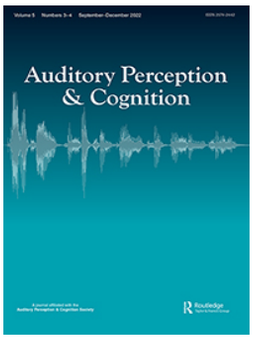

Routledge
Taylor & Francis Group

Auditory Perception & Cognition

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rpac20

An Acoustic Analysis and Comparison between Remotely Collected and In-Person Laboratory Collected Data in Vocal Imitation of Pitch

Chihiro Honda & Peter Q. Pfordresher

To cite this article: Chihiro Honda & Peter Q. Pfordresher (2023): An Acoustic Analysis and Comparison between Remotely Collected and In-Person Laboratory Collected Data in Vocal Imitation of Pitch, Auditory Perception & Cognition, DOI: 10.1080/25742442.2023.2210050

To link to this article: https://doi.org/10.1080/25742442.2023.2210050

Published online: 08 May 2023.

AUDITORY PERCEPTION & COGNITION
https://doi.org/10.1080/25742442.2023.2210050

Routledge
Taylor & Francis Group

An Acoustic Analysis and Comparison between Remotely Collected and In-Person Laboratory Collected Data in Vocal Imitation of Pitch

Chihiro Honda and Peter Q. Pfordresher

Department of Psychology, University at Buffalo, State University of New York, Buffalo, NY, USA

ABSTRACT

Restrictions on face-to-face interactions due to the outbreak of the coronavirus pandemic (COVID-19) in early 2020 have impacted experimental behavioral research. The rapid change from in-person to online data collections has been challenging in many behavioral studies, especially those that require vocal production, and the quality of the remotely collected data needs to be investigated. The current study examines the recording quality and corresponding measures of vocal production accuracy in online and in-person settings using two measurements: harmonic-to-noise ratio (HNR) and fundamental frequency, f_0. Participants imitated pitch patterns extracted from recordings of song or speech, either in a laboratory or via an online platform. The results showed that the recordings from the online setting had higher HNR than those from the in-person setting, whereas the pitch imitation accuracy in both settings did not differ. We also report an experiment that simulated differences between the online and in-person settings within participants, focusing on software used, type of microphone, and presence of ambient noise. Pitch accuracy did not differ according to these variables, except ambient noise, whereas HNR again varied across conditions. Based on these results we conclude that measures of pitch accuracy are reliable across these different types of data collection, whereas finer-grained spectral measures like HNR might be affected by various factors.

ARTICLE HISTORY
Received 18 June 2021
Accepted 25 April 2023

KEYWORDS
Online data collection; COVID-19; vocal production; pitch imitation

In early 2020, the outbreak of the novel coronavirus pandemic (COVID-19) led to school closures and restrictions on face-to-face interactions worldwide, including in the United States, and this has brought challenges to human subject research. Conventional experimental behavioral research on vocal production has typically relied on in-person data collection, in which researchers control extraneous variables, such as background noise and recording devices used to acquire data. However, during the COVID-19, experimental methods have been shifted to remote data collection, protecting the subjects'

CONTACT Peter Q. Pfordresher pqp@buffalo.edu Department of Psychology, University at Buffalo, State University of New York, Buffalo, NY 14260, USA

This material is based on work supported in part by the National Science Foundation under Grant BCS-1848930. We thank Tim Pruitt, Emma Greenspon, Fang Liu, Alice Wang, Chen Zhao, David Vollweiler, Kayden Koh, Swathi Das, Jonathan Jun Kit Liow, Anna Gentile, and Kyle Walsh for assistance in stimulus creation; Esther Song, Chantel Fatorma, Kaithlyn Massiah, Thamaraah Bouaz, and Arshpreet Grewal for help in data collection and data processing; as well as Michael Hall and two anonymous reviewers for helpful comments on an earlier version of this manuscript.

safety while sacrificing the controls over the experimental settings. Many labs have continued this practice as the pandemic has abated. Therefore, it is extremely important to evaluate the advantages and disadvantages of remote data collection, particularly the possibility of differences in the quality of data from in-person versus online settings. The current paper aims to contribute to revealing the validity of remotely collected vocal data in internet-based research.

We here report two studies that address the implications of measureing vocal pitch imitation online. During the pandemic, we initially shifted a study that was originally designed for in-person data collection to data collection via the Internet and ran two forms of data collection in parallel as COVID restrictions eased. As such, this situation provided a unique opportunity to make direct comparisons between data from the controlled lab environment and online data collection, which sacrifices control to a considerable extent.

The change to remote data collection poses a distinct challenge for vocal production studies, due to the need to acquire digital audio data that approximates the quality of online recordings well enough to draw reliable conclusions concerning the parameter(s) of interest. In the example reported here, the most critical parameter is the vocal pitch, measured using fundamental frequency, f_0. Such experiments usually have participants recorded individually in a sound-attenuated booth, with recordings of each participant collected using the same high-fidelity microphone. We know of no papers that compare the accuracy of extracted f_0 in vocal pitch production across lab-based and online recording contexts. Two recent studies (neither appearing in peer-reviewed outlets) reported that recording devices and software can affect absolute acoustic measurements, such as formant structure and vowel duration (Freeman et al., 2020; Sanker et al., 2021), whereas measure of pitch and relative differences may be more robust to different recording environments. These studies did not contrast recordings from actual online and lab-based samples, however. Other studies of online data collection have focused on the validity of stimulus presentation and responses to perceptual tasks (e.g., Bradshaw & McGettigan, 2021; Hartshorn et al., 2019; Honing & Landinig, 2008; Knoll et al., 2011; Lacherez, 2008). We therefore reasoned that online collection may be reasonably well suited to the primary measure of interest in our research, vocal f_0.

In addition to f_0, we also measured the fine-grained quality of recordings using the harmonic-to-noise ratio (HNR), the ratio of the periodic component (harmonic) to the non-periodic component (noise) in auditory signals. Our intention in doing this is to (1) confirm that recording quality varies across laboratory and online data collection settings and (2) to determine if measures of vocal pitch (f_0) are robust to variability in recording quality. There was no difference in sampling for participants across settings that would reasonably influence vocal pitch matching, thus we assumed that similarity in pitch measures should result if online data collection is valid for the purpose of measuring vocal pitch imitation.

Study 1

Study 1 compares data sets collected for a previously unpublished experiment that will be reported in more detail in a forthcoming paper. The experiment was modeled after prior studies in which participants imitated pitch patterns representative of song versus speech

production (Mantell & Pfordresher, 2013; Pfordresher et al., 2022; Wisniewski et al., 2013). In the current study, we compare data collected in two settings: a remote setting utilizing an online platform, FindingFive (FindingFive Corporation, 2021), and the Auditory Perception and Action Laboratory, based at the University at Buffalo. Participants from the in-person setting were tested in a controlled environment, whereas those in the online setting had more variable conditions (e.g., differences in room acoustics) associated with their environments, and more variable recording apparatus. To tease apart the influence of different sources of variability, we report data from a second lab experiment (only reported here) that was designed to simulate different conditions that may distinguish in-person versus online data collection.

