded unnatural and lacked adequate tonality. We found it unlikely that scale would be more salient in these conditions compared to the "Fast, Pure, G_5" condition. As a consequence of omitting the "Fast, Piano, G_5" and "Fast, Piano, C_4" conditions, we cannot assess certain effects that would be estimable using a complete factorial design. Namely, we cannot measure the effect of presentation rate among piano-timbred stimuli, the effect of timbre among fast stimuli, nor the three-way interaction between presentation rate, timbre, and frequency height. Nonetheless, the conditions tested here represent variations on the tone-scramble task that are each more like actual music in some way. Although the 'Fast, Piano, G_5" and "Fast, Piano, C_4" conditions are absent, our results remain conclusive that presentation rate, timbre, and frequency height do not unlock scale for low-performing listeners.

2.4.3 Can we rely on data collected remotely?

We conducted the current study at a time when the COVID-19 global pandemic prompted many institutions, including UCI, to restrict in-person activities such as data collection. In response to these conditions, we conducted the current experiment remotely, allowing listeners to access the experiment via personal web browser. Our method of remote testing relinquished our control over listeners' hardware and limited our ability to perform robust psychophysical measurement. While such limitations might be unacceptable in some areas of research, extensive work has established remote testing as a valid form of study in the areas of music cognition, audiology, and general psychology (Peng et al., 2022). Our headphone screening adapted from Woods et al. (2017) suggests that listeners were largely compliant

with the instructions provided, and the correspondence between our results and previous laboratory-based studies (compare Fig 2.3, "Fast, Pure, G_5", to Fig. 2.1) provides evidence validating the use of remote testing for tone-scramble experiments.

Further, little evidence of non-compliance was found in listeners' completion times or response patterns. For example, a low-performing listener who responded as quickly as possible without attending to the stimuli might complete the experiment faster than the fastest high-performing listener. A basic analysis of response time found that the fastest high-performing listener completed the experiment in 22.4 minutes, and no low-performing listeners completed the experiment faster than this. Non-compliant participants might repeatedly give the same response over a large number of consecutive trials. Across listeners, the longest streak of one repeated response was 14, comparable to the longest streak (generated at random) of one stimulus type observed, 11. For the majority of listeners (29/33), the largest streak of one repeated response was less than 11.

In Figure 2.4 and Figure 2.5, we see that the approximated posterior distributions for the parameters α, p, and q of the "Slow, Pure, G_5" condition do not resemble a normal distribution, unlike the other approximated posterior distributions. This is likely a symptom of the tertiary mode near 75% correct observed in this condition and others. As a consequence of this mode, it seems the binomial-mixture model has two plausible configurations under the "Slow, Pure, G_5" condition. Paired comparisons of d' (not shown here) indicate that these intermediate listeners do not belong to the low-performing group of any condition, so our conclusions do not change in light of this aberration.

Chapter 3

Performance in tone-scramble tasks depends on musical scale, not on individual frequencies.

3.1 Introduction

Hundreds of studies – psychological and physiological – attest to the power of music to express and arouse our feelings (for reviews, see Eerola and Vuoskoski, 2013; Juslin, 2013; Koelsch, 2014). For the composer, a basic question is: How are the features of music related to the experience that music evokes?

Theories of composition suggest that variations in scale are central to music's meaning. For example, many studies find that, on average, music of the major scale sounds happy to listeners, and music of the minor scale sounds sad (e.g., Hevner, 1935; Crowder, 1984; Crowder, 1985a; Kastner and Crowder, 1990; Gerardi and Gerken, 1995; Gagnon and Peretz, 2003; Temperley and Tan, 2013; Bonetti and Costa, 2019). This striking difference in emotional

connotation likely explains the prominence of the major and minor diatonic scales in Western music.

Despite the broad use of major and minor scales, there is mounting evidence that many listeners are only weakly sensitive (if at all) to variations in scale. Halpern (1984) and Halpern et al. (1998) found that listeners rated melodies differing only in scale as more similar than melodies differing in rhythm or contour. When asked to *discriminate* major and minor melodies, non-musicians performed near chance on average, and musicians performed only slightly better (Halpern et al., 1998, Experiment 2; Leaver and Halpern, 2004, Discrimination Task, Experiments 1-3). The apparent lack of scale sensitivity among listeners may not be limited to melodies. Roughly one-third of the listeners studied by Blechner (1977) and Crowder (1985a) failed to differentiate major and minor chords, producing flat psychometric functions as the chords were parametrically manipulated between prototypical major and minor triads. Although the mechanisms that underlie the perception of chords almost certainly differ from those used for melodies, such results nonetheless point to a disparity among listeners in sensitivity to scale-derived musical properties.

More recent work has highlighted this disparity using a class of stimuli called "tone-scrambles." Tone-scrambles are rapid, randomly ordered sequences of brief pure tones ("pips") whose frequencies are drawn without replacement from a specific histogram. In a typical tone-scramble task, the listener is presented on each trial with a tone-scramble and strives (with trial-by-trial feedback) to classify it as one of two possible types according to its histogram. For example, in the "3-task" (Dean and Chubb, 2017), the two types of tone-scrambles to be classified both contain 8 pips of each of the notes G_5, D_6 and G_6, and in addition, major tone-scrambles contain 8 B_5-pips whereas minor tone-scrambles contain 8 $B\flat_5$-pips.

The 3-task partitions adult listeners into two distinct groups. Fig. 3.1 plots proportion correct in the 3-task, pooling results across Chubb et al. (2013), Ho et al. (2022), Dean and Chubb (2017), and Mednicoff et al. (2018). As this figure shows, most listeners ($\approx$ 70%)

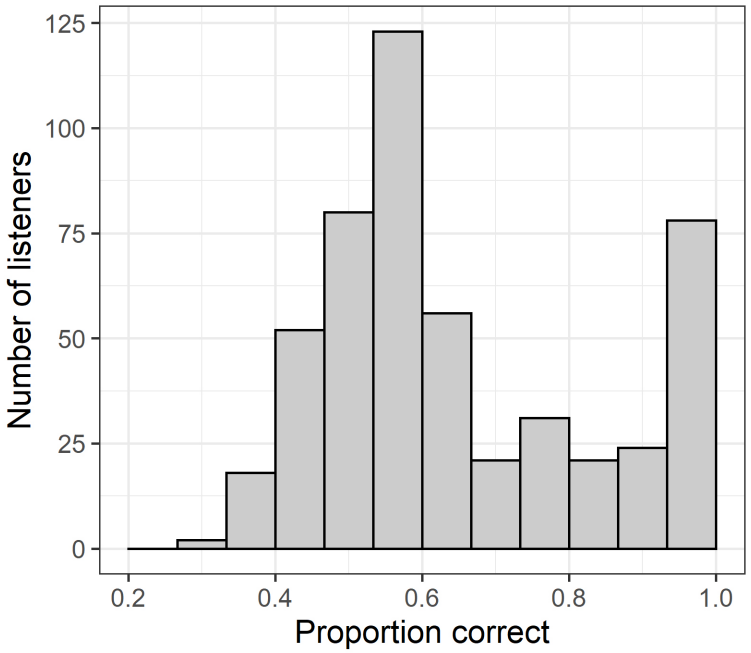

Figure 3.1: Histogram of proportion correct in the 3-task pooled over Chubb et al. (2013), Ho et al. (2022), Dean and Chubb (2017), Mednicoff et al. (2018). In each of these studies, proportion correct is based on 50 trials which were preceded by at least 40 practice trials.

perform near chance; the rest achieve near perfect accuracy. A similar bimodal distribution in performance has been observed among 6-month-old infants (Adler et al., 2020). What do high-performing listeners in the 3-task perceive that low-performing listeners seemingly cannot?

3.1.1 The current study

It is important to stress that the terms "major" and "minor" refer to the *relationships* between notes. A note on its own has no major or minor properties. Notes of a scale are imbued with such properties by their relation to the scale's tonal center, or "tonic." The notes of the G major scale, for example, are regarded with respect to the tonic G. The music-theoretic function of each note depends critically on the number of semitones that separate it from the tonic (i.e., the "interval" it forms with the tonic). In this way, B and

note:	$\hat{1}_{low}$	$\flat\hat{2}$	$\natural\hat{2}$	$\flat\hat{3}$	$\natural\hat{3}$	$\hat{4}$	$\sharp\hat{4}$	$\hat{5}$	$\flat\hat{6}$	$\natural\hat{6}$	$\flat\hat{7}$	$\natural\hat{7}$	$\hat{1}_{high}$
k:	0	1	**2**	3	**4**	5	6	7	8	**9**	10	**11**	12

Table 3.1: *Notation.* The notes indicated in the top row correspond to thirteen frequencies satisfying $f_k = f \times 2^{\frac{k}{12}}$, for some frequency f and $k = 0, 1, \cdots, 12$. In Experiment 1, f will be fixed across all trials. In Experiment 2, f will be varied randomly across trials. Each of the frequencies f_k is separated from its neighbor(s) by a twelfth of an octave (i.e., a semitone). We will use the numbers in the second row to refer to the notes in the first row. As suggested by the notation used in the top row, all of our stimuli will be constructed so that the note $0 \equiv 12$ plays the role of the tonic (where "$\equiv$" indicates that 0 and 12 have the same chroma). Those intervals marked with $\natural$ ($\flat$) symbols are called "major" ("minor"). Throughout the paper, we will embolden the numbers that refer to major intervals.

$B\flat$ are not inherently major or minor: B forms a major third interval with G (a distance of four semitones), and $B\flat$ forms a minor third interval with G (three semitones). Similarly, E ($E\flat$) functions as the major sixth (minor sixth) only with respect to G. As the tonic shifts in frequency, so do the notes that form these relationships. It is well-established that such relationships have a psychological reality and are a fundamental structuring principle in music (see Krumhansl and Cuddy, 2010, for a review). Using the notation in the bottom row of Table 3.1, the major and (natural) minor scales are composed of the following notes:

$$\text{Major} : 0(\equiv 12), \mathbf{2}, \mathbf{4}, 5, 7, \mathbf{9}, \mathbf{11},$$

$$\text{Minor} : 0(\equiv 12), \mathbf{2}, 3, 5, 7, 8, 10.$$

The first note of each scale is the tonic, and each subsequent note is a "degree" named according to its position in the scale relative to the tonic ("second," "third," "fourth," etc.). The two scales differ in their third, sixth, and seventh degrees. Going forward, we will refer to each variant of the tone-scramble task by the notes that distinguish its two stimulus types, e.g., we will hereby refer to the "3-task" described above as the 3v**4**-task.

Do listeners make use of major and minor relationships to perform the tone-scramble task? It is not immediately obvious that they do. Consider how one might design a simple sensor to discriminate the major and minor tone-scrambles used in the 3v**4**-task. Such a device would

only need to detect the presence or absence of note **4** (or alternatively of note 3), making no use of its musical relationship to other notes. All previous tone-scramble experiments have used tasks that admit strategies of this sort (Chubb et al., 2013; Dean and Chubb, 2017; Mednicoff et al., 2018; Ho and Chubb, 2020). Thus, previous results do not tell us whether high-performing listeners in the 3v**4**-task (and other tone-scramble tasks) base their judgments (1) on scale-derived qualities of the stimulus like majorness-vs-minorness (a strategy that requires the listener to relate notes to one another) or (2) on the presence vs absence of a single frequency in the stimulus (a strategy that allows the listener to ignore all but one note).

