Author/s:
Tan, YT;Peretz, I;McPherson, GE;Wilson, SJ

Title:
Establishing the Reliability and Validity of Web-Based Singing Research

Date:
2021-04-01

Citation:
Tan, Y. T., Peretz, I., McPherson, G. E. & Wilson, S. J. (2021). Establishing the Reliability and Validity of Web-Based Singing Research. Music Perception, 38 (4), pp.386-405. https://doi.org/10.1525/MP.2021.38.4.386.

Persistent Link:
https://hdl.handle.net/11343/275074

Establishing the Reliability and Validity of Web-based Singing Research

Yi Ting Tan
University of Melbourne, Parkville, Australia

Isabelle Peretz
Université de Montréal, Montreal, Canada

Gary E. McPherson
University of Melbourne, Southbank, Australia

Sarah J. Wilson
University of Melbourne, Parkville, Australia

In this study, the robustness of an online tool for objectively assessing singing ability was examined by: (1) determining the internal consistency and test-retest reliability of the tool; (2) comparing the task performance of web-based participants ($n = 285$) with a group ($n = 52$) completing the tool in a controlled laboratory setting, and then determining the convergent validity between settings, and (3) comparing participants' task performance with previous research using similar singing tasks and populations. Results indicated that the online singing tool exhibited high internal consistency (Cronbach's alpha = .92), and moderate-to-high test-retest reliabilities (.65–.80) across an average 4.5-year-span. Task performance for web- and laboratory-based participants ($n = 82$) matched on age, sex, and music training were not significantly different. Moderate-to-large correlations ($|r|$ =.31–.59) were found between self-rated singing ability and the various singing tasks, supporting convergent validity. Finally, task performance of the web-based sample was not significantly different to previously reported findings. Overall the findings support the robustness of the online tool for objectively measuring singing pitch accuracy beyond a controlled laboratory environment and its potential application in large-scale investigations of singing and music ability.

Received: February 18, 2020, accepted November 26, 2020.

Key words: singing ability, online assessment, reliability, validity, web-based research

Web-based experimental research has flourished since the beginning of the internet revolution in the 1990s (Gosling & Mason, 2015; Reips, 2002). Initially, the use of the internet was the prerogative of a technologically savvy minority, leaving most researchers skeptical about the feasibility and utility of conducting research online. However, with more than half the world population (and more than 80% of developed countries) now having internet access (Internet World Stats, 2019), and the growing indispensability of internet usage in our daily lives, web-based research promises to transform and strengthen the impact of empirical research, particularly in the social sciences.

Proponents of web-based research have recounted numerous advantages of online studies (Gosling & Mason, 2015; Gosling, Vazire, Srivastava, & John, 2004; Honing & Ladinig, 2008; Musch & Reips, 2000; Reips, 2002), while others have raised caution about inherent challenges primarily relating to the limited control of the experimental setting in which the research takes place (Reips, 2002). Table 1 summarizes the advantages and disadvantages of web-based research canvassed in this debate.

Fortunately, many of the challenges of web-based research outlined in Table 1 can be prevented or have their effects minimized. For instance, multiple submissions may be detected by collecting limited identifying information from participants. Participant dropout can be minimized by providing personal feedback or financial incentives upon study completion. Potential for misunderstanding instructions can be reduced by piloting the online tool and using suggestions from testers to improve the wording and layout of its design (see Reips, 2002, for more suggested solutions). In addition, a number of psychological research studies have reported that results are comparable from web- and lab-based settings (for a summary, see Krantz & Dalal, 2000). This suggests that, in general, the internal validity of web-based research may not be greatly compromised.

Compared to web-based research that has straightforward implementation (such as answering an online survey), web-based music research has only grown in the past decade following advances in web audio technology. In 2008, Honing and Ladinig noted the potential of web-

Music Perception, volume 38, issue 4, pp. 386–405, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2021.38.4.386

TABLE 1. *Advantages and Disadvantages of Web-based Research*

Advantages of web-based research	Disadvantages of web-based research
Access to large, diverse pools of participants, as opposed to convenience samples of university students common in laboratory-based research → good population validity.	Potential participant's fraud (e.g., multiple submissions) and non-serious responses.
High voluntary participation from individuals who are likely to be intrinsically motivated.	Possibility of participant dropout that necessitates keeping the length of the online program (and the amount of data collected) brief.
Access to specific populations that may be challenging to recruit otherwise.	Possible exclusion of participants who do not have the necessary technical knowledge, specialized equipment/ software, or internet access to run the online experiment, introducing bias in the sample. Possibility of other unrecognized or unmeasurable sampling bias.
Reduced demand on time, space, and human resources and the associated costs to run experiments (e.g., numerous participants can take part simultaneously regardless of location as long as they have internet access and the server has the capacity to handle the web traffic).	Development of a new technical skill set for researchers in creating psychometrically robust online tools and ensuring secure storage of data.
High degree of standardization of the experimental procedure built into the online program.	End user technical variance (e.g., hardware and software used, internet speed), leading to variable delivery of stimuli and increased possibility of missing data due to internet or computer problems.*
Reduced experimenter bias. The absence of the experimenter ensures that participants' responses will not be influenced by variability in the experimenter's behavior towards different participants.	Instructions and stimuli may not be understood or perceived as intended, producing inaccurate or imprecise responses.*
Good ecological validity by bringing the experiment to the participant and enabling participants to complete tasks in a familiar, natural and less stressful environment, eliciting more authentic behavior.	Variation in the participant's environment, introducing noise into the data.*
Good external validity (as a result of greater population and ecological validity), ensuring generalizability of the findings.	Reduced data quality due to reasons marked with an asterisk above.
High statistical power.	Potential threats to internal validity due to reasons marked with an asterisk above.

based music perception and cognition research and encouraged researchers to harness this rich resource. As anticipated, concerns were raised about the internal validity and reliability of online music perception tests, particularly due to potential variability in end-user audio settings and thus, the testing environment (Kendall, 2008; Pfeifer & Hamann, 2015). However, others have noted that technical and environmental variability, if occurring randomly within conditions, only leads to greater error variance, making detection of a significant effect more difficult (i.e., reducing study power) without threatening internal validity (Lacherez, 2008). Importantly also, if a significant effect is detected despite increased real-world and technical variance, the case for the effect and the finding's generalizability to real-world settings is strengthened (Honing & Reips, 2008). Moreover, accessing larger samples from the internet can serve to increase statistical power that can, in turn, offset the increased error variance (Lacherez, 2008).