As noted, we here compare the recording quality (i.e., HNR) and vocal production (i.e., pitch imitation accuracy) obtained from these two settings. We expected the HNR for in-person data to be higher (i.e., a greater portion of harmonic energy and less noise) than for online data because we did not control the recording environment for the online setting. Based on the robustness we had previously observed in extracting f_0 in different recording environments (e.g., Pfordresher & Demorest, 2020), we also expected that pitch imitation accuracy would be similar in the online and in-person groups if f_0 extraction is robust to differences in HNR caused by different recording environments.

Method

Participants

Sixty-five undergraduate students from the University at Buffalo, SUNY (age $M = 19.37$, $SD = 2.00$, n female = 33) participated in this study in exchange for a course credit. All participants, regardless of setting, received course credit through the research experience program associated with the University at Buffalo's Introduction to Psychology course, using the online Sona system to register for an experiment time (https://www.sona-systems.com/). Participants were given two options for the experimental setting, in-person or online. Although the setting was self-selected, by necessity, the factors that likely influenced participants' decisions for choosing the setting were of a practical nature (e.g., some students remained out of town during the semester) and not likely to cause differences in the quality of vocal pitch imitation ability. Additional three participants in the online setting were excluded from this sample due to recording failure ($N = 2$) or the issues in f_0 extractions ($N = 1$). Table 1 shows demographic statistics for participants in the online and in-person settings. This study was conducted with approval from the Institutional Review Board of the University at Buffalo.

Stimuli

The same stimuli were used for both in-person and online settings. Stimuli were designed to represent pitch patterns representative of English speech, Mandarin speech, and Song. The following section describes the process of the stimuli creation; further details of the stimuli and a discussion of their theoretical importance will be presented in a forthcoming paper.

The speech stimuli were constructed based on recordings of 48 short phrases produced by two speakers of each gender. Two speakers (one male, one female) produced sentences in English, and the other two produced matched sentences in Mandarin. Each

Table 1. Demographic variables by setting.

	In-Person		Online		
Metric variables	*M*	*SD*	*M*	*SD*	*p*(difference)
Age	19.24	1.81	19.43	2.1	.728
Years music training	2.33	3.41	3.91	4.67	.144
Years music experience	4.29	4.58	4.31	4.24	.915
Pitch discrimination	94.40%	4.50%	92.30%	7.40%	.409
Count variables	*N*		*N*		*p*(difference)
Sample size	21		44		**.006**
% female	71%		41%		**.006**
% English L1	86%		75%		.431

Values in the *p*(difference) column reflect two-sample t-tests comparing means across the in-person and online research settings. All participants in each sample were either native speakers of English (English L1) or native speakers of Mandarin. Musical experience is defined as any experience performaning a musical instrument whereas musical training is defined as having private lessons for a musical instrument.

phrase contained three to five syllables, and we manipulated the pitch contour of each phrase by varying the place of emphasis on words and the form of the phrase (falling contours in statements vs rising contours in questions). The duration of each phrase was adjusted based on the number of syllables (three syllables = 2.25 sec, four syllables = 3.00 sec, and five syllables = 3.75 sec).

We generated phonetically neutral stimuli from recordings of speech in the following way. First, we extracted the f_0 contour of each sentence using the auto-correlation function of Praat (Boersma & Weenik, 2013; for details of the pitch extraction algorithm, see Boersma, 1993). Next, we synthesized a frequency-modulated tone based on fluctuations in the extracted f_0 over time, again in Praat. These tones were given a voice-like timbre by using the Praat "hum" setting, which approximates a reduced *schwa* vowel with five formant frequencies that remain consistently spaced across changes in pitch (see https://www.fon.hum.uva.nl/praat/manual/PointProcess__To_Sound__hum____.html). The accuracy of the extracted pitch was determined by comparing the original recording to this synthetic recording. Artifacts from pitch extraction were removed by adjusting the maximum and minimum acceptable f_0 Hz values in Praat (with starting values of 450 and 50 Hz, respectively). Participants imitated only the phonetically neutral pitch contours.

The song stimuli consisted of 48 melodies created based on the speech stimuli described above. First, the f_0 that best reflected the perceived pitch in each syllable was identified by two researchers and was mapped onto the closest diatonic pitch (musical note) in the G major scale using a Praat script. Pre-recorded notes produced individually by male and female vocalists (used for the Seattle Singing Accuracy Protocol, Demorest & Pfordresher, 2015; Demorest et al., 2015; Pfordresher & Demorest, 2020), which were manipulated in pitch to reflect optimal tuning of each stable diatonic note and standardized in duration (750 ms each), were combined in sequences to create melodies. Sequences of notes were concatenated based on the order of the extracted pitch in each phrase mentioned above, without silence between them, to form each target song stimulus (melody). As in the speech stimuli, the song stimuli consisted of three to five notes with varying pitch contours, and the total duration varied based on the number of the notes (three notes = 2.25 sec, four notes = 3.00 sec, and five notes = 3.75 sec). Song

stimuli had the same phonetically neutral content and the same timbre as speech stimuli and varied only in their acoustic pitch/time structure, including stable tonal pitches and isochronous note durations. All stimuli are available online (https://osf.io/9rmha/).

Tasks

Audiometry task (in-person only): Participants were instructed to indicate when they hear a sound from their left or right ear by raising their corresponding hand. The practice trial was given with 1000 Hz at 50 dB from either left or right, and then the test trials consisting of 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz were given at 20 dB from both the left and right sides. The order of the frequencies was the same (from low to high) for all participants, but the order of the side was random for each frequency. All participants detected at least 3 out of 4 tones presented to the right ear. Three participants (14% of the sample) missed the lowest tone (500 Hz), which was attributed to the presence of ambient noise in the room used for screening (the HVAC system). We disregarded left ear responses after learning that the wire to the left earphone was compromised.

Pitch Discrimination task: Each trial consisted of two pure tones (1 sec each) with 500 ms silence between them, and the task was to identify whether the second tone was higher or lower than the first tone. The frequency of the first tone was always 500 Hz, and the second tone was selected from one of the following values: 300 Hz, 350 Hz, 400 Hz, 450 Hz, 475 Hz, 525 Hz, 550 Hz, 600 Hz, 650 Hz, and 700 Hz. Each pair was presented five times, and the order of the trials was randomized. Because the focus of the current study is on recording quality for vocal production, we do not report analyses of discrimination data here.