The two experiments reported here were designed to decide between these two possibilities. Experiment 1 follows the pattern of previous tone-scramble studies in testing a large group of listeners in 7 different tone-scramble tasks. Experiment 2 tests two high-performing listeners extensively in 15 different tone-scramble tasks.

3.2 Experiment 1.

In Experiment 1, we made use of results from Dean and Chubb (2017) who tested listeners in six different tone-scramble tasks. In five of their tasks, each stimulus contained the same 24 "context" pips (eight each of notes 0, 7 and 12). In addition, in each task, any given stimulus included 8 pips of a single "signal" note. In the 1v**2**-task, the signal note was either 1 or **2**; in the 3v**4**-task, the signal note was either 3 or **4**; in the 5v6-task, the signal note was either 5 or 6; in the 8v**9**-task, the signal note was either 8 or **9**, and in the 10v**11**-task, the signal note was either 10 or **11**. [1]

[1] In Dean and Chubb (2017), the 1v**2**-, 3v**4**-, 5v6-, 8v**9**- and 10v**11**-tasks were called the 2-task, 3-task, 4-task, 6-task and 7-task, respectively,.

Dean and Chubb were able to account for the observed performance of 139 listeners across these five tasks in terms of a single processing resource, R. Under the bilinear model of Dean and Chubb, each listener s possessed some quantity R_s of R, and performance in a given task t was facilitated with strength F_t by the resource R. The value of d' achieved by listener s in task t was thus predicted to be

$$d'_{s,t} = R_s F_t. \tag{3.1}$$

Strikingly, Dean and Chubb found that performance in the 1v**2**-, 3v**4**- and 8v**9**-tasks was facilitated by the resource R with approximately equal strength, i.e., $F_{1\text{v}\mathbf{2}} \approx F_{3\text{v}\mathbf{4}} \approx F_{8\text{v}\mathbf{9}}$.

All listeners in the current study were tested in the 1v**2**-, 3v**4**-, and 8v**9**-tasks as well as in four "hybrid" tone-scramble tasks: the 13v**24**-task, the 1**4**v**2**3-task, the 38v**49**-task, and the 3**9**v**4**8-task. All tone-scrambles in every task included the same fixed set of context pips: eight 0-pips, eight 7-pips and eight 12-pips. Each stimulus also included eight pips of one or more of the following signal notes: 1, **2**, 3, **4**, 8 and **9**. The histograms of the tone-scrambles used across these tasks are listed in Table 3.2. As discussed below, these seven tasks provide powerful leverage into the question of whether or not high-performing listeners rely on scale-derived qualities to make their judgments in tone-scramble tasks. Before we unpack the logic of the experiment, however, we must dispatch a potential issue in our design.

3.2.1 How we know that listeners do not base responses on mean pitch-height.

The reader will note that in each of the 1v**2**-, 3v**4**-, 8v**9**-, 13v**24**-, and 38v**49**-tasks, Type-2 stimuli are higher in mean pitch-height than Type-1 stimuli. The mean pitch-heights of Type-1 and Type-2 stimuli in all of these tasks differ by a quarter semitone (25 cents).

By contrast, Type-1 vs Type-2 stimuli in the 1**4**v**2**3- and 3**9**v**4**8-tasks have equal mean pitch-height. This raises the possibility that high-performing listeners might use stimulus mean pitch-height to perform the 1v**2**-, 3v**4**-, 8v**9**-, 13v**24**-, and 38v**49**-tasks. However, this strategy is not available in either of the 1**4**v**2**3- and 3**9**v**4**8-tasks. Thus, if high-performing listeners were basing their judgments on stimulus mean pitch-height, we would expect them to perform better in the 3v**4**-, 8v**9**-, 13v**24**-, and 38v**49**-tasks than they do in the 1**4**v**2**3-, and 3**9**v**4**8-tasks. Indeed, our results will conform precisely to this pattern.

Previous results, however, allow us to rule out the possibility that high-performing listeners base their judgments on the difference in mean pitch-height in Type-1 vs Type-2 stimuli. First, in the experiment of Dean and Chubb (2017), this strategy should have yielded equal performance in all of the tone-scramble tasks tested (in each task, Type-1 and Type-2 stimuli differed in mean pitch-height by 25 cents). However, performance was substantially lower in the 5v6- and 10v**11**-tasks than in the 1v**2**-, 3v**4**- and 8v**9**-tasks. Second, Chubb et al. (2013) tested listeners in a variant of the 3v**4**-task in which stimulus mean pitch-height was varied randomly across trials. Specifically, instead of including eight each of the notes 0 and 12 on each trial, a given stimulus included (randomly) either nine 0's and seven 12's or seven 0's and nine 12's. This manipulation injected a random perturbation of mean pitch-height of ± 37.5 cents on each trial in addition to the ± 12.5 cents produced by the random, trial-by-trial variations between major and minor scale. If high-performing listeners were making their judgments by comparing the mean pitch-height of each stimulus to a fixed criterion, this manipulation would ensure chance performance. Instead, roughly the same proportion of listeners performed near ceiling in Experiment 2 as in Experiment 1 (in which the stimuli were identical to those used in the current 3v**4**-task). We conclude that high-performers do not base their judgments on stimulus mean pitch-height.

Predictions assuming listeners base judgments on stimulus majorness-vs-minorness in the 3v4-, 8v9-, 38v49- and 39v48-tasks.

The major diatonic scale includes both **4** and **9**; in addition, all four common minor scales include 3, and three of these (the natural minor scale, the harmonic minor scale, and the descending melodic minor scale) include 8. Thus if listeners high in R use scale (majorness-vs-minorness) to perform the 3v**4**- and 8v**9**-tasks, **4**-pips and **9**-pips should both tend to heighten the perceived "majorness" of tone-scrambles in which they occur, and similarly 3-pips and 8-pips should both tend to heighten perceived "minorness."

Imagine that **4**-pips (**9**-pips) tend to alter the majorness-vs-minorness of a tone-scramble in which they occur by some amount $m_{\mathbf{4}}$ ($m_{\mathbf{9}}$), and 3-pips (8-pips) tend to alter the majorness-vs-minorness by m_3 (m_8), and suppose that $m_{\mathbf{4}} - m_3 = k$. Thus, the difference in majorness-vs-minorness produced by Type-1 vs Type-2 stimuli in the 3v**4**-task is $8(m_{\mathbf{4}} - m_3) = 8k$. The finding that $F_{3\text{v}\mathbf{4}} \approx F_{8\text{v}\mathbf{9}}$ Dean and Chubb (2017) implies that the difference in majorness-vs-minorness produced by Type-1 vs Type-2 stimuli is equal in the 3v**4**- and 8v**9**-tasks, implying that we also have $8(m_{\mathbf{9}} - m_8) = 8k$ and hence that $m_{\mathbf{9}} - m_8 = k$.

Consider, then, the Type-1 and Type-2 stimuli in the 38v**49**-task. Type-1 (Type-2) stimuli have four each of the notes 3 and 8 (**4** and **9**). Thus, it follows that the difference in majorness-vs-minorness produced by Type-2 vs Type-1 stimuli in the 38v**49**-task is

$$\begin{aligned} & 4(m_{\mathbf{4}} + m_{\mathbf{9}}) - 4(m_3 + m_8) \\ = \; & 4(m_{\mathbf{4}} - m_3) + 4(m_{\mathbf{9}} - m_8) \\ = \; & 4k + 4k = 8k; \end{aligned} \tag{3.2}$$

i.e., the difference in majorness-vs-minorness produced by Type-1 vs Type-2 stimuli in the 38v**49**-task should be equal to the difference in the 3v**4**- and 8v**9**-tasks. Hence, we should

find that $F_{38v\mathbf{49}} \approx F_{3\text{v}\mathbf{4}} \approx F_{8\text{v}\mathbf{9}}$.

By contrast, the difference in majorness-vs-minorness produced by Type-2 vs Type-1 stimuli in the 3**9**v**4**8-task is

$$\begin{aligned} & 4(m_{\mathbf{9}} + m_3) - 4(m_{\mathbf{4}} + m_8) \\ = \; & 4(m_{\mathbf{9}} - m_8) - 4(m_{\mathbf{4}} - m_3) \\ = \; & 4k - 4k = 0, \end{aligned} \tag{3.3}$$

implying that the the strategy of using majorness-vs-minorness to classify stimuli should provide no traction in the 3**9**v**4**8-task. This leads us to expect that performance should be substantially poorer in the 3**9**v**4**8-task than in the 3v**4**-, 8v**9**- or 38v**49**-tasks.

Predictions assuming listeners base judgments on the presence vs absence of signal notes in the 3v4-, 8v9- 38v49- and 39v48-tasks.

On the other hand, suppose that listeners are ignoring the context notes and are instead basing their judgments on the presence vs absence of a *single* signal note. In this case, we predict that the 38v**49**- and 3**9**v**4**8-tasks should be harder than the 3v**4**- and 8v**9**-tasks for the following reason: in contrast to the 3v**4**- and 8v**9**-tasks in which each stimulus includes 8 pips all of the same signal note, each stimulus in either of the 38v**49**- or 3**9**v**4**8-tasks contains only 4 pips of any one signal note.

In addition, two features of the stimuli in the 38v**49**- and 3**9**v**4**8-tasks make it unlikely that the listener might be able to listen for *both* signal notes in a given stimulus simultaneously: in each of these tasks, (1) each of the signal notes used in one stimulus type is separated by a single semitone from a signal note used in the other stimulus type, and (2) the two signal notes in a given stimulus are separated by either 4, 5 or 6 semitones. In other words,

the signal notes *between* stimulus types are closer in frequency than the signal notes *within* each stimulus type. It is thus unlikely that the listener will be able to apply (e.g.) a single-band filter that passes both signal notes in one stimulus type while filtering out both signal notes of the other stimulus type. Such a strategy which relies on selectively listening for the presence vs absence of both signal-note frequencies in one of the two stimulus types is likely to be less effective in the 38v**49**- and 3**9**v**4**8-tasks than in the 3v**4**- and 8v**9**-tasks.