Web-based music research has since gained considerable traction, as evidenced by large-scale surveys on musicality (Müllensiefen, Gingras, Musil, & Stewart, 2014), musical preferences (Greenberg et al., 2016), music engagement (Cogo-Moreira & Lamont, 2017),

and music perception (Peretz & Vuvan, 2017; Ullén, Mosing, Holm, Eriksson, & Madison, 2014). However, only a handful of studies have examined the psychometric properties and validity of online music perception tests. While some have reported robust music perception data (Ullén et al., 2014; Zentner & Strauss, 2017), others have challenged the validity and reliability of these tasks (Harrison & Müllensiefen, 2018; Pfeifer & Hamann, 2015). For instance, Pfeifer and Hamann (2015) found that participants who completed a web-based version of the Montreal Battery for Evaluation of Amusia (MBEA) scored significantly lower on all but the rhythm subtests, which might have led to a greater number of participants being diagnosed with congenital amusia. Similarly, Harrison and Müllensiefen (2018) observed considerably lower test-retest reliability from participants who undertook a beat perception test online as compared to those who took the same test in a laboratory-based setting. These findings suggest that music perception research may not be well-suited to online testing perhaps because a quiet, controlled environment cannot be guaranteed in an online setting.

In this study, we aimed to extend the comparison of web-based and laboratory-based assessments by investigating the validity and reliability of an objective online singing tool. To date, there has been a dearth of studies examining this issue (Pfordresher & Demorest, 2020a, 2020b), which is unsurprising given that recording sound, such as singing "on the fly" directly from an internet browser, has only become possible in recent years. Although the online singing tool described in our paper shares some task similarities with one previous tool, the Seattle Singing Accuracy Protocol (SSAP; Demorest et al., 2015), there were a number of differences between the two. First, our online singing tool includes a perceptual pitch-matching task that is an innovative online adaptation of an existing pitch perception task developed by Hutchins and Peretz (2012, see Materials and Procedure for more information). Moreover, in its current form as an HTML5 web application, our online singing tool differs from the SSAP in terms of its ability to support mobile devices (see Author Note for more information).

Specifically, in this study we examined the robustness of our purpose-built online singing assessment tool by comparing the psychometric properties of this tool completed in a web-based or controlled laboratory environment. Our main research questions were: (1) What is the reliability of the online tool? (2) What is the convergent validity of the online tool? (3) How does the singing task performance obtained from the online tool compare with previously published experimental data using similar singing tasks and target populations?

The reliability of the online tool was examined by: (1) comparing the internal consistency of the singing tasks undertaken in both environments, and (2) computing the test-retest reliability of the tool in a subset of web-based participants. To test the convergent validity of the tool, we compared singing performance data collected in the web-based and laboratory environment, as well as with self-reported singing ability. Convergent validity is established when similar results are obtained using two different methods (Krantz & Dalal, 2000). Therefore, if the online tool measures singing ability robustly, task performance in web-based and lab-based environments should converge. Similarly, performance of the singing tasks might be expected to correlate well with self-reported singing ability, as previous studies have reported links between self-concept of singing or music abilities and actual singing accuracy (Demorest, Kelley, & Pfordresher, 2017; Wise & Sloboda, 2008, but see also Pfordresher & Brown, 2007; Pfordresher & Demorest, 2020b). This approach concords with the two methods of establishing the validity of web-based research proposed by Krantz and Dalal (2000). Finally, we compared our web-based singing data with previously published findings for similar laboratory-based singing tasks by conducting a systematic literature search and meta-analyses.

Method

PARTICIPANTS

The participant group comprised two samples: a large web-based sample ($n = 285$) and a laboratory-based sample ($n = 52$).

The web-based sample was recruited as part of a larger twin study investigating the genetic basis of singing ability. The majority of the web-based sample ($n = 245$) was recruited from Twins Research Australia, a volunteer twin registry that helps to facilitate and support medical and scientific studies involving twins. Study information was sent in an invitation email containing a web link to the study, which interested twins could click to commence participation. The remaining twins in the web-based sample were recruited via advertisement within the University of Melbourne or through personal contacts.

The laboratory-based sample comprised first-year University of Melbourne psychology students who opted to participate in the current study from a selection of research projects within the Melbourne School of Psychological Sciences' Research Experience Program. They were awarded 1% course credit for one hour of research

participation upon completion of the study. The laboratory-based participants completed the online singing tool under laboratory-based experimental conditions. Specifically, each participant completed the assessment individually in a sound-attenuated chamber of the Clinical and Music Neuroscience Lab in the Melbourne School of Psychological Sciences, using a laptop fitted with an external computer microphone. Neither sample was selected on the basis of having music training.

Sixty-six participants from the web-based sample redid the online study four years later to determine test-retest reliability (mean test-retest interval = 4.45 years). Both the web-based and laboratory-based studies were approved by the Human Research Ethics Committee of the University of Melbourne and all participants provided informed consent.

Demographic information of the web-based, laboratory-based and test-retest web-based samples are summarized in Table 2. One-sample *t*-tests comparing the test-retest web-based sample (based on data collected at time 1) with the entire web-based sample revealed that the test-retest web-based sample had a greater proportion of female participants, $t(65) = 3.85$, $p < .001$ (medium effect size $r = .43$), and was significantly older than the web-based sample, $t(65) = 2.53$, $p = .014$ (medium effect size $r = .30$). The samples however, were not significantly different in terms of years of music training. $t(65) = 0.19$, $p = .85$). A chi-square test of independence revealed that there was a greater proportion of female participants in the laboratory-based sample compared to the web-based sample, $\chi^2(1) = 6.17$, $p = .013$. Independent *t*-tests also revealed that both age and years of music training were significantly different between these two samples, with the web-based sample being significantly older, $t(212.27) = 13.32$, $p < .01$ (large effect size $r = .67$), and having more years of music training, $t(106.11) = 3.56$, $p < .01$ (moderate effect size $r = .33$). In view of this, we controlled for the effects of sex, age, and years of music training where relevant in our subsequent analyses.