Pitch Imitation task: On each trial, participants listened to one of the target stimuli and were asked to imitate the sound as best as they could after the stimulus ended. Male participants were presented with the stimuli produced by male speakers, and female participants were given the stimuli produced by female speakers. Participants were randomly assigned to one of the two trial orders.

Procedure

In order to compare data from two different settings, we tried to make the procedure for the in-person and online experiments as similar as possible. The following sections describe the procedure for each setting, and Figure 1 shows an example of the screens shown during the experiment in the online and in-person settings.

In-person setting: Our laboratory re-opening plan that described protocols for safe operations during COVID-19 was submitted to and approved by the University at Buffalo, SUNY, in September 2020. The experimental room was air-purified for at least 30 min between experiments using a HEPA air filter (LEVOIT H13, 24 dB filtration) if there were multiple experiments on the same day. The participants and the experimenter both wore facial masks and kept at least 6-foot distance from each other at all times. Participants were tested individually in a sound-attenuated booth (WhisperRoom Inc.). The booth and devices used in the experiment were sanitized after each experiment. When participants arrived at the lab, they reported their health condition and, if their current conditions met safety protocols (which was the case for each participant), sat inside the sound-booth. First, participants completed the audiometry task (~5 min), the pitch discrimination task (~5 min), and then the pitch imitation (experimental) tasks

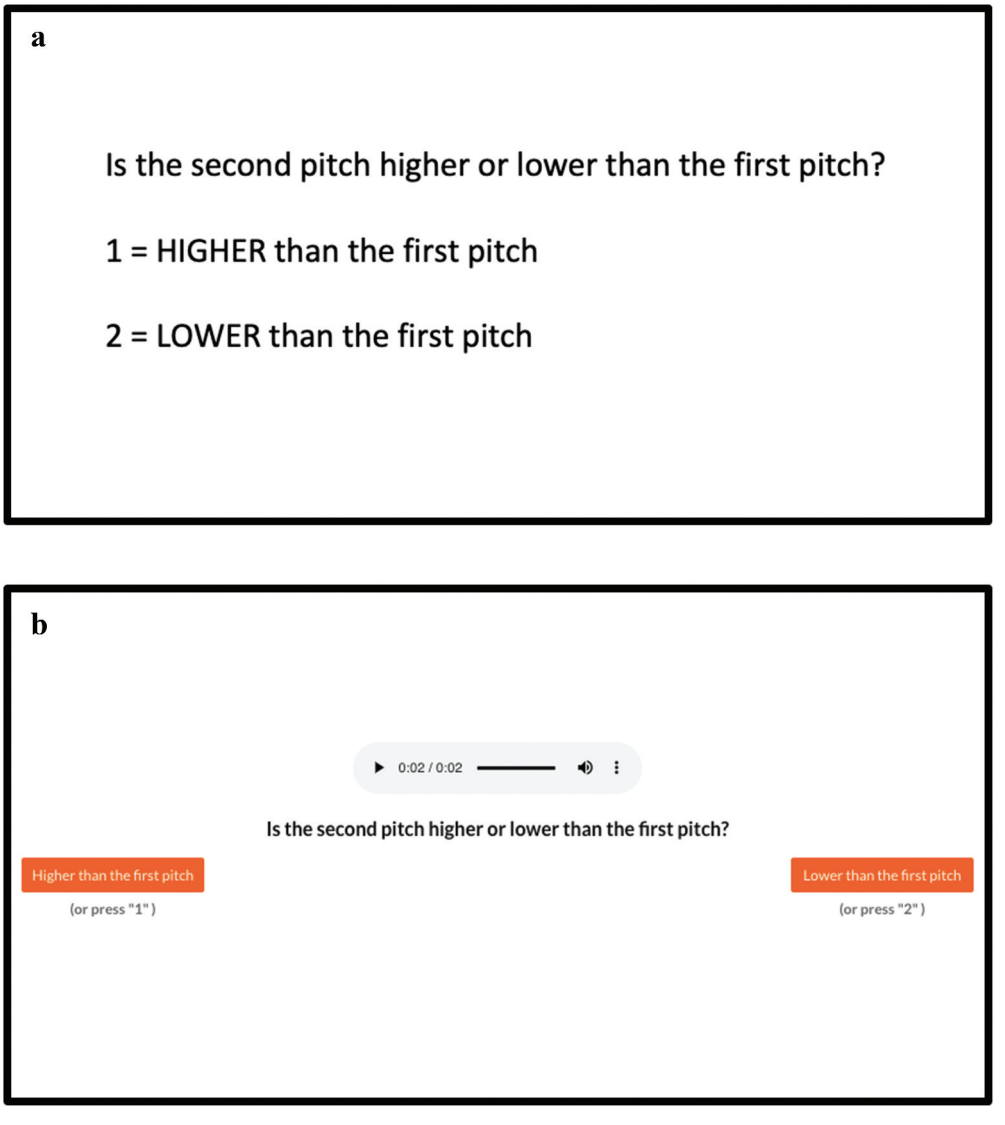

Figure 1. The screen shown during the discrimination task in the online and in-person settings. (a). In-person (b). Online

(~30 min). The experimental stimuli and prerecorded instructions were played over headphones (Sennheiser HD280 Pro) presented via a Matlab script (MathWorks, 2019) implemented on Windows 10. Participants' vocal responses were recorded via a microphone (Shure PG58) shielded by a microphone cover that was discarded after every session (BILIONE Disposable Microphone Sanitary Windscreen, 120 Pcs), and their numeric responses were recorded via a keypad (Targus). After the experimental tasks, participants completed a questionnaire about their language and music background given by the experimenter (~5 min).

Online setting: Participants individually met an experimenter who delivered instructions to the participant on the online-platform virtually via Zoom (Zoom Video

Communications, Inc., 2020). At the beginning of the session, participants were asked to sit in a quiet room. There was no audiometry task for this group, but all other tasks were identical to the in-person setting. Participants accessed the experimental tasks on an online platform, FindingFive (FindingFive Corporation, 2021), via either Chrome or Firefox on their personal computer (we did not accept a smartphone or tablet, in order to retain some control over the recording device participants used). Participants typically used a laptop and the built-in microphone and speakers for auditory processing. An instruction for each task was projected on a screen, and the stimuli were played via the platform. Participants used their own headphones/speaker device and microphone. While participants were taking the experimental tasks, the experimenter remained on Zoom to monitor their progress, but the video interface was disabled (participants were informed of this). After completing the experimental tasks, participants were given the same questionnaire as the in-person experiment by the experimenter via Zoom.