How the 1v2-, 13v24- and 14v23-tasks fit in

The results of Dean and Chubb (2017) imply that the 1v**2**-task should be roughly equal in difficulty to the 3v**4**- and 8v**9**-tasks. However, the signal notes of the 1v**2**-task differ in musical function: **2** is in both the major and minor diatonic scales, and 1 is in neither. This observation suggests that if listeners use some scale-derived quality to make their judgments in the 1v**2**-task, it is not majorness-vs-minorness. We note, however, that, while 1 does not appear in the natural minor scale (the Aeolian mode), it does appear in the Phrygian mode. This dark mode is used in many heavy metal songs (for example, "Symbolic" by Death). Moreover, Temperley and Tan (2013), have shown that listeners tend to rate melodies in the Phrygian mode as sounding less happy than the corresponding melodies in the Aeolian mode. This suggests that 1-pips may work in concert with 3-pips to "darken" the quality of the tone-scrambles in which they occur and, correspondingly, that **2**-pips may work in concert with **4**-pips to "brighten" their quality. To the extent that listeners use the same scale-derived quality to make their judgments in both the 1v**2**- and 3v**4**-tasks, the argument above (Sec. 3.2.1) applies to show that $F_{1\mathbf{4}v\mathbf{2}3}$ should be lower than all of $F_{1v\mathbf{2}}$, $F_{3v\mathbf{4}}$ and $F_{1\mathbf{4}v\mathbf{2}3}$, which should be roughly equal.

Alternatively, it might be the case that high-performing listeners use one scale-derived quality to perform the 3v**4**-task and a different one to perform the 1v**2**-task. For example, instead of a scale-derived quality akin to majorness-vs-minorness, perhaps they use something like

"consonance-vs-dissonance" in the 1v**2**-task, with Type-1 stimuli sounding more dissonant than Type-2 stimuli. In this case, even if we replicate Dean and Chubb (2017) in showing that $F_{1v\mathbf{2}} \approx F_{3v\mathbf{4}}$, we can make no definite predictions concerning $F_{13v\mathbf{24}}$ or $F_{\mathbf{14}v\mathbf{23}}$.

What if listeners are not using mode but are rather basing their judgments on the presence vs absence of a specific signal note (or notes) in the stimulus? Predictions for the 13v**24**- and 1**4**v**2**3-tasks under this hypothesis differ slightly from those of the 38v**49**- and 3**9**v**4**8-tasks. In the 13v**24**-task, the signal notes *within* each stimulus type (two semitones apart) are farther in frequency than the corresponding signal notes *between* stimulus types (a single semitone apart); thus, similar to the 38v**49**- and 3**9**v**4**8-tasks, the hypothesized strategy is likely of limited usefulness here, resulting in low performance. However, in the 1**4**v**2**3-task, the two signal notes used in a Type-2 stimulus are separated by 3 semitones, and the two signal notes used in a Type-1 stimulus are separated by only 1 semitone. Unlike the other hybrid tasks, the close proximity of the two signal notes in the Type-1 stimuli of the 1**4**v**2**3-task raises the possibility that listeners may be able to listen for both signal notes, **2** and 3, simultaneously. If listeners are able to exploit this proximity, then we may find that performance is better in the 1**4**v**2**3-task than in the 13v**24**-, 38v**49**- and 3**9**v**4**8-tasks.

3.2.2 Methods

All methods were approved by the UCI Institutional Review Board.

Participants

The data were collected at UCI between November 2017 and June 2018. One hundred listeners participated, including the first author. All listeners were between 18 and 30 years of age with self-reported normal hearing. The mean years of musical training was 2.8 (popula-

Table 3.2: The number of pips of each note in each type of scramble. Dots stand for zero. In Experiment 1, $n = 4$; in Experiment 2, $n = 2$.

Task	Type	0	1	**2**	3	**4**	5	6	7	8	**9**	10	**11**	12
1v**2**	1	2n	2n	·	·	·	·	·	2n	·	·	·	·	2n
	2	2n	·	2n	·	·	·	·	2n	·	·	·	·	2n
3v**4**	1	2n	·	·	2n	·	·	·	2n	·	·	·	·	2n
	2	2n	·	·	·	2n	·	·	2n	·	·	·	·	2n
8v**9**	1	2n	·	·	·	·	·	·	2n	2n	·	·	·	2n
	2	2n	·	·	·	·	·	·	2n	·	2n	·	·	2n
13v**24**	1	2n	n	·	n	·	·	·	2n	·	·	·	·	2n
	2	2n	·	n	·	n	·	·	2n	·	·	·	·	2n
14v**2**3	1	2n	n	·	·	n	·	·	2n	·	·	·	·	2n
	2	2n	·	n	n	·	·	·	2n	·	·	·	·	2n
38v**49**	1	2n	·	·	n	·	·	·	2n	n	·	·	·	2n
	2	2n	·	·	·	n	·	·	2n	·	n	·	·	2n
39v**4**8	1	2n	·	·	n	·	·	·	2n	·	n	·	·	2n
	2	2n	·	·	·	n	·	·	2n	n	·	·	·	2n

tion standard deviation: 3.8) among the ninety-eight listeners who provided this information (two did not provide years of musical training). Fifty-four of these listeners reported having at least one year of musical training. Listeners were recruited through the UCI School of Social Sciences Subject Pool and were compensated with course credit.

Stimuli

Each stimulus was a tone-scramble comprising 32 consecutive "pips," each pip being a 65-ms pure tone windowed by a raised cosine function with 22.5 ms rise and decay times. Tone-scrambles were thus 2.08 s in duration. All pips had equal peak amplitude. The notes $0, 1, \cdots, 12$ were the notes $G_5, A\flat_5, \cdots, G_6$ (i.e., in Table 3.1, $f = 783.99$ Hz). Table 3.2 gives the histogram of each of Type-1 and Type-2 stimuli in each of the seven tasks tested in Experiment 1. The notes in a given tone-scramble were presented in random sequence.

Stimuli were generated in Matlab at a sampling rate of 50 kHz. The stimuli were presented diotically via JBL Elite 300 noise-cancelling headphones while the listener sat in a quiet lab.

Volume was adjusted manually by the listener to a comfortable level prior to the experiment. The listener read prompts and entered responses via the Matlab command window.

Procedure

The listener was tested in the seven tasks listed in Table 3.2, randomly blocked according to a multiple Latin square. On each trial in each task, the listener was presented with a single tone-scramble and strove to classify it as either Type 1 or Type 2. A single block consisted of a brief example sequence followed by 100 trials (50 each of Type 1 and Type 2, randomly ordered). At the end of the experiment, the listener completed a survey that collected demographic information. The only information from this survey used in this study is age and years of musical training.

Before beginning the experiment, the listener received scripted instructions from a research assistant. The listener was informed that the task was to identify each stimulus as one of two types using examples provided at the start of each block. The instructions included a description of the stimuli, incorporating both subjective terms (e.g., "happy") and music-theoretic terms (e.g., "major"). The listener was also informed that stimuli may be ambiguous in these dimensions. Upon confirming their understanding of the experimental procedure, the listener initiated the experiment.

At the outset of the first block, an announcement appeared on-screen: "*Entering Block 1. You will hear two types of stimuli. Your task is to identify whether each stimulus is Type 1 or Type 2. Press ENTER to begin training.*" Upon pressing ENTER, the listener heard two examples each of the Type-1 and Type-2 stimuli in alternating order with correct labels provided. The examples were self-paced. At the end of the example sequence, listeners saw the following prompt: "*OK. You're ready to start testing. Press ENTER to begin.*"

Stimuli were presented one at a time. The listener responded to each trial by pressing "1"

for Type 1 and "2" for Type 2. At the onset of each trial, a prompt reminded the listener of the response mapping and the current trial number, X ("*Trial [X] of 100: Enter 1 for Type 1 or 2 for Type 2.*"). Feedback ("*CORRECT*" or "*INCORRECT*") was presented after each response. Upon completing the K^{th} block, the listener was presented with the prompt "*Entering Block [K + 1]. Please take a short break. When ready, press ENTER to begin training.*" The above steps were repeated for each task.

3.2.3 Results

The dependent variable in our analysis was sensitivity (d' of signal detection theory). We used only the last 50 trials in each task to calculate d' (treating the first 50 trials as practice). To perform this calculation under our experimental paradigm, we needed to label trials as "signal" and "noise" trials arbitrarily. We chose to treat major (minor) trials as signal (noise) trials. This choice did not influence our d' measures. Each of the 100 listeners provided seven data points: one d' measure for each of the seven tasks, for a total of 700 observations. If a listener achieved a perfect hit rate (i.e., n hits out of n signal trials) in a given task, the proportion of hits was set to $\frac{n-0.5}{n}$ (as recommended by Macmillan and Kaplan, 1985).

Bilinear model

We fit a bilinear model to the observed d' values. This model has the form

$$d'_{s,t} = R_s F_t + \varepsilon_{s,t} \tag{3.4}$$

where $s = 1, \ldots, 100$ indexes the listener, $t = 1, \ldots, 7$ indexes the task, and $\varepsilon_{s,t} \sim N(0, \sigma^2)$ i.i.d. According to this model, a single latent cognitive resource R governs performance in all tasks. Different listeners possess different levels of R (hence R_s), which facilitates perfor-

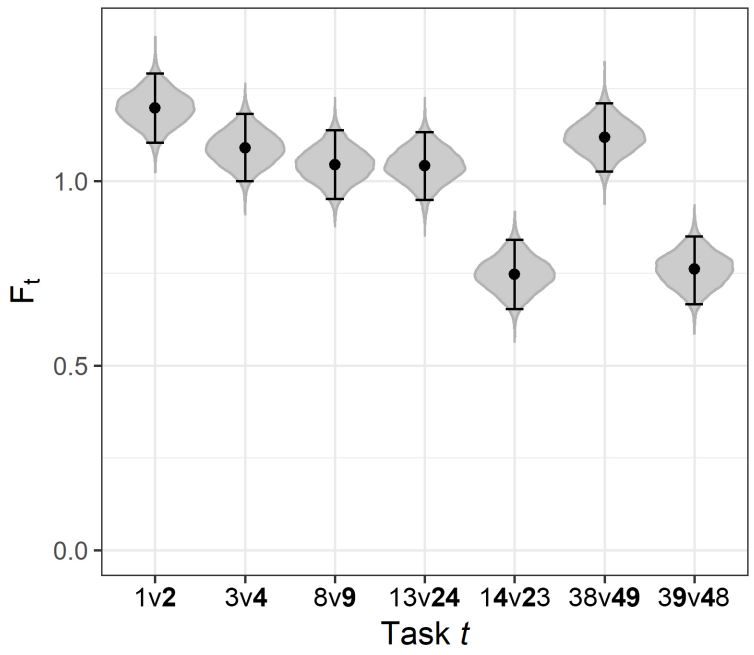

Figure 3.2: Violin plot of the estimated marginal posterior densities of F_t. Points indicate posterior medians, and error bars indicate 95% Bayesian credible intervals.

mance in task t with relative strength F_t. Note that F_t is fixed across listeners. Therefore, aside from measurement error, variation between listeners is attributed only to variation in the amount of R possessed by listeners. Note that R_s and F_t are dimensionless quantities, identifiable only up to a constant (i.e., any scalar multiplied into F_t can be absorbed by R_s to give the same predicted d'). Thus, we imposed the constraint

$$\sum_{t=1}^{7} F_t = 7. \tag{3.5}$$

This admits the following interpretation for F_t: if $F_t > 1$, task t was more strongly facilitated than the average task, and if $F_t < 1$, task t was less strongly facilitated than the average task. While we interpret R_s as the resource level possessed by listener s, under the above constraint, R_s is also the mean d' of listener s predicted by the bilinear model.