MATERIALS AND PROCEDURE

The online singing tool was developed in Adobe Flash and could be run on Windows or Macintosh platforms. Prior to recruitment, the program was extensively piloted for four weeks by 13 volunteer pilot testers and based on their feedback, program bugs, user interface design, and task instructions were rectified and improved. The online singing tool comprised three singing tasks (*Happy Birthday*, *Sing The Note*, *Sing The Tune*), two music perception tasks (*Match The Note* and *Happy Birthday Melody Recognition Task*) and a questionnaire on music and singing experience (Table 3; see Supplementary Materials accompanying the online version of this paper at online.ucpress.edu/mp for screenshots of the tasks and the questionnaire items). The online singing tasks were similar to those typically used in comprehensive singing assessments, such as the Singing Performance Battery (Berkowska & Dalla Bella, 2013) and the Seattle Singing Ability Protocol (Demorest et al., 2015). *Match The Note* was adapted from the perceptual pitch-matching task of Hutchins and Peretz (2012). Instead of matching synthesized voice stimuli with a physical pitch slider, we modified the task such that the pitch stimuli were sine waves and participants matched the stimuli by moving a sine wave pitch slider on the screen.

The estimated time for completion of the study was approximately 30 minutes, with the order of tasks and estimated duration of each shown in Table 3. At the end of the study, feedback on *Match The Note* performance was provided. In addition, participants could opt to receive feedback on how well they performed the singing tasks after the pitch analysis of the singing tasks had been conducted off-line. While it would have been ideal for participants to receive feedback on their singing performance immediately upon completion of the study (which, incidentally, is one of SSAP's features for simple pitch-matching and pitch imitation), off-line pitch analysis was deemed a necessary trade-off particularly for more complex singing tasks, to ensure faster and smoother user experience regardless of user device processing speed and internet bandwidth.

TABLE 2. *Demographic Information of the Web-based and Laboratory-based Samples*

	Web-based sample		Laboratory-based sample
	All	Test-retest	
n	285	66	52
Female (%)	220 (77.2%)	60 (90.9%)*	48 (92.3%)*
Mean age (*SD*)	33.01 (12.01)	37.08 (13.05)*	20.31 (4.47)**
Mean years of music training (*SD*)	4.62 (4.13)	4.70 (3.36)	3.02 (2.63)**

*$p < .05$; **$p < .01$

TABLE 3. *Order of Components of the Online Singing Tool*

Component	Duration (min)	Description
Welcome and instructions	1.5	Brief description of the assessment and instructions for participants before commencing the tasks (which indicated consent). Participation in a quiet setting is strongly advised.
Sound and microphone check		Tests the functionality of speakers and microphone of the participant's computer/device to ensure recording accuracy.
Demographic information		Enter details of demographical information (name, date of birth, sex, level of education, first and second languages)
Happy Birthday	3.5	Sing *Happy Birthday* in several different conditions (e.g. paced/unpaced, with/without lyrics), in a key of choice.
Match The Note	5	Match five different sine tones (pitch classes: B, C#, D#, F, G) to a probe tone by moving a sine wave pitch slider on the screen (15 trials; tone range for females = 246.94–397 Hz; tone range for males=123.47–196 Hz)
Sing The Note	5	Sing back five sine tones (same stimuli as above) (15 trials; tone range for females =246.94-397 Hz; tone range for males =123.47–196 Hz)
Sing The Tune	5	Sing back five different 7-note piano tunes* (15 trials; tone range for females =220–440 Hz; tone range for males =110–220 Hz)
*Happy Birthday Melody Recognition Task***	2	Identify the correct version of *Happy Birthday* from two mistuned versions.
Music questionnaire	5	Items assessing singing and music background.
Concluding remarks		Option to leave expression of interest for participation in future studies. Feedback on performance in *Match The Note* provided.
Thank-you page		
Total duration	**27**	

*See Figure S6 in Supplementary Materials at online.ucpress.edu/mp. **This task is not discussed in the paper because 99% of the participants performed at ceiling.

MEASURES

Singing data were downloaded from a secure online FTP server and processed using *Tony*, an open-source pitch transcription software (Mauch et al., 2015). The software automatically determines stable pitch segments in the participant's singing and estimates the median fundamental frequency of each pitch segment using the pYin algorithm (Mauch & Dixon, 2014) and Viterbi-decoded hidden Markov model. The estimated pitch segments were then inspected visually and aurally within the *Tony* graphical user interface and manually adjusted where necessary (e.g., merging or splitting pitch segments if the number of estimated segments was more or less than the expected number of pitch segments, respectively). The reliability of this automatic pitch segmentation method was high (intraclass correlation = .999), evaluated by comparing the pitch segmentation outcome in *Tony* with a manual pitch segmentation method we performed using the speech analysis software, *Praat* (version 5.4.01; Boersma & Weenink, 2014) on the sound data of 20 randomly selected web-based participants.

In the two single pitch-imitation tasks (*Match The Note*, *Sing The Note*), accuracy was assessed by computing the absolute difference (in cents) between each sung pitch and the corresponding expected pitch before summing and averaging across the entire task. This measure (herein referred to as Pitch Deviation, PD) was