Data Analyses

The audio recordings were analyzed in the same way for the in-person and online settings. However, as noted in the beginning, the recordings collected from the online setting were encoded in the Ogg vorbis compressed format, which is a nonproprietary format for high-fidelity compression (44.1–48.0 kHz, 16+ bit, polyphonic, https://ccrma.stanford.edu/guides/planetccrma/Sound_Compression.html). Therefore, before analyzing these data, we uncompressed these audio files and converted them to the same format (wav) as the data from in-person setting by using a converter, FFmpeg (https://ffmpeg.org/), to a 44,100 Hz sampling rate. In-person participants were recorded at a 22,050 Hz sampling rate directly to a wav file. The following sections describe the initial data processing for each dependent measure.

Pitch Deviation: Participants' pitch imitation accuracy was assessed using the average difference in f_0 between the targets and their imitations. In the initial data processing, the f_0 values of each imitation were extracted from each recording by using Praat (Boersma & Weenink, 2013). An adaptive pitch-extracting function allowed users to adjust floor and ceiling settings, threshold values for silence, and octave jump costs. The accuracy of pitch extraction was assessed by comparing recorded imitations of participants with a synthesized pitch pattern based on the extracted f_0. Extractions judged to be artifactual based on audio and visual inspection, even after adjustment of parameters, were removed from the data analyses (4.62%). Auditory artifacts were determined when a synthesized frequency-modulated tone based on the extracted f_0 values did not match the original audio recording, when each was played in succession. Visual artifacts were determined when a plot of f_0 values displayed abrupt jumps in pitch that could not be generated by the human voice. The extracted f_0 and target f_0 were converted from hertz to cents with a baseline hertz (98 Hz for males and 215 Hz for females). The resulting vectors of f_0 values were used for the remaining stages of the pitch deviation analysis.

To compare the f_0 of the imitation with the target f_0, the duration of each target was adjusted to match the duration of the imitative production. First, the time stamps associated with each sampled f_0 value in the target vector were adjusted based on the ratio of the total number of samples in the target to the total number of samples in the imitation; the resulting *time transformation* value serves as a measure of how closely the timing of imitation matches the timing of imitation targets. The target vector was then re-

sampled so that each value was matched to the nearest sample from the imitation, using linear interpolation when necessary. It is important to note that the vector of f_0 values from imitative performances was never altered in this process. Time transformation values were close to and slightly higher than one ($M = 1.13$, $SD = 0.15$), indicating a slightly faster imitation relative to the target and did not vary across in-person and online settings ($p = .70$, two-sample t-test). Importantly, the time transformation ratio always stayed inside a 2:1 ratio, thus obviating the need to impute values outside the range of the two corresponding samples.

The f_0 values were converted from Hz to cents. Pitch deviations for each sample in a trial were then calculated by subtracting the f_0 values of the imitation from the corresponding target f_0 values and taking the absolute values for each difference. The resulting vector of mean absolute deviation scores were then averaged within a trial and averaged across all trials for a participant. The mean deviation for each group was computed by averaging across participants in each group (online or in-person).

Harmonicity-to-Noise Ratio (HNR): In the initial data processing, the intensity of each recording was adjusted to 70 dB. The HNR for each trial was calculated using the Praat cross-correlation function with default parameter settings and averaged across all trials for each participant, and then the mean HNR for each group (online or in-person) was computed.

Results

We first present results, shown in Table 1, that provide an assessment of how comparable the in-person and online samples are to each other based on critical measures that do not relate to audio recordings of vocal productions. The most important variables were measures relating to musical experience and training (i.e., private lessons) for vocal or instrumental performance, and pitch discrimination, given that these variables often correlate with pitch accuracy in singing (e.g., Pfordresher & Demorest, 2021). None of these variables yielded significant differences across settings as shown in Table 1. Likewise, the samples did not differ with respect to age across settings. On the other hand, the online sample was significantly larger in number and less dominated by female participants than the in-person sample. It is not clear how either of these measures would influence measures of vocal production; nevertheless, we bore these differences in mind as we interpreted other results.

We then analyzed the mean absolute difference between produced and target f_0, which is a measure of pitch accuracy in vocal imitation. This is the most critical variable for the program of research represented in this case study. Means and distributions of pitch deviation scores in different settings are shown in Figure 2(a). We analyzed these results along with the factor gender (male versus female) in a two-way between-subjects Analysis of Variance (ANOVA), which yielded a main effect of gender, $F(1, 61) = 8.90$, $p = .004$, $\eta_p^2 = .127$, with female participants yielding lower (more accurate) deviation scores ($M = 157.84$, $SD = 40.44$) than male participants ($M = 210.60$, $SD = 88.15$), but no main effect of setting ($p = .884$, $\eta_p^2 < .001$) and no interaction ($p = .824$, $\eta_p^2 = .001$).[1] Thus, as shown in Figure 2(a), differences in settings did not affect the primary measure of performance used in this line of research. Measures of vocal f_0 thus appear to be robust to differences

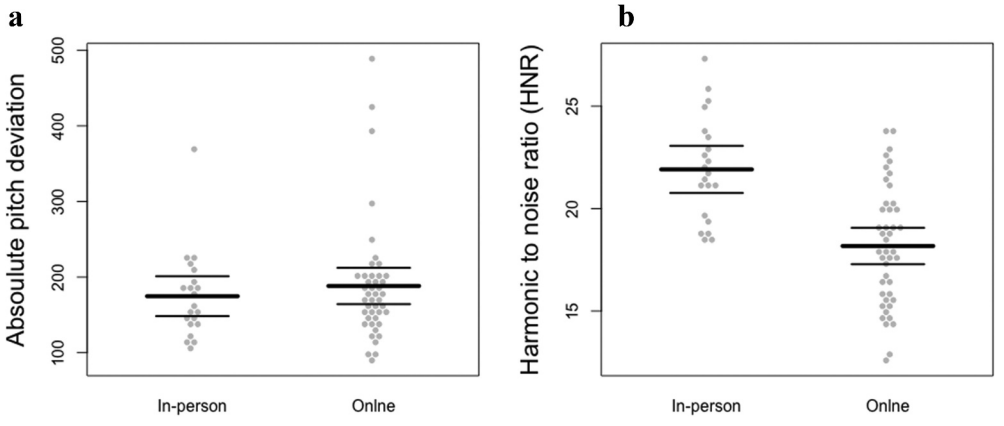

Figure 2. Differences in mean absolute pitch deviation (a) and mean HNR (b) across settings. In each panel, bold central black lines represent mean scores, surrounding black lines represent 95% confidence intervals, and gray dots represent means for individual participants. Units for HNR are dB-SPL, and units for pitch deviations are cents.

Table 2. Results by setting and participant gender.