The model has a total of 108 free parameters. We estimated these parameters using Bayesian methods, assuming the following priors: $R_s \sim \text{Uniform}(-100, 100)$ i.i.d, $F_t \sim \text{Uniform}(-100, 100)$

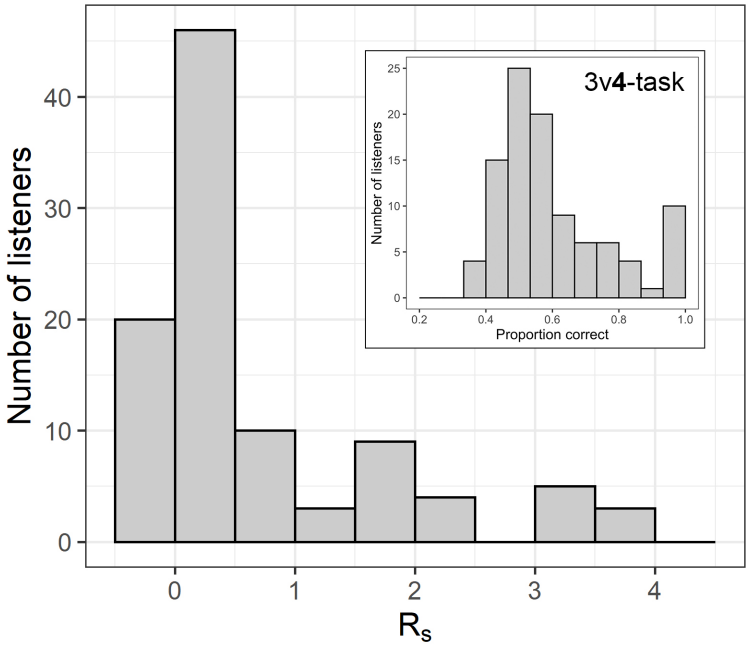

Figure 3.3: Histogram of the median posterior R_s estimates. An R of 1 corresponds to a mean d' of 1 (which corresponds to a proportion correct around 0.69). **Inset**: Histogram of proportion correct in the 3v**4**-task, based on the last 50 trials of each condition. The distribution is bimodal as in previous studies (Fig. 3.1).

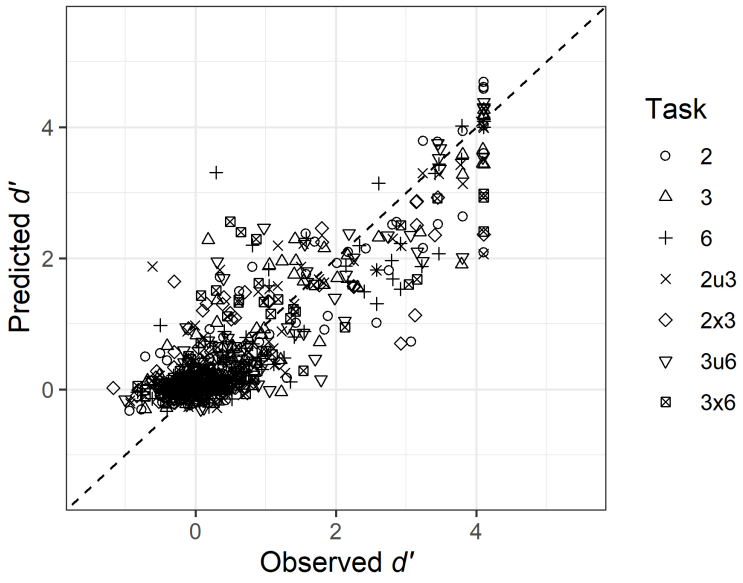

Figure 3.4: Observed d's plotted against d's predicted by the bilinear model. The dashed $x = y$ line represents a perfect fit of the data.

i.i.d., and $\sigma^2 \sim \text{Uniform}(0, 100)$. Point estimates were calculated by taking the median of 10,000,000 MCMC samples, thinned to every 1000^{th} sample. These were drawn after first taking 10,000,000 burn-in samples.

Fig. 3.2 provides a violin plot of the approximate posterior marginal densities of F_t. Fig. 3.3 shows the histogram of R_s estimates. The model explains 78.3% of the variation in observed d's (i.e., $r^2 = 0.783$).

1v2-, 3v4- and 8v9-tasks

As seen in Fig. 3.2, facilitation was approximately equal across the 1v**2**-, 3v**4**- and 8v**9**-tasks. The 95% credible interval for $F_{1v\mathbf{2}} - F_{3v\mathbf{4}}$ was [−0.035, 0.248] with a median of 0.108. The 95% credible interval for $F_{8v\mathbf{9}} - F_{3v\mathbf{4}}$ was [−0.188, 0.095] with a median of −0.045. However, the 95% credible interval for $F_{1v\mathbf{2}} - F_{8v\mathbf{9}}$ was entirely positive [0.011, 0.296] with a median of 0.153 and a posterior probability of 0.983 that this difference is greater than zero.

Hybrid tasks

The 38v**49**-task mixes the minor-notes 3 and 8 in Type-1 stimuli and the major-notes **4** and **9** in Type-2 stimuli; thus, if listeners are using stimulus majorness-vs-minorness to perform the 38v**49**-task, they should do roughly as well as they do in the 3v**4**-task and the 8v**9**-task. Indeed, this is what we find: the 95% credible interval for $F_{38\text{v}\mathbf{49}} - F_{3v\mathbf{4}}$ was [−0.113, 0.168] with a median of 0.029, and the 95% credible interval for $F_{38\text{v}\mathbf{49}} - F_{8v\mathbf{9}}$ was [−0.068, 0.215] with a median of 0.074 (i.e., $F_{38\text{v}\mathbf{49}} \approx F_{3\text{v}\mathbf{4}}$, and $F_{38\text{v}\mathbf{49}} \approx F_{8\text{v}\mathbf{9}}$). By contrast, The 3**9**v**4**8-task mixes minor-note 3 and major-note **9** in Type-1 stimuli and major-note **4** and minor-note 8 in Type-2 stimuli; thus, Type-1 and Type-2 stimuli should register similarly along the major-minor continuum. This implies that if listeners are using stimulus majorness-vs-minorness to perform the 3**9**v**4**8-task, they should perform poorly in comparison to all three of the 3v**4**-,

8v**9**- and 38v**49**-tasks. The results confirm this prediction: the median posterior facilitation in the 3**9**v**4**8-task was reduced by 32.0% relative to the 38v**49**-task. The 95% credible interval for $F_{\mathbf{39v48}} - F_{\mathbf{38v49}}$ was entirely negative [−0.499, −0.216] with a median of −0.358.

A comparable effect was observed between the 13v**24**- and 1**4**v**2**3-tasks. The 95% credible interval for $F_{\mathbf{13v24}} - F_{\mathbf{3v4}}$ was [−0.188, 0.092] with a median of −0.047 (i.e., $F_{\mathbf{13v24}} \approx F_{\mathbf{3v4}}$). However, the 95% credible interval for $F_{\mathbf{13v24}} - F_{\mathbf{1v2}}$ was entirely negative [−0.302, −0.016] with a median of −0.155 (i.e., $F_{\mathbf{13v24}} < F_{\mathbf{1v2}}$). The median posterior facilitation for the 1**4**v**2**3-task was 28.3% lower than that of the 13v**24**-task. The 95% credible interval for $F_{\mathbf{14v23}} - F_{\mathbf{13v24}}$ was entirely negative [−0.436, −0.152] with a median of −0.296.

There was no evidence of an interaction between scale degrees and their manner of hybridization: the 95% credible interval for $(F_{\mathbf{13v24}} - F_{\mathbf{14v23}}) - (F_{\mathbf{38v49}} - F_{\mathbf{39v48}})$ contained zero [−0.265, 0.137] with a median of −0.063.

3.2.4 Discussion

Why does Experiment 1 include both high- and low-performing listeners?

Given that the current project seeks to understand how high-performers perform tone-scramble tasks, the reader may wonder why Experiment 1 does not focus exclusively on high-performers. Why, one might ask, didn't we screen out all except those listeners whose performance was above some relatively high threshold? The answer rests in the least-squares fit of the bilinear model to the data. Namely, this fit automatically gives more weight to the data of high-performing listeners. If $R_s = 0$ for a particular listener s, then s's data will not influence the estimates of F_t at all: in this case, the bilinear model (as stated by Eq. 3.4) predicts a d' of 0 regardless of F_t, so s's data do not constrain F_t. In practice, no listener is estimated to have R exactly equal to zero, but for listeners estimated to have relatively low

R, F is capable of explaining relatively little variance because R mediates the influence of F. While we did not explicitly minimize squared error when fitting the bilinear model, the bilinear model's maximum-likelihood estimates are equivalent to its least-squares estimates, and our Bayesian estimates are approximately equal to the maximum-likelihood estimates due to our relatively flat prior. Thus, there is no reason to throw away the data from listener s when R_s is low. In fact, if R_s is high, then listener s's performance may saturate (i.e., s may perform perfectly) in some subset of the tasks. In this case, s's data tell us nothing about the relative difficulty of the tasks in this subset. Thus, in the current experiment, we derive most of our information about the relative difficulty of the different tasks from listeners with R intermediate between 0 and ceiling. Accordingly, we retain the data from all listeners.

Do high-performers use scale-derived qualities to classify tone-scrambles?

The current results replicate Dean and Chubb (2017) in finding that $F_{3v\mathbf{4}} \approx F_{8v\mathbf{9}}$ (Fig. 3.2). Under the assumption that listeners use stimulus majorness-vs-minorness to make their judgments in both the 3v**4**- and 8v**9**-tasks, this finding implies (as discussed in Sec. 3.2.1) that (1) $F_{38v\mathbf{49}}$ should be roughly equal to $F_{3v\mathbf{4}}$ and $F_{8v\mathbf{9}}$, and (2) $F_{\mathbf{39}v\mathbf{4}8}$ should be substantially lower. The estimates of $F_{38v\mathbf{49}}$ and $F_{\mathbf{39}v\mathbf{4}8}$ in Fig. 3.2 show precisely this pattern. The current results are thus consistent with the proposal that listeners base their judgments predominantly on majorness-vs-minorness.

In the 3**9**v**4**8-task, however, it seems that listeners use some scale-derived quality other than stimulus majorness-vs-minorness. As shown by Eq. 3.3, if listeners use only majorness-vs-minorness to make their judgments in the 3v**4**- and 8v**9**-tasks, then the finding that $F_{3v\mathbf{4}} \approx F_{8v\mathbf{9}}$ implies that $F_{\mathbf{39}v\mathbf{4}8}$ should be approximately 0, yielding chance performance regardless of listeners' R. Contrary to this prediction, however, $F_{\mathbf{39}v\mathbf{4}8}$ is significantly greater than 0.