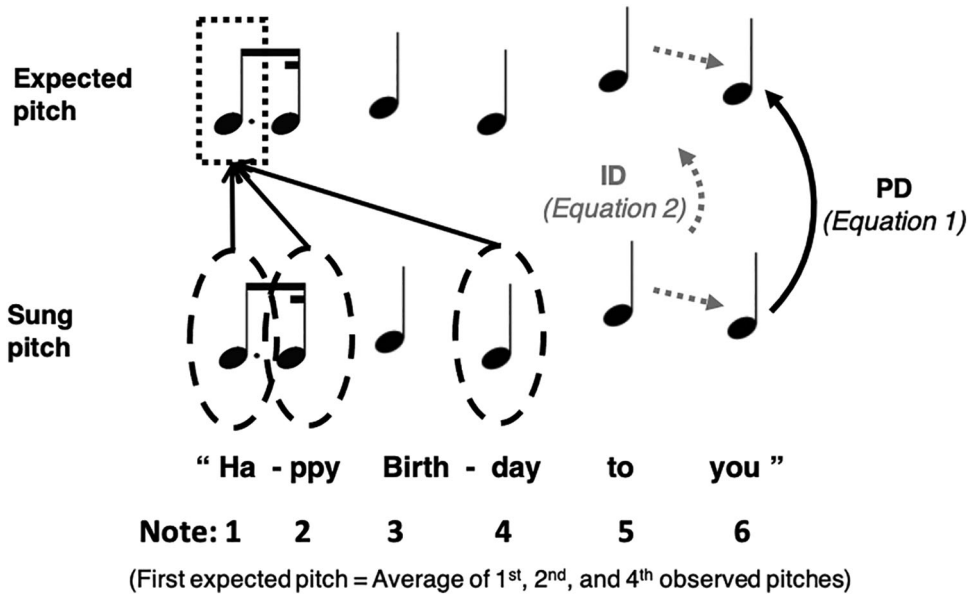

FIGURE 1. Estimation of pitch deviation (PD) and interval deviation (ID) in the first phrase of *Happy Birthday*. Fundamental frequencies of the 1st, 2nd, and 4th pitches sung by the participant were averaged to form the first expected pitch, from which other pitches were computed using the known intervals of the song. PD was computed using Equation 1 and ID using Equation 2.

computed with the following equation:

$$PD = \frac{\sum_{i=1}^{n} |1200 \times \log_2 \frac{f_i}{r_i}|}{n} \quad (1)$$

where f_i is the fundamental frequency (in Hz) of the i^{th} pitch produced by the participant, r_i is the fundamental frequency of the corresponding reference pitch, and n is the total number of pitches in a trial or a task. The smaller the PD, the better the pitch accuracy.

For *Sing The Tune*, an additional accuracy measure was computed from the absolute difference between each sung interval (i.e., distance between two adjacent pitches) and the corresponding expected interval, averaged across the entire task. This measure (herein referred to as Interval Deviation, ID) was computed with the following equation:

$$ID = \frac{\sum_{i=2}^{n} |\left(1200 \times \log_2 \frac{f_i}{f_{i-1}}\right) - \left(1200 \times \log_2 \frac{r_i}{r_{i-1}}\right)|}{n} \quad (2)$$

where f_i is the fundamental frequency (in Hz) of the i^{th} pitch produced by the participant, r_i is the fundamental frequency of the corresponding reference pitch of the stimulus, and n is the total number of pitches in the stimulus of the task. Both PD and ID have been commonly used in previous studies as indicators of singing proficiency (Dalla Bella, 2015).

Finally, in *Happy Birthday*, ID was similarly computed using the above equation. As for PD, the expected pitches of the task were estimated as follows: the first expected pitch was estimated by averaging the first, second and fourth observed pitches from the first phrase of *Happy Birthday* sung by the participant (Figure 1). As these three notes have the same expected pitch height, their average fundamental frequency yields a more stable estimate of the expected starting pitch than using the first sung note alone. Then, using the known intervallic information of *Happy Birthday*, the subsequent expected pitches of the melody were derived using the following equation:

$$r_i = 2^{\frac{int}{1200}} \times r_{i-1} \quad (3)$$

where r_i is the fundamental frequency (in Hz) of the expected pitch in the i^{th} position of the song, r_{i-1} is the fundamental frequency of the preceding expected pitch, and *int* is the interval (in cents) between the current and previous expected pitches (see Figure 1). Only the PD and ID of the final two trials (singing at a prescribed tempo of 120 bpm without lyrics) were computed and averaged, as previous research has found that singing on a neutral syllable at a prescribed tempo elicits the best singing accuracy (Berkowska & Dalla Bella, 2013; Dalla Bella, Giguère, & Peretz, 2007).

SYSTEMATIC LITERATURE REVIEW

To identify previous studies using similar singing tasks and accuracy measures in similar populations (i.e., the general public or university students), a literature search was conducted in Scopus using (1) "singing ability" or "singing proficiency," (2) "accura*" and (3) "pitch" as search terms, applied to publication titles, abstracts, and keywords. This produced 29 results, of which 23 were excluded using the following exclusion criteria:

A. The study did not have participants (experimental or control) from the general population (e.g., the study sample only included expert musicians or