		In Person		Online		
Measure	Gender	*M*	*SD*	*M*	*SD*	*p*(difference)
Pitch Deviation	Both	174.66	57.93	188.18	79.05	0.883
	Female	157.56	37.99	158.06	43.46	0.972
	Male	217.4	79.46	209.03	91.43	0.837
HNR	Both	21.93	2.53	18.18	2.91	**<.001**
	Female	22.55	2.58	20.18	2.85	**0.019**
	Male	20.33	1.65	16.79	2.05	**<.001**

Values in the *p*(difference) column reflect two-sample t-tests comparing means across the in-person and online research settings. Bold values are significant at α = .025 based on two non-orthogonal contrast (Both genders versus each within-gender contrast).

in recording quality across in-person and online settings that are suggested by HNR results. Due to differences between genders in overall accuracy, we also analyzed the effect of setting within each gender. As shown in Table 2, the effect of setting was not significant within male or within female participants.

Next, we considered whether measures of vocal production varied across settings. We focused on two key measures. The first measure was the harmonic-to-noise (HNR) ratio in recordings. This measure reflects the amount of periodic energy in the signal, which may be influenced by the quality of the audio file, room acoustics, and voice quality of the participant. Means and distributions of HNR in different settings are shown in Figure 2(b). We analyzed these results along with the factor gender (male versus female) in a two-way between-subjects Analysis of Variance (ANOVA), which yielded significant main effects of setting, $F\ (1,\ 61) = 17.91$, $p < .001$, $\eta_p^2 = .227$, and gender, $F\ (1,61) = 24.07$, $p < .001$, $\eta_p^2 = .284$, but no interaction ($p = .400$, $\eta_p^2 = .012$). Recordings of female participants were associated with higher HNR values ($M = 21.25$, $SD = 2.95$) than male participants ($M = 17.46$, $SD = 2.42$), as has been noted previously (Goy et al., 2013).

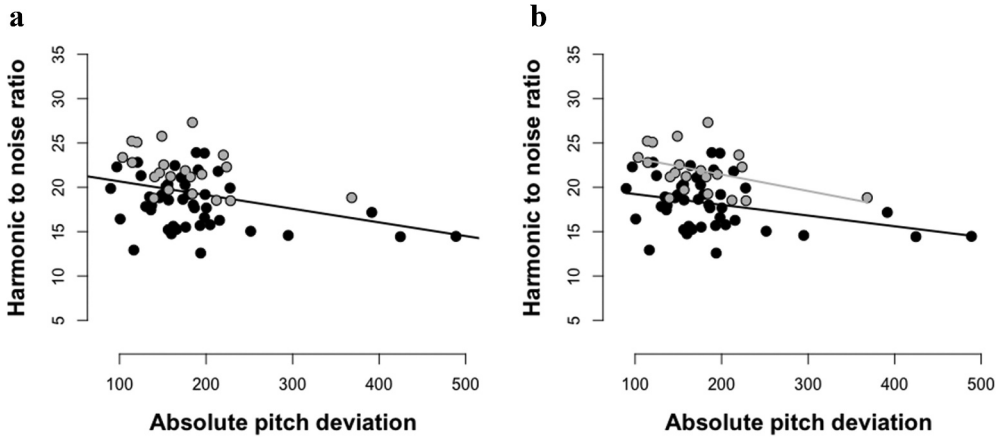

Figure 3. Correlation between mean HNR = and mean absolute pitch deviation aggregating across both settings (a) and within each setting (b). In each panel, each point dot represents the mean for individual participants; black dots represent online data and gray dots represent in-person data. Units for HNR are dB-SPL, and units for pitch deviations are cents.

However, this difference did not qualify the critical effect for the present study, which was the fact that in-person recordings yielded higher HNR values than those collected online, as shown in Figure 2(b). As can be seen in Table 2, the effect of setting was significant within female and within male participant groups.

A critical question for the present research is whether it is possible to collect pitch accuracy data online with comparable precision to data collected in person. Although the null result reported above is consistent with this claim, a better test involves the use of the Bayes Factor (BF), which can address the degree to which a result is consistent with the prior assumption of the null or alternative hypotheses.[2] Based on default (i.e., uninformative) priors in the software package JASP (version 0.11.1, https://jasp-stat.org), pitch deviation results provided modest support for the null hypothesis ($BF_{01} = 3.042$) and no support for the alternative hypothesis ($BF_{10} = 0.329$). In contrast, the HNR data provided no support for the null hypothesis ($BF_{01}<0.001$) and robust for the alternative hypothesis ($BF_{10} =$ 3,570).

We next report further analyses that address the degree of independence versus association between HNR and pitch accuracy using linear regression. When aggregating participants across both settings, there was a significant negative correlation between absolute pitch deviations and HNR, $r(63) = -.34$, $p < .001$, shown in Figure 3(a). Low HNR is thus associated with less accurate (more deviant) pitch imitation overall. Furthermore, significant associations are found within each setting [in-person, $r(19) = -.43$, $p = .027$, online, $r(42) = -.33$, $p = .015$], shown in Figure 3(b), suggesting that the association is not an artifact of significantly lower HNR in recordings from the online setting as compared to in person setting.

Discussion

Study 1 results suggest that reliable recordings of vocal pitch (f_0) can be obtained via recordings done online as well as in-person, even when sacrificing considerable control over the acoustic conditions and hardware used for recordings. Moreover, measures of pitch appeared to be robust to differences in recording quality measured using HNR. At the same time, results thus far leave open an important question. Although HNR appears to vary based on the experimental setting, some variability also seems related to individual differences that likely reflect vocal quality. In an attempt to better understand how factors contributing to the experimental setting contribute to HNR, independent of vocal quality, we attempted to simulate these differences through a follow-up experiment carried out in the lab.

Study 2

Study 2 is a controlled experiment designed to address the causes of differences observed in the data comparison reported in Study 1. Manipulated variables were designed to simulate factors that may lead to measured differences across in-person and online data collection. These included the presence of background noise (quiet for in-person recordings, possible ambient noise for online), type of microphone (external for in-person, laptop for online), and software system (Matlab recording of *.wav files for in-person, Finding Five recording of *.ogg files for online). All conditions were manipulated within subjects. We predicted that the type of microphone used would lead to differences in the HNR reported above, based on results of Sanker et al. (2021), and we anticipated that measures of pitch accuracy would not vary across different conditions.

Method

Participants

Forty-two undergraduate students at the University at Buffalo, SUNY (age $M = 18.95$, $SD = 1.12$, n female = 23) participated in this study in exchange for a course credit. All participants received course credit through the research experience program associated with the University at Buffalo's Introduction to Psychology course, using the online Sona system to register for an experiment time (https://www.sona-systems.com/). The majority of participants were native English speakers ($n = 31$, 74% of the sample). Only one participant was a native speaker of Mandarin and two more participants reported Mandarin as L2 (both were English L1).