The results of the 1v**2**-, 3v**4**-, 13v**24**- and 1**4**v**2**3-tasks parallel those of the 3v**4**-, 8v**9**- and 38v**49**- and 3**9**v**4**8-tasks. Namely, $F_{1\mathbf{4}v\mathbf{2}3}$ is substantially lower than $F_{1v\mathbf{2}}$, $F_{3v\mathbf{4}}$, and $F_{13v\mathbf{24}}$. This pattern is consistent with the proposal that, across the 1v**2**-, 3v**4**-, and 13v**24**-tasks, listeners base their judgments on the same scale-derived quality. Consequently, if listeners are using majorness-vs-minorness to perform the 3v**4**-task, then they must also be using majorness-vs-minorness to perform the 1v**2**- and 13v**24**-tasks. This is noteworthy because the notes 1 and **2** do not distinguish the major scale from the minor scales. The scale-derived quality that listeners use across these tasks must therefore be different than (but closely akin to) music-theoretic majorness-vs-minorness. This scale-derived quality might be more accurately referred to as "brightness-vs-darkness" (for example) even though we refer to it here as "majorness-vs-minorness." Moreover, analogous to the case of the 3**9**v**4**8-task, the fact that $F_{1\mathbf{4}v\mathbf{2}3}$ is significantly greater than 0 suggests that listeners were using some scale-derived quality other than brightness-vs-darkness in the 1**4**v**2**3-task.

Unlike Dean and Chubb (2017), we find that $F_{1v\mathbf{2}} > F_{8v\mathbf{9}}$. However, this effect appears to be relatively small. Consider an intermediate listener s with $R_s = 2$. Assuming that s uses an ideal signal-detection criterion, this difference in F means that s is expected to answer 1.63 more trials correctly in the 1v**2**-task than in the 8v**9**-task. By comparison, listener s is expected to answer 3.94 (4.58) more trials correctly on the 13v**24**-task than the 1**4**v**2**3-task (on the 38v**49**-task than the 3**9**v**4**8-task). The practical implications of $F_{1v\mathbf{2}} > F_{8v\mathbf{9}}$ are thus relatively small compared to our effects of interest, and taking into account the prior results of Dean and Chubb, we believe that this finding does not meaningfully change our conclusions.

Can we rescue the idea that listeners attend to specific frequencies to perform tone-scramble tasks?

We argued in Sec. 3.2.1 that if listeners perform tone-scramble tasks by selectively listening for the presence vs absence of one or more specific frequencies in the stimulus, performance should be reduced in the 38v**49**-task, for example, compared to the 3v**4**-task because the signal notes in any stimulus in the 38v**49**-task are divided between two different frequencies. Our results are at odds with this prediction and thus suggest that listeners do not attend to specific frequencies. However, there remain other possible strategies based on selective listening that we have not yet considered. What if the process used to sense the presence of a particular target frequency f_{targ} is no more effective if the stimulus contains eight pips with frequency f_{targ} than if it contains only four pips? Then, performance should be equally good across hybrid tasks. Instead, we find a difference in performance (1) between the 13v**24**- and 1**4**v**2**3-tasks and (2) between the 38v**49**- and 3**9**v**4**8-tasks. What if listeners listen simultaneously for *both* signal notes in a stimulus type using (e.g.) some multi-band filter that does not pass the signal notes of the other stimulus type? This theory, too, predicts equal performance across hybrid tasks, which is not what we observe. We conclude that listeners do not base their judgments in tone-scramble tasks on the presence vs absence of specific frequencies in the stimulus.

3.3 Experiment 2

Experiment 2 adopts a different strategy to test whether high-performing tone-scramble listeners use scale-derived qualities to perform tone-scramble tasks. This experiment tests two high-performing listeners (the two authors) in 15 different tone-scramble tasks. To rule out the possibility that the listeners base their responses on the presence vs absence of specific frequencies in the stimulus, the tonic is randomly roved from trial to trial in each task. This

manipulation also serves to increase the difficulty of the tasks to ensure that neither listener performs any task perfectly.

3.3.1 Methods

All methods were approved by the UCI Institutional Review Board.

Participants

The participants were the two authors.

3.3.2 Tasks

Each listener was tested in fifteen different tone-scramble tasks. Seven of these tasks were the 1v**2**-, 3v**4**, 8v**9**, 13v**24**-, 1**4**v**2**3-, 38v**49**- and 3**9**v**4**8-tasks of Experiment 1 (histograms listed in Table 3.2); however, in Experiment 2, stimuli contained only half as many pips as they did in Experiment 1; i.e., instead of $n = 4$ in Table 3.2, $n = 2$. The histograms of the Type-1 and Type-2 stimuli used in the other eight tasks are listed in Table 3.3. On a given trial in any of these 15 tasks, the listener heard a single tone-scramble, classified it as Type-1 or Type-2 by entering "1" or "2" on the keyboard, and received immediate, visual, correctness feedback.

As in Experiment 1, each tone-scramble was a sequence of 65-ms pure-tone pips, each pip windowed by a raised cosine function with 22.5 ms rise and decay times. In contrast to Experiment 1, on each trial, the pitch f_0 of note 0 was $f_0 = 698.46 \times 2^{\frac{X}{12}}$ Hz for $X \sim \text{Unif}(0, 4)$. Thus, the pitch of note 0 was uniformly distributed on the four-semitone interval between $F\sharp_5$ (whose frequency is 698.46 Hz) and $A\sharp_5$. The pitches of notes $1, \mathbf{2}, \cdots, 12$ increased from

Table 3.3: The number of pips of each note in the Type-1 and Type-2 tone-scrambles used in the additional tasks of Experiment 2. Dots stand for zero. The number of pips in the Type-1 and Type-2 tone-scrambles in the 1v**2**-, 3v**4**-, 8v**9**-, 13v**24**-, 1**4**v**2**3-, 38v**49**- and 3**9**v**4**8-tasks are listed in Table 3.2.

Task	Type	0	1	**2**	3	**4**	5	6	7	8	**9**	10	**11**	12
2v3	1	4	·	4	·	·	·	·	4	·	·	·	·	4
	2	4	·	·	4	·	·	·	4	·	·	·	·	4
4v8	1	4	·	·	·	4	·	·	4	·	·	·	·	4
	2	4	·	·	·	·	·	·	4	4	·	·	·	4
18v**29**	1	4	2	·	·	·	·	·	4	2	·	·	·	4
	2	4	·	2	·	·	·	·	4	·	2	·	·	4
19v**2**8	1	4	2	·	·	·	·	·	4	·	2	·	·	4
	2	4	·	2	·	·	·	·	4	2	·	·	·	4
28 v3**9**	1	4	·	2	·	·	·	·	4	2	·	·	·	4
	2	4	·	·	2	·	·	·	4	·	2	·	·	4
29v38	1	4	·	2	·	·	·	·	4	·	2	·	·	4
	2	4	·	·	2	·	·	·	4	2	·	·	·	4
23v**4**8	1	4	·	2	2	·	·	·	4	·	·	·	·	4
	2	4	·	·	·	2	·	·	4	2	·	·	·	4
24v38	1	4	·	2	·	2	·	·	4	·	·	·	·	4
	2	4	·	·	2	·	·	·	4	2	·	·	·	4

f_0 in semitone increments (i.e., $f_k = f_0 \times 2^{\frac{k}{12}}$).

Each listener completed thirty, 60-trial blocks on each of three successive days. The first 15 blocks on each day tested the listener in each of the 15 tasks. The order of the tasks in these first fifteen blocks was selected from one row of a 15×15 Latin square matrix. A different row was used for each subject on each day. The second 15 blocks on each day retested the listener in all 15 tasks in the reverse order. Prior to each block, the listener heard three examples each of the Type-1 and Type-2 stimuli from that block, alternating between the two types. The first 10 trials in the block were treated as practice (i.e., they were not included in the analysis). These first 10 trials always included five Type-1 and five Type-2 stimuli in random order.

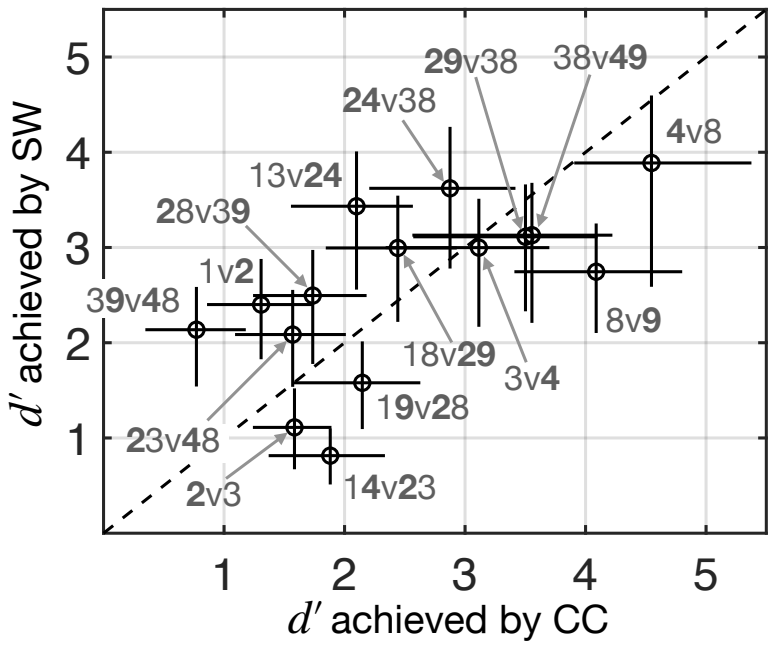

Figure 3.5: Scatter plot of d' values achieved by CC vs SW. The horizontal (vertical) line through each point gives the 95% confidence interval for the estimated d' value for CC (SW).