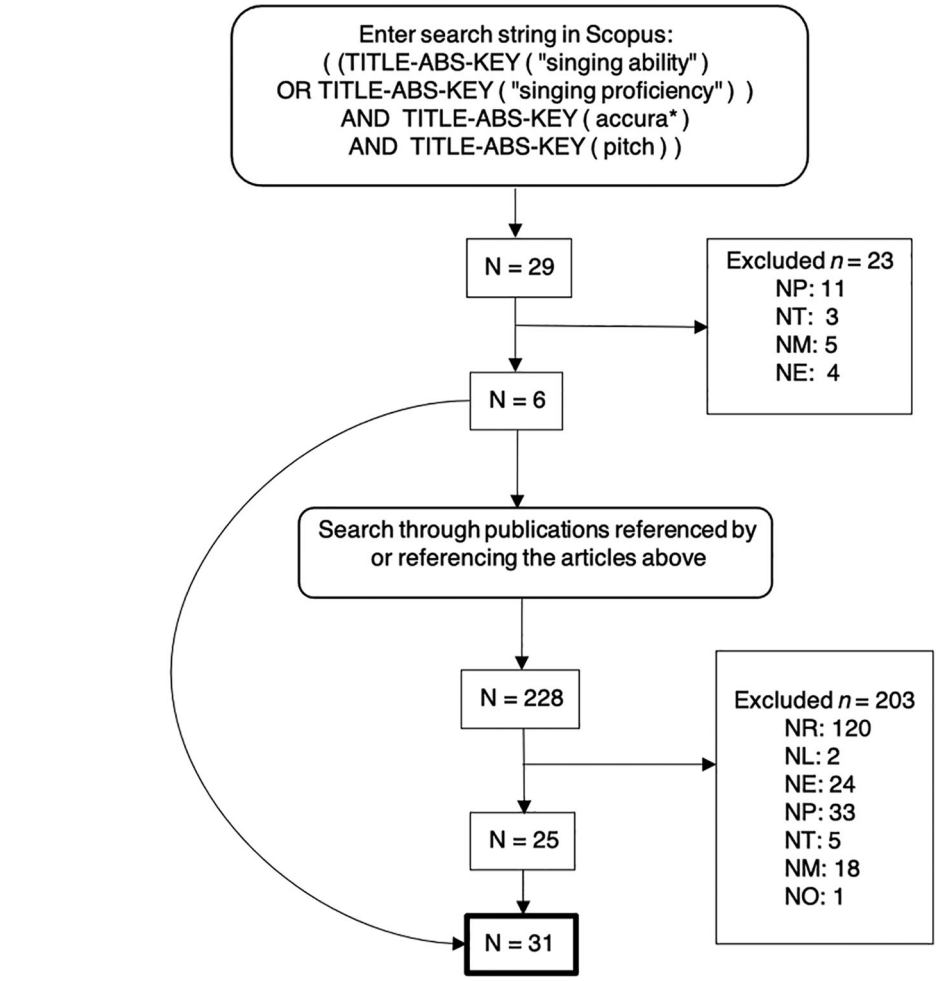

FIGURE 2. Flowchart highlighting the study selection process.

singers, individuals with congenital amusia or absolute pitch, or clinical populations)

B. The study did not use similar singing tasks to those in the online singing tool.
C. The study did not use objective measures that were comparable to those used in the current study to assess singing ability, or the study did not sufficiently report means and standard deviations of the objective measures to allow comparison. To minimize the number of suitable studies excluded by this criterion, for the latter scenario good-faith estimates were made from data in plots or tables where possible. For studies that were published fairly recently (i.e., in the past decade) where the required information was not available from the published content, requests for more information were made via email communication to the corresponding authors.
D. The study was not an empirical study (e.g., theoretical paper).

Of the 6 remaining articles, 228 distinct publications referenced by or referencing these articles were manually canvassed to identify relevant studies. In this step, three more conditions were added to the aforementioned exclusion criteria:

E. The study was not related to singing ability research.
F. The study was not published in English.
G. The study had a highly overlapping participant sample with another study.

After excluding 203 publications, 31 relevant studies remained. Figure 2 presents a flowchart of how relevant studies were selected and Tables 4 to 6 present the characteristics of the relevant studies as organized by different singing tasks (single pitch-matching, novel tune imitation and familiar song singing, respectively).

For each selected study, if there was more than one participant sample or more than one condition for a singing task, the study sample or the singing condition that most closely matched the current study's web-based sample or condition was chosen. For studies that comprised samples of both expert musicians and non-musician controls, we reported the accuracies of the two groups separately in Tables 4 to 6 for completeness.

ANALYSES

Unless otherwise stated, analyses were conducted using SPSS statistical package version 25 (IBM Corp., 2010), with an alpha criterion of .05. As the skewness and kurtosis values for the task measures were less than -1 or greater than +1 and thus indicated non-normality (George & Mallery, 2010), all variables were transformed using the natural log (ln) transformation, such that the skewness and kurtosis values were within ± 1. All subsequent references to the task variables refer to the transformed variables, although for interpretability, the mean and median values are presented in their raw form in Table 7.

To address the first research question, the internal consistency of the online singing tool in each sample was computed using Cronbach's alpha, evaluating the extent to which performance on individual tasks (e.g., "Sing the Note PD") predicted the participant's performance on other tasks (e.g., "Happy Birthday ID"). Test-retest reliabilities of each singing task in the online tool were computed following the guidelines from Koo and Li (2016). Specifically, intraclass correlation estimates were calculated based on a single-measurement, absolute-agreement, two-way mixed-effects model. Single measurement was selected because for each task, performance from the first attempt was used as the basis of the assessment rather than the mean performance of the test and retest. In addition, absolute agreement was selected rather than consistency because we were concerned with the extent to which the second attempt is similar to the first attempt in absolute terms, rather than in a correlational manner. Finally, a two-way mixed effects model was chosen instead of a two-way random-effects model because repeated attempts using the same test battery imply that the measurements are not randomized and thus the reliability results cannot be generalized to other test batteries with similar characteristics.

For the second research question, the singing accuracy of the web-based and laboratory samples was compared for each singing task using a general linear model (GLM) approach. Each task variable was entered as the dependent variable, group (web-based vs. laboratory-based) as the fixed factor and age, sex, and years of music training as covariates. Interaction terms between each covariate and the fixed factor were also entered into the model to explore potential interactions between group and covariates. In addition, correlations between the task measures and self-reported singing ability (in 7-point Likert scale; see Question 20 from Table S1 in Supplementary Materials at online.ucpress.edu/mp) were computed to investigate the relationship between self-assessed and objective measures of singing ability.

For the third research question, to compare the task performances of the web-based sample with relevant studies canvassed from the literature search, random-effects meta-analyses (DerSimonian Laird method)

TABLE 4. *Description of Previous Studies Using Comparable Populations and Accuracy Measures for a Single Pitch-matching Task.*