Apparatus

All sessions took place in the Auditory Perception and Action Lab, with conditions varying to simulate differences in recordings that are likely to occur naturally during the online and in-person data collection reported in Study 1. First, half the trials were conducted using Matlab (as for in-person data from Study 1) and half were collected using Finding Favier (as for online data from Study 1). This variation in *platform* may have influenced recordings given the different file formats that were generated. Second, half the trials used an external professional-quality microphone (as for in-person data

from Study 1), and half used a built-in microphone from a Dell Latitude E5540 laptop. This variation in *microphone* was used to simulate the fact that most online participants probably used their laptop microphones in Study 1. Third, half the trials were recorded in silence (as for in-person data from Study 1), and half were recorded while white noise (created with Praat using a formula: randomGauss(0.01)) was playing at 70 dB from two speakers (CR3 Creative Reference Multimedia Monitor) placed on the left and right sides of the front monitor. This variation in *noise* was used to simulate the potential for ambient noise during online recordings in Study 1. All recordings were conducted in a Soundroom Solutions acoustic chamber.

All participants underwent the same audiometric screening as used in Study 1. All participants detected at least 3 out of 4 tones presented to the right ear. Nine participants (21% of the sample) missed the lowest tone (500 Hz), which was attributed to the presence of ambient noise in the room used for screening (the HVAC system). We disregarded left ear responses after learning that the wire to the left earphone was compromised.

Procedure and Design

The procedure was identical to the in-person sessions from Study 1. Crossing of the three variables associated with the recording environment (platform, microphone, and noise) was organized into eight blocks of trials, the order of which was counterbalanced across participants. To minimize experimenter entrances into the booth, trials with and without noise were grouped into the first and second halves of all trials, with the order of these conditions counterbalanced across participants.

Results

We first report differences in absolute pitch deviation (the primary measure for this line of research) with differences in recording platform, microphone, and ambient noise. One participant yielded deviation scores that were considerably higher than the other participants (M deviation for that participant = 900.52 cents, M for next highest participant = 346.84 cents) and was removed from all analyses. Pitch deviation scores were analyzed using a mixed-model ANOVA with the between-subjects factor gender (male/female), and within-subjects factors ambient noise (present/absent), recording platform (Matlab/Finding Five), and microphone (external/laptop). It is important to note that the within-subjects factors may be partitioned according to those that simulate features of laboratory data collection (no noise + Matlab + external mic), versus those that simulate features of online data collection (noise + Finding Five + laptop mic).

Figure 4 displays these results using the same format used for Figure 2, averaging across genders. There was a significant main effect of ambient noise, F (1,37) = 6.47, p = .015, η_p^2 = .149. Pitch deviation scores were significantly higher in the presence of ambient noise (M = 208.18, SD = 71.89) than in its absence (M = 198.34, SD = 63.71). No other main effects or interactions were significant among the within-subjects factors used to simulate different data collection environments (p > .250 in each case). It is important to note that the effect of ambient noise found here was not apparent in comparisons across recording settings from Study 1.

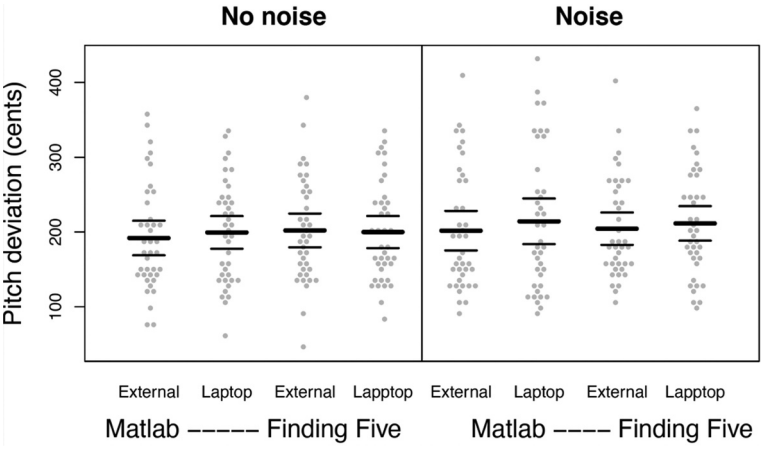

Figure 4. Differences in mean absolute pitch deviation as a function of variables associated with online and in-person recording environments. In each panel, bold central black lines represent mean scores, surrounding black lines represent 95% confidence intervals, and gray dots represent means for individual participants. Units for pitch deviations are cents. Differences in ambient noise separate left and right panels, differences in recording platforms designate the left and right halves within each panel, and differences in microphones determine grouping within each panel. Plots remove one outlying participant (see text for details).

There was also a main effect of gender, $F\ (1,37) = 5.53$, $p = .024$, $\eta_p^2 = .124$, with male participants yielding higher pitch deviation scores ($M = 224.77$, $SD = 68.19$) than female participants ($M = 180.73$, $SD = 58.82$). There were also two significant higher-order interactions with gender: gender × recording platform × ambient noise, $F\ (1,37) = 14.03$, $p < .001$, $\eta_p^2 = .275$, and gender × ambient noise × microphone, $F\ (1,37) = 7.15$, $p = .011$, $\eta_p^2 = .162$. Both interactions reflected the fact that differences across genders were not stable. With respect to the interaction with setting and ambient noise, gender differences were significant for recordings using Finding Five in a quiet environment or Matlab in a noisy environment ($p = .04$ in each case, two-sample t-tests), and slightly above the criterion for significance in the other conditions. With respect to the interaction with ambient noise and microphone type, gender differences were significant for recordings with the external mic in a quiet

Table 3. Absolute pitch deviations by participant gender, ambient noise, software, and microphone conditions.

			Female		Male	
Noise	Software	Microphone	*M*	*SD*	*M*	*SD*
Absent	Matlab	External	165.498	60.175	217.079	73.511
		Laptop	185.457	65.214	212.664	68.488
	Finding Five	External	182.82	64.902	220.453	70.589
		Laptop	174.819	56.988	223.737	66.812
Present	Matlab	External	174.747	73.036	227.443	82.596
		Laptop	185.663	62.952	241.511	111.057
	Finding Five	External	197.149	68.388	211.348	66.303
		Laptop	178.088	62.862	243.218	64.969