3.3.3 Results

The question motivating Experiment 2 is: Do high-performing listeners use scale-derived qualities to produce their responses in tone-scramble tasks? By roving the tonic across trials, we seek to rule out the strategy of listening for a specific frequency in the stimulus. Nonetheless, this manipulation leaves open other degenerate strategies. In particular, a listener might be able to deploy a "single-note" strategy in which they listen for the presence vs absence of a specific interval relative to the tonic. For example, the listener might be able to perform the 38v**49**-task by detecting the presence vs absence of **4**'s. Although one might consider "contains-**4**'s" to be a scale-derived property, a strategy based on "contains-**4**'s" is degenerate because the listener assigns all weight in their judgment to a single interval. Conceivably, one might be able to detect this quality without experiencing any of the emotional coloration produced by scale variations like major and minor. In contrast to single-note scale-derived qualities like "contains-**4**'s," we expect those scale-derived qualities that impart emotional coloration to music to be influenced by multiple notes of the scale. Thus,

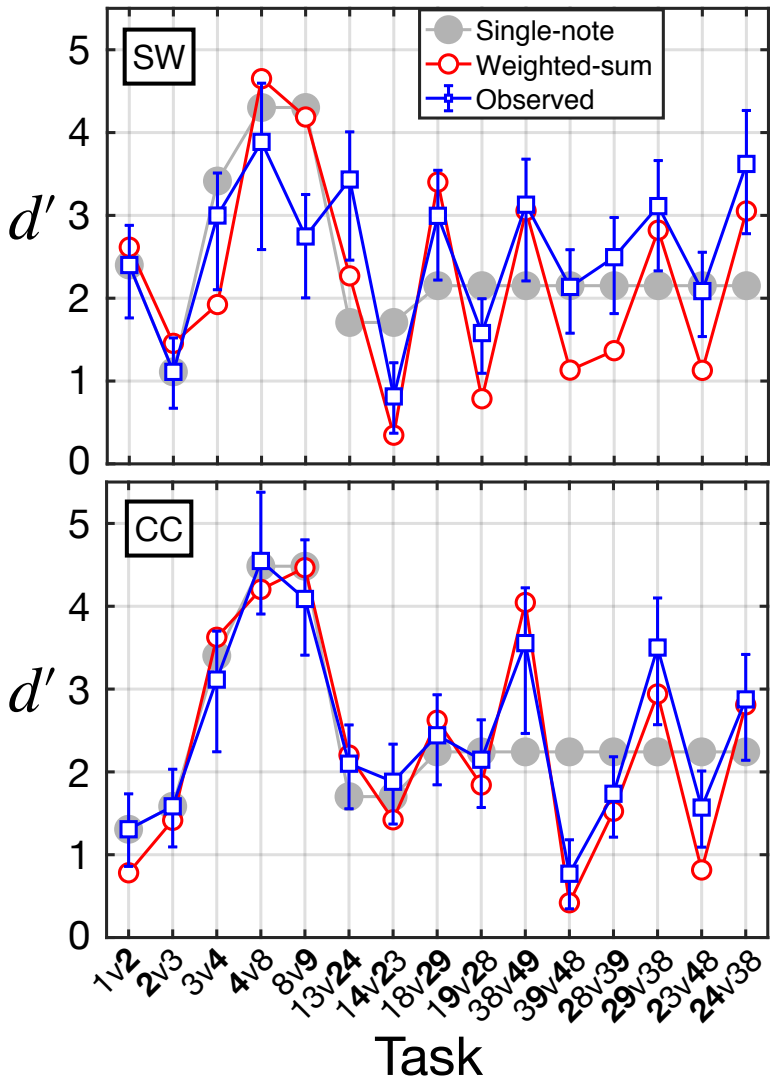

Figure 3.6: Estimated d' values achieved by SW (top) and CC (bottom) in all 15 tasks (blue lines with square markers; error bars are 95% confidence intervals). Large gray disks show the predicted d' values under the "single-note" model. Red line with circular markers gives the predicted d' values under the "weighted-sum" model.

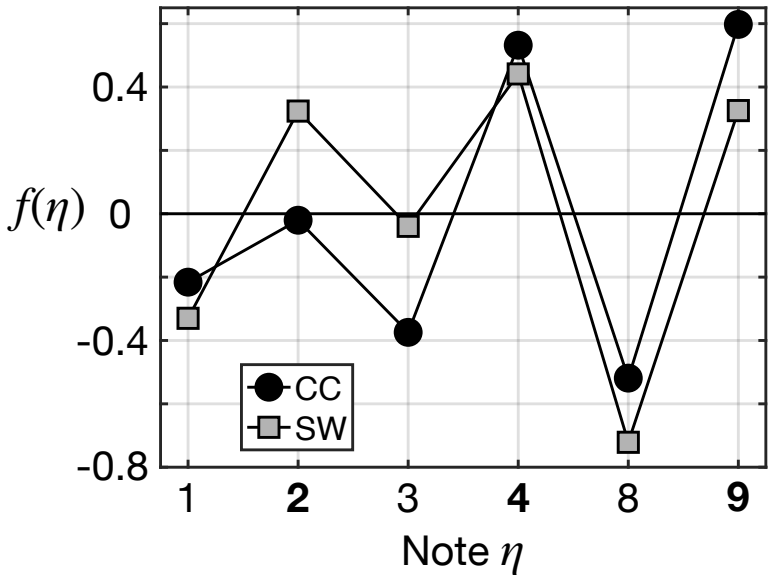

Figure 3.7: The functions f for SW (gray squares) and CC (black circles). The red line in the upper (lower) panel of Fig. 3.6 gives the predicted d' values for all tasks under the assumption that d' in each task is equal to $|f \bullet \Delta_{\text{task}}|$ for Δ_{task} equal to the difference between the histograms of the Type-2 vs Type-1 stimuli in the task.

for example, in the 38v**49**-task, if a listener is using majorness-vs-minorness to make their judgments, we expect both **4**'s and **9**'s to promote "major" responses and both 3's and 8's to promote "minor" responses. Accordingly, the analysis we undertake focuses on the question of whether SW and CC use single-note strategies or strategies that are simultaneously sensitive to multiple notes that might occur in the stimulus. To investigate this issue, we fit two different models to the data from each of SW and CC.

The single-note model.

The single-note model assumes that in every task, the listener picks out a single note (i.e., a single interval relative to the tonic) to listen for and bases their response to each stimulus on the presence vs absence of that note. We further assume that in each task, the listener listens for the particular note that will maximize performance (i.e., the note that will produce the highest value of d' in that task). Finally, for any note η, and any task T, we assume that

there exists a number $g(\eta)$ such that the value of d' that the listener can achieve by selectively listening for the presence vs absence of η's in the stimulus is equal to $n_{T,\eta}g(\eta)$, where $n_{T,\eta}$ is the (nonnegative) number of η-pips by which the histograms of Type-2 and Type-1 stimuli differ in task T. Altogether, the single-note model predicts that $d'_T = \max_{\eta}\{n_{T,\eta}g(\eta)\}$.

The weighted-sum model

The weighted-sum model assumes that the scale-derived qualities evoked by the tone-scrambles used in our 15 tasks are analogous to the colors that might be experienced by a creature with a single class of chromatic receptor. The overall color experienced by such a creature under a given light would be determined by

$$\text{Receptor response to light with spectrum } S = \int f(\lambda)S(\lambda)d\lambda = f \bullet S \tag{3.6}$$

where the integral is across all wavelengths λ in the visible range, $f(\lambda)$ is the sensitivity of the receptor-class to light of wavelength λ, and $S(\lambda)$ (the spectrum of the light) reflects the number of quanta of wavelength λ that impinge on the receptor per unit time.

Analogously, under the weighted-sum model, the scale-derived quality produced by a given tone-scramble is determined by a single system M whose response to a tone-scramble with histogram h can be characterized as follows:

$$M\text{-response to tone-scramble with histogram } h = \sum_{\eta \in \text{Notes}} f(\eta)h(\eta) = f \bullet h \tag{3.7}$$

where $f(\eta)$ reflects the sensitivity of system M to different notes η. Under this model, the d' value achieved by the listener in a given task in Experiment 2 is equal to the magnitude

of the difference in activation produced by Type-2 vs Type-1 stimuli in M. That is,

$$d'_{\text{task}} = |f \bullet (h_2 - h_1)| \tag{3.8}$$

for h_1 and h_2 the histograms of the Type-1 and Type-2 stimuli used in a given task. Note that the current results only allow us to determine the function $f(\eta)$ for the notes $\eta = 1$, **2**, 3, **4**, 8, **9** because $h_2(\eta) - h_1(\eta) = 0$ for all notes η other than these six. Note also that we can only determine f up to (1) an arbitrary sign inversion (because of the absolute value in Eq. 3.8) and (2) an arbitrary additive constant (because in every task, $\sum_\eta h_2(\eta) - h_1(\eta) = 0$). Accordingly, in order to uniquely determine f, we impose the additional constraints that (1) $f(\mathbf{9}) > f(8)$ and (2) the average value of f is 0.

The fits of the single-note and weighted-sum models to the data

We fit both the single-note and weighted-sum models to the data from each listener. Each of these two models has six parameters (although only five are free in the weighted-sum model because f must sum to 0). The parameters of the single-note (weighted-sum) model are the values $g(\eta)$ ($f(\eta)$) for $\eta = 1$, **2**, 3, **4**, 8, **9**. For each model, we find the parameters that minimize the sum of squared deviations of the predicted d''s from the observed d''s (i.e., the d''s estimated directly from the data).

Fig. 3.6 plots the d' values estimated from the data (blue lines with square markers) and their 95% confidence intervals for SW (upper panel) and CC (lower panel). The gray disks (red circles) plot d' values predicted by the single-note (weighted-sum) model. For both listeners, the single-note model does a reasonable job of capturing the d''s in the tasks in which Type-1 and Type-2 stimuli each contain only a single type of signal note (e.g., the 3v**4**-task); however, it fails to capture the variation in performance across the hybrid tasks (e.g., the 38v**49**-task).

The weighted-sum model does a strikingly good job of describing the results for CC, accounting for 93% of the the variance in d' across the 15 tasks. This model is less successful for SW, accounting for 62% of the variance. For each of SW and CC, however, the weighted-sum model provides a better description of the data (i.e., accounts for more variance) than does the single-note model: the single-note model accounts for only 29% (59%) of the variance in the estimated d''s for SW (CC).

Fig. 3.7 plots the function f (from the weighted-sum model) for SW (gray squares) and CC (black circles). In line with what we might expect from a system sensitive to majorness-vs-minorness, the function f for CC gives strongly positive (negative) weights to **4** and **9** (3 and 8) and gives weight near 0 to both 1 and **2**. By contrast, although SW seems to be less sensitive than CC to the difference between 3 and **4**, SW is more sensitive to the difference between 1 and **2**; $f(\mathbf{2}) - f(1)$ is roughly three times greater for SW than for CC. The two listeners show roughly the same pattern of sensitivity to **4**, 8 and **9**.

3.3.4 Discussion

The weighted-sum model decisively outperforms the single-note model in fitting the data for both SW and CC. This suggests that, instead of basing their responses on the simple presence vs absence of specific notes in the stimuli, the two listeners each base their responses on stimulus properties that are influenced by multiple notes. We take this as evidence that our listeners use non-degenerate, scale-derived qualities to perform the tasks in Experiment 2. How should we conceptualize these scale-derived qualities?

As discussed in Sec. 3.3.3, the weighted-sum model treats scale-derived qualities as analogous to colors experienced by a creature whose visual system possesses only a single chromatic sensor class. Many animals, however, possess more than a single chromatic sensor class. For example, human photopic vision has three: the long-, medium-, and short-wavelength

cone classes. Accordingly, we say that human color vision is 3-dimensional. Indeed, the 3-dimensionality of human color vision was discovered using behavioral experiments by Maxwell (1855) before the number of different cone types was known. Is the perception of scale-derived qualities also multidimensional?

Experiment 2 provides an ambiguous answer to this question. The current results suggest that CC may possess only a single system M that is sensitive to the scale-derived qualities produced by the tone-scrambles in Experiment 2. The remarkably good fit of the weighted-sum model to CC's data is consistent with the idea that CC used the same single system characterized by the sensitivity function f shown in Fig. 3.7 in all fifteen tasks. Perhaps CC possesses multiple systems sensitive to scale-derived qualities but applies only one to the particular set of tone-scrambles explored in Experiment 2 (this set is restricted to tone-scrambles that include four each of the context notes 0, 7 and 12 and whose signal notes are restricted to 1, **2**, 3, **4**, 8, and **9**). Yet, CC was free to use any system available to him (or any combination thereof) to optimize performance in a given task, so the fact that his results conform nearly perfectly to the weighted-sum model argues that he has only a single system available, the one characterized by the function plotted in Fig. 3.7.