Study	*N*	Sample description	Mean age (range)	Stimuli	Pitch Deviation [*M* (*SD*) cents]		
					Musicians	Nonmusicians	All
Murry (1990)	10	5 experienced singers (mean professional experience = 6 years); 5 nonmusicians	27.4 (25–33)	Sine tones	102.00 (46.00)[b]	326.00 (259.00)[b]	214.00 (211.40)[a]
Amir, Amir, & Kishon-Rabin (2003)	26	13 musicians (mean years of playing =13) and 13 nonmusicians	26 (20–34)	Sine tones	46.40 (52.90)[b]	193.20 (257.30)[b]	119.79 (196.80)[a]
He & Zhang (2017)	32	University poor-pitch singers with intact pitch perception (mostly nonmusicians)	23.3 (18–31)	Sine tones	-	-	184.00 (116.00)
Watts & Hall (2008)	19	University female students (nonmusicians)	20–35	Violin	-	-	98.00 (28.00)[c]
Watts, Moore, & McCaghren (2005)	21	University students (nonmusicians)	19–35	Synthesized piano tones	-	-	159.50 (102.30)[a]
Estis, Dean-Claytor, Moore, & Rowell (2011)	40	20 trained singers ($\geq$ 3 years formal singing training); 20 controls (no singing training)	23.1 (19–32)	Synthesized piano tones	12.50 (5.70)	168.40 (172.00)	90.42 (143.71)[a]
Greenspon & Pfordresher (2019)	216	University students (not selected for music training)	NA	Piano tones	-	-	102.18 (132.10)[d]
Moore, Keaton, & Watts (2007)	30	University students (nonmusicians)	20–30	Synthesized complex tones	-	-	112.13 (128.62)[b]
Moore, Estis, Gordon-Hickey, & Watts (2008)	20	University students (no singing training)	20–30	Synthesized non-vocal complex tones	-	-	70.65 (115.84)[b]
Estis, Coblentz, & Moore (2009)	32	10 trained singers (mean training = 6.15 years); 22 untrained singers	23.1 (20–30)	Synthesized complex tones	16.00 (6.00)[c]	47.20 (52.60)[c]	37.44 (45.86)[a]
Lévêque, Giovanni, & Schön (2011)	32	General public (18 self-reported poor singers)	21-51	Synthesized voice-like complex tones	-	-	72.75 (87.91)[a,c]
Berkowska & Dalla Bella (2013)	50	University students (mostly nonmusicians)	25.1 (19–39)	Synthesized voice-like complex tones	-	-	140.20 (143.40)
Pfordresher & Brown (2007)	78	University students (nonmusicians)	NA	Vocaloid	-	-	63.32 (107.56)[d]

(continued)

TABLE 4. *(continued)*

Study	*N*	Sample description	Mean age (range)	Stimuli	Pitch Deviation [*M* (*SD*) cents] Musicians	Nonmusicians	All
Pfordresher & Halpern (2013)	136	University students (not selected for music training)	19 (18–27)	Vocaloid	-	-	176.72 (122.41)[d]
Wise & Sloboda (2008)	29	University students (12 self-reported "tone-deafness"); mean years of music involvement = 3.5(4.3)	19.3 (18–24)	Voice model	-	-	24.14 (16.48)[c]
Zarate, Delhommeau, Wood, & Zatorre (2010)	19	University students (nonmusicians)	22 (17.6–26.4)	Voice model	-	-	30.34 (17.74)[a]
Hutchins, Larrouy-Maestri, & Peretz (2014)	22	University students (nonmusicians)	23 (18–30)	Participant's own voice	-	-	38.00 (4.83)

[a] Computed using formulae (e.g., for combining two or more groups, or to estimate *SD* from *SE* or from range). [b] Computed from raw data. [c] Good-faith estimates from plots. [d] Obtained through personal communication with study authors.

were performed using Open Meta-Analyst (Wallace, Schmid, Lau, & Trikalinos, 2009), conducting one analysis for each task variable. A random-effects model was chosen over fixed-effects because this takes into account both within-study sampling error and between-study variance to explain the overall variability in study results. This approach was appropriate here as there was evidence of heterogeneity among studies that could not be readily explained by methodological or participant sample differences (Deeks et al., 2011).

For those studies that comprised samples of both expert musicians and nonmusician controls, the samples were combined to obtain a pooled estimate of the mean and standard deviation of task performance for the subsequent meta-analysis. Combining the samples served to maximize the comparability with the current study's web-based sample, which comprised participants with music training and nonmusicians. It is also the case that studies varied considerably in their definition of musicianship, thereby precluding reliable comparison of musicians and nonmusicians in the meta-analyses. All study results were extracted and converted to the desired format (e.g., deriving pooled estimates or converting from standard error to standard deviation) following the guidelines from the *Cochrane Handbook for Systematic Reviews of Intervention* (Higgins & Green, 2011).

Thus, the overall estimated mean and standard deviation derived from each meta-analysis reflect how well a singing task was generally performed regardless of variation in participant pool and task conditions. These values were then compared with the mean and standard deviation of the corresponding task variable from the current online study using an independent *t*-test with summary statistics in R (via the tsum.test function from the BSDA package).

Results

INTERNAL CONSISTENCY AND TEST-RETEST RELIABILITY OF THE ONLINE TOOL

Overall, the internal consistency of the online tool for the web-based and laboratory-based samples was high, with Cronbach's $\alpha = .90$ and $\alpha = .86$, respectively. When the internal consistency of the singing tasks was considered separately, Cronbach's α for the web- and laboratory-based samples were .92 and .83, respectively. These values demonstrate that irrespective of the sample, there was high internal consistency across all singing tasks, even when a music perception task (*Match The Note*) was included.

The α values were corroborated by the significant Pearson correlation coefficients found across all task measures (Table 8), with moderate to large positive effect sizes in both web-based and laboratory-based samples ($r = .33$–$.96$). The largest correlations occurred between the PD and ID scores for a given task, which were similar across the two samples ($r = .88$–$.89$ for *Happy Birthday* and $r = .94$–$.96$ for *Sing The Tune*). In addition, the correlations between the perceptual and vocal single-pitch matching tasks (i.e., *Match The Note* and *Sing The Note*) were also similar for the two samples ($r = .50$ and $.52$).