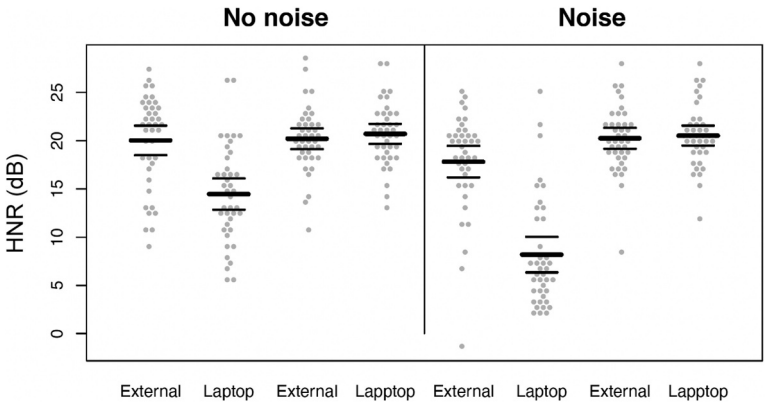

Figure 5. Differences in harmonic-to-noise ratio as a function of variables associated with online and in-person recording environments. In each panel, bold central black lines represent mean scores, surrounding black lines represent 95% confidence intervals, and gray dots represent means for individual participants. Units for HNR are dB-SPL Differences in ambient noise separate left and right panels, differences in recording platforms designate the left and right halves within each panel, and differences in microphone determine grouping within each panel.

environment ($p = .04$), or for recordings of the laptop mic in a noisy environment ($p = .01$), but not for other conditions. It is not clear if any of these differences suggest a systematic problem in conducting online experiments concerning vocal pitch imitation. Table 3 displays mean values across all conditions broken down by participant gender.

Next, we analyzed the mean HNR (harmonic-to-noise ratio) per trial in the same way; Figure 5 shows means and confidence intervals across within-subjects conditions. Every ANOVA effect was significant ($p < .001$ in every case), reflecting the dominant effect of a significant three-way ambient noise × recording platform × microphone interaction, $F\ (1,37) = 37.87$, $p < .001$, $\eta_p^2 = .506$. Post-hoc pairwise comparisons with a Bonferroni comparison (familywise $\alpha = .05$) revealed that this interaction was due to conditions involving the use of the laptop microphone while recording via Matlab. These two conditions yielded lower HNR (indicating poorer recording quality) than all other conditions. In addition, HNRs during Matlab recordings with the laptop microphone were lower in the presence of ambient noise than when ambient noise was absent. No other pairs of conditions differed significantly from each other. These effects also do not clearly replicate results from Study 1, which would have led to a difference between the combination of Finding Five recording with the laptop microphone (the standard setup for online recordings) versus Matlab recording with the external microphone (the setup used for in-person recordings).

Although the main effect of gender was not significant ($p = .381$), unlike Study 1, several interactions with gender were significant: gender × microphone, $F\ (1,37) = 5.96$, $p = .020$, $\eta_p^2 = .139$, and gender × recording platform × microphone, $F\ (1,37) = 5.22$, $p = .028$, $\eta_p^2 = .124$. A contrast analysis of the three-way interaction, similar to analyses performed for absolute pitch deviation scores, suggested that differences across genders were not significant for recordings using Matlab and an external microphone ($p = .53$)

Table 4. Harmonic-to-noise ratios by participant gender, ambient noise, software, and microphone conditions.

			Female		Male	
Noise	Software	Microphone	*M*	*SD*	*M*	*SD*
Absent	Matlab	External	21.173	4.869	19.402	4.246
		Laptop	13.296	5.398	15.972	4.154
	Finding Five	External	21.873	2.974	19.079	2.527
		Laptop	22.509	2.725	19.378	2.48
Present	Matlab	External	18.049	6.176	17.913	3.927
		Laptop	6.575	5.493	9.999	5.624
	Finding Five	External	22.002	2.724	19.165	2.212
		Laptop	21.929	2.932	19.635	2.458

but were significant for other combinations of microphones and recording platforms ($p < .04$ in each case). Furthermore, there was not a consistent difference across male and female participants by conditions, as can be seen in Table 4.

Study 1 also revealed a significant association between absolute pitch deviation scores and HNR scores across participants both within and across data collection settings. Study 2 did not replicate this effect. Although there was a significant association when all participants were included, $r(40) = -.32$, $p = .039$, this correlation was entirely attributable to the outlier that was removed in all analyses reported so far and became non-significant when this participant was removed, $r(39) = .04$, $p = .821$. A similar pattern was found when computing correlations based on data within each of the eight cells resulting from the three manipulated variables (i.e., significant associations that are due to the single outlier).

Discussion

Many results from Study 1, when comparing data that were collected online versus in-person due simply to circumstance, were not replicated in Study 2, which included experimental manipulations of conditions that varied across the data settings in Study 1. Whereas measures of pitch accuracy and absolute pitch deviation scores decreased in the presence of ambient noise in Study 2 (a manipulation meant to simulate ambient noise that can be present during online data collection), a similar decrement was not found for online data in Study 1. A more puzzling difference was that HNR varied in both studies across conditions but in somewhat different ways. Specifically, Study 2 found a more complex association in which HNR decreased for a combination of conditions that were not present in Study 1, namely data collection through Matlab (characteristic of in-person data collection in Study 1) while recording with the built-in laptop microphone (characteristic of online data collection in Study 1). Furthermore, Study 2 did not replicate the association between absolute pitch deviation and HNR found in Study 1.

These differences likely reflect the difficulty of simulating the complexity of factors that may vary in online data collection, as well as differences in the design of each study. First, the difference across studies in the effect of ambient noise may reflect the fact that the white noise (at approximately 70 dB) used in Study 2 was an exaggerated version of possible ambient noise in Study 1. Second, the difference across studies in HNR may be

due to a difference in HNR across participant groups in Study 1 that did not carry over to the within-subjects design in Study 2.

The interaction between data collection platform (Finding Five versus Matlab) and microphones is more complex. After further analysis and listening to the files, our tentative conclusion is that this interaction occurs because the uncompressed audio files collected by Matlab were more sensitive to differences in microphone recording quality than the compressed *.ogg files created by Finding Five. Thus, microphone differences yielded different HNR values for Matlab that were not evident from Finding Five recordings. It is important to note again that this difference would not have led to the different HNR measurements found in Study 1.

Aside from these differences, the most important test of this research was replicated in Study 2. Like Study 1, Study 2 yielded no convincing evidence that online data collection compromises the quality of pitch matching accuracy data. In fact, Study 2 suggests that the apparent differences in HNR across settings in Study 1 may have been an artifact related to the between-subjects design.

General Discussion

The rapid shift to online data collection after the breakout of COVID-19 enabled us to collect data from those who are unable to be physically present in a laboratory setting. However, due to the lack of controls during online data collection, the analysis and interpretation of remotely collected data must be assessed with caution, especially for studies that involve vocal productions. Current studies have investigated the quality of remotely collected data, specifically vocal pitch (fundamental frequency, f_0) and HNR.

In the first study, we examined the difference between in-person and remote settings in these two measures by using the data collected during the pandemic. In the second study, we further investigated the differences between the two settings by manipulating the factors related to the variabilities in online and in-person data collection (i.e., background noise, type of microphone, and software system) within participants. Since these two studies had different approaches in investigating the data quality (see Discussion in Study 1 and 2), here we focus on the results of accuracy f_0 and HNR.