Although the data of SW is fit moderately well by the weighted-sum model, this fit is markedly poorer than for CC's data. It is possible that SW was using the same single system across different tasks and that the fit is poorer because of between-task variations in the effectiveness with which this system was applied. Alternatively, it might be the case that SW has available more than a single system that is sensitive to scale-derived qualities in the stimuli of Experiment 2. If so, then the variations in d' seen in SW's data may reflect SW's use of different weighting functions across tasks.

3.4 General discussion

3.4.1 Listeners use scale-derived qualities to perform tone-scramble tasks

The current results strongly suggest that high-performing listeners in tone-scramble tasks base their judgments on non-degenerate, scale-derived qualities such as majorness-vs-minorness. They do not base their judgments on the presence vs absence of a particular note in the stimulus. For the arguments supporting this claim with reference to Experiment 1, see Sections 3.2.1, 3.2.1 3.2.1, 3.2.4, 3.2.4, and 3.2.4. For the arguments supporting this claim with reference to Experiment 2, see Sections 3.3.3 and 3.3.4.

3.4.2 Reconciling the results of Experiments 1 and 2

Although the results of Experiment 1 are well-described by the bilinear model, this model is violated by the results of Experiment 2 in several important ways. Under the bilinear model (as implied by Eq. 3.4), if listener A achieves a d' significantly greater than listener B in some task, then listener A should outperform listener B in *all* tasks. This condition is strongly violated by the results of Experiment 2. As shown in Fig. 3.5, each listener achieved significantly higher d''s than the other in several different tasks. The results of Experiment 2 depart from those of Experiment 1 in other ways. For example, the fact that $F_{3v\mathbf{4}} \approx F_{8v\mathbf{9}} \approx F_{13v\mathbf{24}} \approx F_{38v\mathbf{49}}$ in Experiment 1 suggests that each of SW and CC should achieve d''s that are approximately equal in all of the 3v**4**-, 8v**9**-, 13v**24**- and 38v**49**-tasks. Although this condition seems to be roughly satisfied by the results for SW, it is decisively violated by those for CC whose d' value in the 13v**24**-task is substantially lower than the d''s he achieves in the 3v**4**-, 8v**9**- and 38v**49**-tasks. How can we reconcile these seeming contradictions?

It is important to realize that the models used in Experiments 1 and 2 have very different purposes. The purpose of the bilinear model is to provide a coarse description of the distribution of sensitivity to various types of tone-scrambles across the general population. The model separates the distribution of R_s (the sensitivity of listener s to tone-scrambles of all types) from the relative strengths F_t (the degree to which different tasks t are, on average, facilitated by sensitivity) as shown in Fig. 3.3 and Fig. 3.2, respectively. Undoubtedly, many of the deviations of individual $d'_{s,t}$'s from the predicted values R_sF_t are due to variations in the individual patterns of sensitivity to tone-scrambles across different listeners s (as we discover is true of SW and CC in Experiment 2). These deviations from the predicted values, however, tend to be small in comparison to the large variations resulting from differences in R_s across different listeners s. For this reason, the bilinear model is able to provide a good description of the distribution of sensitivity to various types of tone-scrambles across the general population even though it ignores individual differences in sensitivity to different types of tone-scrambles.

By contrast, the weighted-sum model attempts to reveal the sensitivity of an individual listener to the different notes that occur in the stimuli. This model assumes that the notes $\eta = 1$, **2**, 3, **4**, 8, **9**, influence the scale-derived quality of a tone-scramble with weights $f(\eta)$ that are fixed across all tasks. The estimated $f(\eta)$ thus reflects the average influence exerted by each note η on the scale-derived quality that the listener uses to make their judgments (at least, to the extent that the predicted d''s capture the observed d''s). As reported in Sec. 3.3.3, the weighted-sum model fits the data very well for CC and moderately well for SW, and in each case, the sensitivity function $f(\eta)$ resembles what we would expect from a system sensitive to majorness-vs-minorness.

The contrasting purposes of the bilinear and the weighted-sum models is mirrored by important differences in the stimuli used in Experiments 1 and 2. The stimuli used in Experiment 1 all had the same fixed tonic on every trial and included 32 tones; in Experiment 2, the tonic

was randomly roved from trial to trial, and stimuli included only 16 tones. This makes the tone-scramble tasks of Experiment 2 substantially harder than the tasks used in Experiment 1. Indeed, both CC and SW routinely perform at ceiling or very close to ceiling in all of the tasks used in Experiment 1. To take one example, CC achieved $d' = 0.77$ in the 3**9**v**4**8-task in Experiment 2 (the lowest d' achieved by CC across all tasks); however, in the variant of the 3**9**v**4**8-task used in Experiment 1, CC achieved $d' = 4.11$ in a recent test (2 incorrect responses out of 100 trials).

Bibliography

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. *Journal of the Acoustical Society of America*, 147:3758–3764.

Bent, T., Bradlow, A. R., and Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 32(1):97.

Best, C. C. and McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, 46(2-3):183–216.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2):161–188.

Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLOS ONE*, 8(4).

Bidelman, G. M. and Lee, C.-C. (2015). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage*, 120:191–200.

Blechner, M. J. (1977). Musical skill and the categorical perception of harmonic mode. *Haskins Laboratories Status Report on Speech Perception*, SR-51/52:139–174.

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

Boothroyd, A. and Mackersie, C. (2017). A "Goldilocks" approach to hearing-aid self-fitting: User interactions. *American Journal of Audiology*, 26(3S):430–435.

Bradley, E. D. (2016). Phonetic dimensions of tone language effects on musical melody perception. *Psychomusicology: Music, Mind, and Brain*, 26(4):337.

Chang, D., Hedberg, N., and Wang, Y. (2016). Effects of musical and linguistic experience on categorization of lexical and melodic tones. *Journal of the Acoustical Society of America*, 139(5):2432–2447.

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. *Journal of the Acoustical Society of America*, 134(4):3067–3078.

Cooke, M. and García Lecumberri, M. L. (2021). How reliable are online speech intelligibility studies with known listener cohorts? *Journal of the Acoustical Society of America*, 150(2):1390–1401.

Crowder, R. G. (1984). Perception of the major/minor distinction: I. Historical and theoretical foundations. *Psychomusicology*, 4(1/2):3–12.

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. *Bulletin of the Psychonomic Society*, 23(4):314–316.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

Dean, T. and Chubb, C. (2017). Scale-sensitivity: A cognitive resource basic to music perception. *Journal of the Acoustical Society of America*, 142(3):1432–1440.

Deutsch, D., Henthorn, T., Marvin, E., and Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *Journal of the Acoustical Society of America*, 119:719–722.

Eerola, T. and Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception*, 30:307–340.

Gagnon, L. and Peretz, I. (2003). Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies. *Cognition and Emotion*, 17(1):25–40.

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

Giuliano, R. J., Pfordresher, P. Q., Stanley, E. M., Narayana, S., and Wicha, N. Y. (2011). Native experience with a tone language enhances pitch discrimination and the timing of neural responses to pitch change. *Frontiers in Psychology*, 2:146.

Halpern, A. R. (1984). Perception of structure in novel music. *Memory and Cognition*, 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 Perception*, 15:335–356.

Hevner, K. (1935). The affective character of the major and minor modes in music. *American Journal of Psychology*, 47:103–118.

Hilton, C. and Mehr, S. (2021). Citizen science can help to alleviate the generalizability crisis.

Ho, J. and Chubb, C. (2020). How rests and cyclic sequences influence performance in tone-scramble tasks. *Journal of the Acoustical Society of America*, 147:3859–3870.

Ho, J., Mann, D. S., Hickok, G., and Chubb, C. (2022). Inadequate pitch-difference sensitivity prevents half of all listeners from discriminating major vs minor tone sequences. *Journal of the Acoustical Society of America*, 151(5):3152–3163.

Hove, M. J., Sutherland, M. E., and Krumhansl, C. L. (2010). Ethnicity effects in relative pitch. *Psychonomic Bulletin & Review*, 17(3):310–316.

Juslin, P. N. (2013). What does music express? Basic emotions and beyond. *Frontiers in Psychology*, 4.

Kan, I. P. and Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, 83:243–253.

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

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience*, 15(3):170–180.

Kothinti, S. R., Huang, N., and Elhilali, M. (2021). Auditory salience using natural scenes: An online study. *Journal of the Acoustical Society of America*, 150(4):2952–2966.

Kraus, N. and Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, 11(8):599–605.

Kraus, N., Slater, J., Thompson, E. C., Hornickel, J., Strait, D. L., Nicol, T., and White-Schwoch, T. (2014). Music enrichment programs improve the neural encoding of speech in at-risk children. *Journal of Neuroscience*, 34(36):11913–11918.

Krumhansl, C. L. and Cuddy, L. L. (2010). A theory of tonal hierarchies in music. In et al., M. R. J., editor, *Music Perception*, volume 36 of *Springer Handbook of Auditory Research*, chapter 3. Springer.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843.

Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101.

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

Liu, J., Hilton, C. B., Bergelson, E., and Mehr, S. A. (2021). Language experience shapes music processing across 40 tonal, pitch-accented, and non-tonal languages. *bioRxiv*.

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

Maxwell, J. C. (1855). Experiments on colour, as perceived by the eye with remarks on colour-blindness. *Transactions of the Royal Society of Edinburgh*, XXI, Part II:275–298.

Mednicoff, S., Mejia, S., Rashid, J., and Chubb, C. (2018). Many listeners cannot discriminate major vs. minor tone-scrambles regardless of presentation rate. *Journal of the Acoustical Society of America*, 144(4):2242–2255.

Mehr, S. A., Singh, M., York, H., Glowacki, L., and Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, 28(3):356–368.

Merchant, G. R., Dorey, C., Porter, H. L., Buss, E., and Leibold, L. J. (2021). Feasibility of remote assessment of the binaural intelligibility level difference in school-age children. *JASA Express Letters*, 1(1):014405.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., and Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, 53(4):1551–1562.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research*, 308:98–108.

Peng, Z. E., Waz, S., Buss, E., Shen, Y., Richards, V., Bharadwaj, H., Stecker, G. C., Beim, J. A., Bosen, A. K., Braza, M. D., et al. (2022). Remote testing for psychological and physiological acoustics. *Journal of the Acoustical Society of America*, 151(5):3116–3128.

Peretz, I. and Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics*, 25(5):625–630.

Pfordresher, P. Q. and Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics*, 71(6):1385–1398.

Pike, K. (1948). Tone languages, university of michigan press. *Ann Arbor, Michigan*.

Rameau, J. P. (1971–orig., 1722). *Treatise on Harmony*. Dover Press, New York.

Schoenberg, A. (1978–orig. 1922). *Theory of Harmony*. University of California Press.