TABLE 5. *Description of Previous Studies Using Comparable Populations and Accuracy Measures for Tune Imitation Tasks*

Study	*N*	Sample description	Mean age (range)	Stimuli	Pitch Deviation *M* (*SD*) cents	Interval Deviation *M* (*SD*) cents
Price (2000)	141	University students (non-music majors)	NA	Two-note stimuli (voice and sine tones)	37.41 (45.00)	-
Granot, Israel-Kolatt, Gilboa, & Kolatt (2013)	18	Nonmusicians who question their singing ability	34.6 (17–55)	Two-note piano stimuli	184.40 (185.39)[a]	172.90 (155.90)[a]
Yang et al. (2014)	12	University students (nonmusicians)	22.5 (19–26)	Three-note piano stimuli	-	140.00 (70.00)
Greenspon, Pfordresher, & Halpern (2017)	20	Mostly university students with a mean of 3.1 years of music training	20.8 (18–30)	Three- to four-note vocaloid stimuli	109.90 (106.60)[a]	-
Pfordresher and Brown (2007)	78	University students (nonmusicians)	NA	Four-note vocaloid stimuli	122.24 (87.72)[d]	141.07 (75.34)[d]
Pfordresher & Brown (2009)	24	University students with little or no music training	20 (17–39)	Four-note vocaloid stimuli	107.52 (29.49)[a]	-
Greenspon & Pfordresher (2019)	216	University students (Not selected for music training)	NA	Four-note voice stimuli	84.31 (93.32)[d]	-
Pfordresher, Brown, Meier, Belyk, & Liotti (2010)	45	University students (mostly nonmusicians)	20.5 (17–31)	Five-note vocaloid stimuli	54.80 (87.88)[a]	-
Belyk, Johnson, & Kot (2018)	34	Not selected for music training (2-15 years of formal music training)	21 (18–29)	Five-note vocaloid stimuli	166.16 (156.67)[b]	128.06 (85.67)[b]
Wise & Sloboda (2008)	29	University students (12 self-reported "tone-deafness"); mean years of music involvement = 3.5(4.3)	19.3 (18–24)	Five-note voice stimuli	55.09 (43.22)[c]	-
Zarate et al. (2010)	19	University students (nonmusicians)	22 (17.6–26.4)	Five-note voice stimuli	74.04 (65.75)[a]	44.76 (31.19)[a]
Berkowska & Dalla Bella (2013)	50	University students (mostly nonmusicians)	25.1 (19–39)	Six-note voice-like complex tone stimuli	126.40 (145.10)	-

[a] Computed using formulas (e.g., for combining two or more groups, or to estimate *SD* from *SE* or from range). [b] Computed from raw data. [c] Good-faith estimates from plots. [d] Obtained through personal communication with study authors.

However, while *Sing The Note* was observed to have strong correlations with the other two singing tasks in the web-based sample ($r = .64. –.74$), in the laboratory sample the correlations were moderate ($r = .33–.49$). In contrast, *Match The Note* was observed to have stronger correlations with the singing tasks in the laboratory sample ($r = .50–.63$) than in the web-based sample ($r = .40–.53$).

As shown in Table 9, the test-retest reliabilities for the 66 web-based participants who repeated the online test ranged from .65 (*Happy Birthday* PD and ID) to .80 (*Sing The Tune* PD).

TABLE 6. *Description of Previous Studies Using Comparable Populations and Accuracy Measures (ID) for Familiar Song Tasks*

Study	*N*	Sample description	Mean age (range)	Stimuli	Interval Deviation *M* (*SD*) cents
Pfordresher & Brown (2009)	46	University students with little or no musical training	21.9	*Happy Birthday* (unpaced and with lyrics)	111.57 (50.64)[a]
Hutchins, Larrouy-Maestri, & Peretz (2014)	22	University students (nonmusicians)	23 (18–30)	*Happy Birthday* (unpaced and with lyrics)	45.00 (23.08)
Larrouy-Maestri, Lévêque, Schön, Giovanni, & Morsomme (2013)	166	General public (not selected for music training)	29.9 (14–76)	*Happy Birthday* (unpaced and with lyrics)	55.97 (48.21)[d]
Larrouy-Maestri & Morsomme (2014)	63	Female (nonmusicians)	29.8 (15–75)	*Happy Birthday* (unpaced and with lyrics)	50.83 (26.11)
Erdemir & Rieser (2016)	42	University students (12 singers, 12 instrumentalists, 18 nonmusicians)	22.7 (18–29)	*Happy Birthday* (unpaced and with lyrics)	37.14 (11.96)[a,c]
Dalla Bella, Giguère, & Peretz (2007)	62	General public (not selected for music training)	35.8 (18–75)	*Gens du Pay* (unpaced and with lyrics)	80.32 (60.38)[a]
Tremblay-Champoux, Dalla Bella, Phillips-Silver, Lebrun, & Peretz (2010)	21	11 Controls for congenital amusia participants and 10 French University exchange students (mean years of training = 1.31 years)	44.8	*Gens du Pay* (unpaced and with lyrics)	103.33 (51.01)[a]
Dalla Bella, Giguère, & Peretz (2009)	11	Control group for congenital amusia participants	56	*Gens du Pay* (neutral syllables and unpaced)	50.00 (14.72)[a]
Berkowska & Dalla Bella (2013)	50	University students (mostly nonmusicians)	25.1 (19–39)	Three familiar songs: *Brother John, Jingle Bells* and *Stolat* (neutral syllables and paced)	44.20 (19.70)
Dai, Mauch, & Dixon (2015)	39	University students (mostly amateur musicians or singers from university's music society)	23.3 (20–27)	Three familiar songs from *The Sound of Music: Edelweiss, Do-Re-Mi and My Favourite Things* (neutral syllables and paced)	34.00 (46.00)

[a] Computed using formulas (e.g., for combining two or more groups, or to estimate *SD* from *SE* or from range). [b] Computed from raw data. [c] Good-faith estimates from plots. [d] Obtained through personal communication with study authors.