As noted at the outset, this research focuses on the accuracy of vocal pitch imitation and as such vocal f_0 constitutes the primary variable of interest. Our analyses found no difference in vocal pitch accuracy (measured using pitch deviations from target to imitation) across settings, and a Bayesian analysis indicated moderate support for the null hypothesis of no difference across settings. Thus, we conclude that online data collection is sufficient for reliable measures of vocal pitch accuracy. This is a benefit not only for the pandemic but also with respect to broadening the scope of research beyond traditional college-age populations.

In contrast, HNR in Study 1 varied significantly across settings, and was lower for online recordings than recordings collected in person. We also found that HNR varied significantly across variables in Study 2 though not in ways that reflected different measurements in Study 1. Although it is hard to draw comparisons across studies, the data do suggest overall that HNR is more vulnerable to differences in the details of recording than pitch accuracy; a point that is worth noting for auditory researchers interested in online data collection.

Based on these results, there are a few implications and suggestions for future online-based studies in signing and speech production. First, as mentioned above, the f_0 in remotely collected data can be as reliable as those in lab-collected data. However, it is important to note that the present study examined pitch imitations without any phonetic variations and did not involve much articulation of speech that would affect f_0. Further studies are needed to investigate the interaction between f_0 extraction of natural speech and noise in remote settings. Second, the noise component in the signals can be reduced by using a high-quality recording device. Results from Study 2 further suggest that uncompressed file formats (e.g., *.wav files) may be more sensitive to the effects of noise and microphone quality than compressed (e.g., *.ogg) formats when it comes to HNR though there was no effect on measures of pitch accuracy. Therefore, we believe that online data collection can be used as an alternative to in-person data collection for analyses of pitch production accuracy, although research questions that require fine-grained spectral analyses may not be as well suited to online data collection.

Notes

1. Levene's test for homogeneity of variance was not significant for either of the ANOVAs reported here.
2. Bayesian tests are also robust to deviations from normalcy, and the distribution of pitch deviation scores in Figure 2(b) do deviate from normalcy according to a Shapiro–Wilk test ($p < .001$).

Acknowledgments

This research was supported in part by NSF Grant BCS-1848930. We thank Tim Pruitt, Emma Greenspon, Fang Liu, Alice Wang, Chen Zhao, David Vollweiler, Kayden Koh, Swathi Das, Jonathan Jun Kit Liow, Anna Gentile, and Kyle Walsh for assistance in stimulus creation; Esther Song, Chantel Fatorma, Kaithlyn Massiah, Thamaraah Bouaz, and Arshpreet Grewal for help in data collection and data processing, as well as Michael Hall and two anonymous reviewers for helpful comments on an earlier version of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the National Science Foundation of United States [BCS-1848930].

References

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, *17*, 97–110. The University of Amsterdam.

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer [Computer program]*. Version 5.4.09, Retrieved August 25, 2014, from http://www.praat.org/

Bradshaw, A. R., & McGettigan, C. (2021). Convergence in voice fundamental frequency during synchronous speech. *PloS One*, *16*(10), e0258747. https://doi.org/10.1371/journal.pone.0258747

Demorest, S. M., & Pfordresher, P. Q. (2015). Singing accuracy development from K-adult: A comparative study. *Music Perception*, *32*(3), 293–302. https://doi.org/10.1525/mp.2015.32.3.293

Demorest, S. M., Pfordresher, P. Q., Dalla Bella, S., Hutchins, S., Loui, P., Rutkowski, J., & Welch, G. F. (2015). Methodological perspectives on singing accuracy: An introduction to the special issue on singing accuracy (Part 2). *Music Perception*, *32*(3), 266–271. https://doi.org/10.1525/mp.2015.32.3.266

FindingFive Team (2021). FindingFive: A web platform for creating, running, and managing your studies in one place. FindingFive Corporation (nonprofit). https://www.findingfive.com

Freeman, V., De Decker, P., & Landers, M. Suitability of self-recordings and video calls: Vowel formants and nasal spectra. (2020). *The Journal of the Acoustical Society of America*, *148*(4), 2714–2715. Published Abstract. https://doi.org/10.1121/1.5147526

Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative voice data for younger and older adults. *Journal of Voice*, *27*(5), 545–555. https://doi.org/10.1016/j.jvoice.2013.03.002

Hartshorn, J. K., de Leeuw, J. R., Gooman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z

Honing, H., & Landinig, O. (2008). The potential of the internet for music perception research: A comment on lab-based versus web-based studies. *Empirical Musicology Review*, *3*(1), 4–7. https://doi.org/10.18061/1811/31692

Knoll, M. A., Uther, M., & Costall, A. (2011). Using the internet for speech research: An evaluative study examining affect in speech. *Behaviour & Information Technology*, *30*(6), 845–851. https://doi.org/10.1080/0144929X.2011.577192

Lacherez, P. F. (2008). The internal validity of web-based studies. *Empirical Musicology Review*, *3*(3), 161–162. https://doi.org/10.18061/1811/34107

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of speech and song. *Cognition*, *127*(2), 177–202. https://doi.org/10.1016/j.cognition.2012.12.008

The MathWorks Inc. (2019). MATLAB (R2019a). https://www.mathworks.com

Pfordresher, P. Q., & Demorest, S. M. (2020). Construction and validation of the seattle singing accuracy protocol (SSAP): An automated online measure of singing accuracy. In F. Russo, B. Ilari, & A. Cohen (Eds.), *Routledge companion to interdisciplinary studies in singing: Vol. 1 development* (pp. 322–333). Routledge.

Pfordresher, P. Q., & Demorest, S. M. (2021). The prevalence and correlates of accurate singing. *Journal of Research in Music Education*, *69*, 5–23.

Pfordresher, P. Q., Mantell, J. T., & Pruitt, T. A. (2022). Effects of intention in the imitation of sung and spoken pitch. *Psychological Research*, *86*(3), 792–807. https://doi.org/10.1007/s00426-021-01527-0

Sanker, C., Babinski, S., Burns, R., Evans, M., Kim, J., Smith, S., Weber, N., & Bowern, C. (2021). (Don't) try this at home! The effects of recording devices and software on phonetic analysis. *Lingbuzz*. https://ling.auf.net/lingbuzz/005748

Wisniewski, M. G., Mantell, J. T., & Pfordresher, P. Q. (2013). Transfer effects in the vocal imitation of speech and song. *Psychomusicology: Music, Mind, and Brain*, *23*(2), 82–99. https://doi.org/10.1037/a0033299

Zoom Video Communications Inc. (2020). Zoom. https://zoom.us/