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

Turner, A. M., Engelsma, T., Taylor, J. O., Sharma, R. K., and Demiris, G. (2020). Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1230. American Medical Informatics Association.

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

van der Hulst, H., Goedemans, R., and van Zanten, E. (2010). *A survey of word accentual patterns in the languages of the world*. De Gruyter Mouton Berlin/New York.

Viswanathan, V., Shinn-Cunningham, B. G., and Heinz, M. G. (2021). Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble. *Journal of the Acoustical Society of America*, 150(4):2664–2676.

Vyas, D., Brummet, R., Anwar, Y., Jensen, J., Jorgensen, E., Wu, Y.-H., and Chipara, O. (2022). Personalizing over-the-counter hearing aids using pairwise comparisons. *Smart Health*, 23:100231.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

Woods, K. J., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception and Psychophysics*, 79(7):2064–2072.

Appendix A

Derivation of maximum-likelihood estimators and confidence intervals for single-resource model with missing data

A.1 Single-resource model

Suppose that we have d' measures for N listeners across T conditions. The single-resource model is described by the following equation:

$$d'_{s,t} = R_s F_t + \varepsilon_{s,t} \tag{A.1}$$

where $s = 1, ..., N$ indexes each subject, $t = 1, \ldots, T$ indexes each task, and $\varepsilon_{s,t} \sim N(0, \sigma)$ independently and identically distributed. In the analysis reported in Section 1.3.3, $T = 8$, and N depends on the subset of listeners being considered. Note that any scalar can be

divided from R and multiplied into F to produce the same predicted d', so for identifiability, the following constraint is imposed:

$$\sum_{t=1}^{T} F_t = T. \tag{A.2}$$

The model has a total of $N + T + 1$ parameters.

A.2 Maximum-likelihood estimators

In Section 1.3.3, several of the conditions were tested between subjects. In this case, our likelihood function is not simply expressed as a product over every combination of s and t.

Let us use the following notation: Φ is the set of all observed subject-task combinations, and (s,t) is any particular subject-task combination. Further, $(t|s)$ is the set of observed t for a given s, and likewise, $(s|t)$ is the set of observed s for a given t. In set notation, $(t|s) = \{t : (s,t) \in \Phi|s\}$, and $(s|t) = \{s : (s,t) \in \Phi|t\}$. The number of observed subject-task combinations is denoted by $\|\Phi\|$. Assuming complete data, $\|\Phi\| = 2N$ in Section 1.3.3.

Using the notation introduced above, we may state the likelihood function for the model as

$$L(R, F, \sigma) = \prod_{(s,t)\in\Phi} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{d'_{s,t} - R_s F_t}{\sigma}\right)^2\right\}. \tag{A.3}$$

The log-likelihood is thus

$$\ell(R, F, \sigma) = \|\Phi\| \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{(s,t)\in\Phi}\left(\frac{d'_{s,t} - R_s F_t}{\sigma}\right)^2. \tag{A.4}$$

Taking the first derivative with respect to each parameter,

$$\frac{\partial \ell}{\partial R_s} = \frac{1}{\sigma^2} \sum_{t \in (t|s)} (d'_{s,t} - R_s F_t) F_t \tag{A.5}$$

$$\frac{\partial \ell}{\partial F_t} = \frac{1}{\sigma^2} \sum_{s \in (s|t)} (d'_{s,t} - R_s F_t) R_s \tag{A.6}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{\|\Phi\|}{\sigma} + \frac{1}{\sigma^3} \sum_{(s,t) \in \Phi} \left(d'_{s,t} - R_s F_t\right)^2 \tag{A.7}$$

Setting each of these to zero, we find that our maximum-likelihood estimators satisfy the following equations:

$$R_s = \frac{\sum_{(t|s)} d'_{s,t} F_t}{\sum_{(t|s)} F_t^2} \tag{A.8}$$

$$F_t = \frac{\sum_{(s|t)} d'_{s,t} R_s}{\sum_{(s|t)} R_s^2} \tag{A.9}$$

$$\sigma = \sqrt{\frac{\sum_{\Phi} \left(d'_{s,t} - R_s F_t\right)^2}{\|\Phi\|}} \tag{A.10}$$

We can use conditional maximization to find the parameters that satisfy these equations. When performing conditional maximization, we partition the parameter space (in this case, into R parameters and F parameters) and iteratively find the best parameter estimates of one partition conditioned on given values of the other partition(s). In this case, we initialize the conditional maximization procedure with a naive guess at the parameters F: that $F_t = 1$ for all t. Based on this naive guess, we can use Equation A.8 to calculate an informed guess at the parameters R. In turn, we can use the informed guess at R to calculate an informed guess at F using Equation A.9. At each step, the updated parameter estimates will tend to be better than the estimate on the previous step, and we can repeat this process iteratively until convergence. In this case, after updating R and F once, the estimates of F are multiplied by a scalar to satisfy $\sum_t F_t = T$; the estimates of R are divided by the same scalar so that

the predicted values of d' (and thus the likelihood of the current parameter estimates) are unchanged. Also note that an estimate of σ is not needed to update R and F, so σ can be computed from R and F after convergence. Simulations were performed to verify that this procedure would recover input parameters to a simulated data set.

A.3 Confidence intervals

Asymptotic likelihood theory tells us that as $n \to \infty$, the distribution of the maximum-likelihood estimator, $\hat{\vec{\theta}}$ (a $p \times 1$ vector of parameter estimates), is approximately $N_p\left(\vec{\theta}, \mathcal{I}^{-1}(\vec{\theta})\right)$, where $\mathcal{I}(\vec{\theta})$ is the Fisher information matrix. In practice, we can use the *observed* Fisher information matrix, $\mathbf{I}(\vec{\theta}) = -\left[\frac{\partial^2}{\partial\vec{\theta}\partial\vec{\theta}^T}\ell(\vec{\theta})\right]$, to approximate the asymptotic covariance of $\hat{\vec{\theta}}$.

If we proceed in the usual manner to compute the observed Fisher information matrix using the above results, we will find that the observed Fisher information matrix is non-invertible, presumably because the likelihood function stated above does not incorporate the constraint $\sum_t F_t = T$ which is required for identifiability.

To incorporate the constraint, we can introduce it into the likelihood function with a Lagrange multiplier, K:

$$L(R, F, \sigma, K) = \prod_{(s,t)\in\Phi} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{d'_{s,t} - R_s F_t}{\sigma}\right)^2\right\} + K\left(T - \sum_{t=1}^{T} F_t\right). \quad \text{(A.11)}$$

Now, if we proceed as before, the first derivatives of the log-likelihood function are very

similar:

$$\frac{\partial \ell}{\partial R_s} = \frac{1}{\sigma^2} \sum_{t \in (t|s)} (d'_{s,t} - R_s F_t) F_t \tag{A.12}$$

$$\frac{\partial \ell}{\partial F_t} = \frac{1}{\sigma^2} \sum_{s \in (s|t)} (d'_{s,t} - R_s F_t) R_s - K F_t \tag{A.13}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{\|\Phi\|}{\sigma} + \frac{1}{\sigma^3} \sum_{(s,t) \in \Phi} \left(d'_{s,t} - R_s F_t\right)^2 \tag{A.14}$$

$$\frac{\partial \ell}{\partial K} = T - \sum_{t=1}^{T} F_t \tag{A.15}$$

The only difference is that $\frac{\partial \ell}{\partial F_t}$ now includes a $-KF_t$ term. Clearly, the maximum-likelihood estimator that we derived previously satisfies these equations when $K = 0$. When $K = 0$, the new $-KF_t$ term disappears. Moreover, we constrained $\sum_{t=1}^{T} F_t = T$ during our conditional maximization procedure, so $\frac{\partial \ell}{\partial K} = 0$. Thus our point estimates do not need to be changed.

Now, we may calculate the second derivatives of ℓ to get the observed Fisher information, $\mathbf{I}(\vec{\theta})$, keeping in mind that these are evaluated at our maximum-likelihood estimator (which

includes $K = 0$):

$$\frac{\partial^2 \ell}{\partial R_s^2} = -\frac{1}{\sigma^2} \sum_{(t|s)} F_t^2 \tag{A.16}$$

$$\frac{\partial^2 \ell}{\partial R_s \partial R_{s'}} = 0 \tag{A.17}$$

$$\frac{\partial^2 \ell}{\partial F_t^2} = -\frac{1}{\sigma^2} \sum_{(s|t)} R_s^2 - K \tag{A.18}$$

$$\frac{\partial^2 \ell}{\partial F_t \partial F_{t'}} = 0 \tag{A.19}$$

$$\frac{\partial^2 \ell}{\partial R_s \partial F_t} = \begin{cases} 0 & \text{if } t \notin (t|s) \\ \frac{d'_{s,t} - 2R_s F_t}{\sigma^2} & \text{otherwise} \end{cases} \tag{A.20}$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{\|\Phi\|}{\sigma^2} - \frac{3}{\sigma^4} \sum_{(s,t)\in\Phi} (d'_{s,t} - R_s F_t)^2 \tag{A.21}$$

$$\frac{\partial^2 \ell}{\partial \sigma \partial R_s} = -\frac{2}{\sigma^3} \sum_{(t|s)} (d'_{s,t} - R_s F_t) F_t \tag{A.22}$$

$$\frac{\partial^2 \ell}{\partial \sigma \partial F_t} = -\frac{2}{\sigma^3} \sum_{(s|t)} (d'_{s,t} - R_s F_t) R_s \tag{A.23}$$

$$\frac{\partial^2 \ell}{\partial K^2} = 0 \tag{A.24}$$

$$\frac{\partial^2 \ell}{\partial K \partial R_s} = 0 \tag{A.25}$$

$$\frac{\partial^2 \ell}{\partial K \partial F_t} = -F_t \tag{A.26}$$

$$\frac{\partial^2 \ell}{\partial K \partial \sigma} = 0 \tag{A.27}$$

$$\tag{A.28}$$

Reversing the order of the derivatives on the left-hand side of each equation does not change the result on the right-hand side. Thus, the above equations represent all of the elements of the observed Fisher information matrix, $\mathbf{I}(\vec{\theta})$. We can take the inverse of $\mathbf{I}(\vec{\theta})$, plug it in for the asymptotic covariance of the maximum-likelihood estimators, and construct confidence intervals based on the approximate asymptotic distribution: $\hat{\vec{\theta}} \dot{\sim} N_p\left(\vec{\theta}, \mathbf{I}^{-1}(\vec{\theta})\right)$. In practice,

we get a simultaneous confidence region for the F parameters using the (conservative) Bonferroni adjustment. It is worth noting that the matrix $\mathbf{I}(\vec{\theta})$ is of size $(N+T+1)\times(N+T+1)$. Thus, as the sample size used to fit the single-resource model increases, computing the inverse of $\mathbf{I}(\vec{\theta})$ numerically will become more resource intensive. In the case of fitting the single-resource model to the subset of native English speakers in Section 1.3.3, this would require us to invert a $33,386 \times 33,386$ matrix, which was not possible for us with the hardware available. Thus, confidence intervals were only computed for the language subsets smaller than this.