TASK PERFORMANCE AND CONVERGENT VALIDITY OF THE WEB- AND LABORATORY-BASED SAMPLES

General linear models revealed no main effect of sample for *Happy Birthday* PD, $F(1, 292) = 2.87$, $p = .09$, *Happy Birthday* ID, $F(1, 292) = 2.66$, $p = .10$, and *Match The Note* PD, $F(1, 298) = 0.26$, $p = .61$. In contrast, a significant main effect of group was found for the singing imitation tasks: *Sing The Note* PD, $F(1, 293) = 4.26$, $p = .04$, *Sing The Tune* PD, $F(1, 293) = 6.54$, $p = .01$, and *Sing The Tune* ID, $F(1, 293) = 3.99$, $p = .047$, with the laboratory-based sample performing significantly better than the web-based sample, albeit with small effect sizes (partial $\eta^2 = .01$–$.02$, see Tables S3 to S5 in Supplementary Materials at online.ucpress.edu/mp). Although significant group effects were found, Levene's tests showed that the assumption of equality of residual variances between the groups was violated for *Sing The Note* PD, $F(1, 299) = 4.38$, $p = .037$, and *Sing The Tune* PD, $F(1, 299) = 4.01$, $p = .046$, hence the significant findings warranted further investigation. Overall, no significant interaction effects were observed in any of the general linear

TABLE 7. *Performance of the Web-based and Laboratory-based Samples (in Cents) for the Singing and Music Tasks Prior to Transformation*

Task measure	Sample	Mean (SD)	Median	Range	Skewness	Kurtosis
MatchTheNote PD	Web	36.62 (52.89)	14.52	1.87–487.30	3.71	21.56
	Lab	33.94 (48.57)	14.70	2.64–220.22	2.72	7.18
SingTheNote PD	Web	86.31 (116.61)	28.52	5.88–504.28	1.77	2.09
	Lab	63.29 (94.64)	24.81	8.75–424.46	2.59	6.76
SingTheTune PD	Web	103.41 (121.95)	56.06	9.92–715.29	2.55	7.39
	Lab	58.39 (55.13)	39.65	13.75–323.54	2.94	11.08
SingTheTune ID	Web	97.33 (80.02)	69.06	10.75–519.23	1.43	2.43
	Lab	67.12 (54.26)	48.38	13.99–238.34	1.44	1.53
HappyBirthday PD	Web	68.78 (67.08)	41.31	10.84–591.91	2.94	14.33
	Lab	56.86 (48.42)	37.57	15.26–222.92	2.06	4.22
HappyBirthday ID	Web	52.57 (39.92)	41.02	12.16–366.59	3.84	24.03
	Lab	48.92 (30.34)	37.06	14.03–127.46	1.33	0.95

models, but main effects of years of music training, sex, or age were present in one or more of the models (see Tables S2 to S7 in Supplementary Materials at online.ucpress.edu/mp).

As a post hoc analysis, we conducted propensity score matching (PSM) analysis to generate two subsets of web-based and laboratory-based samples that were well-matched on the three covariates (i.e., age, sex, years of music training) to facilitate unbiased comparison between the two samples. After propensity scores were estimated using logistic regression, a 1:1 nearest-neighbor matching algorithm was employed to match a laboratory-based participant with a web-based participant with the most similar estimated propensity score, using a caliper (the maximum allowable difference between matched participants) of 0.25 *SD*. The PSM analysis was performed using Propensity Score Matching for SPSS (version 3.0.4), an SPSS R-menu developed by Thoemmes (2012). The matching process resulted in two subsamples each with 41 participants and dependent *t*-tests revealed that they did not differ significantly in age and years of music training (see Table S8 in Supplementary Materials at online.ucpress.edu/mp). A chi-square test of independence also confirmed that there was no significant difference in the proportions of males (9.8%) and females (90.2%) in the two subsamples, $\chi^2(1) = .00$, $p = 1.00$. This demonstrates the suitability of using dependent *t*-tests to compare the two subsamples on the imitative singing tasks. No significant differences were found between the matched participants from the two subsamples on all three task measures ($p > .05$, Table 10). Subsequent comparisons for the other singing and music tasks also showed no significant differences in task performance, confirming the broader set of findings from the general linear models ($p > .05$, Table 10).

Significant correlations between self-reported singing ability and the singing task measures were observed after partialling out the effects of sex, age and years of music training (r = -.31 to -.59, Table 11). In both samples, the highest correlation was observed between

TABLE 8. *Pearson Correlation Coefficients Between the Task Measures*

	MatchTheNote PD	SingTheNote PD	SingTheTune PD	SingTheTune ID	HappyBirthday PD	HappyBirthday ID
MatchTheNote PD		.52**	.52**	.53**	.40**	.43**
SingTheNote PD	.50**		.74**	.65**	.66**	.64**
SingTheTune PD	.63**	.49**		.94**	.75**	.74**
SingTheTune ID	.57**	.33*	.96**		.71**	.72**
HappyBirthday PD	.60**	.42**	.59**	.56**		.89**
HappyBirthday ID	.51**	.40**	.50**	.47**	.88**	

Values above the diagonal belong to the web-based sample and the values below the diagonal belong to the laboratory-based sample. *$p < .05$; **$p < .01$

TABLE 9. *Test-retest Reliability of the Singing Task Measures (n = 66).*

	Test-retest reliability (Intraclass correlation)
SingTheNote PD	.79 [.67, .87]
SingTheTune PD	.80 [.67, .88]
SingTheTune ID	.71 [.49, .83]
HappyBirthday PD	.65 [.49, .77]
HappyBirthday ID	.65 [.49, .77]

Values in brackets are the 95% confidence intervals.

self-rated singing ability and *Sing The Tune* PD ($r = .55$ and -.59). The correlation trends for the other tasks were somewhat different. In the web-based sample, the correlations between the singing task measures and self-rated singing ability tended to be high and of similar magnitude ($r = -.47$ to -.55), whereas in the laboratory-based sample the correlations were strongest for *Sing The Tune* ($r = -.59$ and -.57), followed by *Happy Birthday* ($r = -.52$ and -.39), with a relatively weaker correlation for *Sing The Note* ($r = -.31$).

Different-sized correlations were also observed between self-rated singing ability and pitch perception ability (as measured by *Match The Note*) in both samples. A low correlation ($r = -.14$) between self-rated singing ability and *Match The Note* was observed in the web-based sample, which was reasonable as one would expect the correlations between self-rated singing ability and the singing tasks to be higher. In contrast, however, the correlation between self-rated singing ability and *Match The Note* was fairly large ($r = -.48$) in the laboratory-based sample, and higher than the correlations with *Sing The Note* PD ($r = -.31$) and *Happy Birthday* ID ($r = -.39$).

COMPARING THE ONLINE SINGING TOOL TO PREVIOUS RESEARCH

The forest plots derived from the principal meta-analyses of the relevant published singing research for the single pitch-matching task (PD), tune imitation (PD

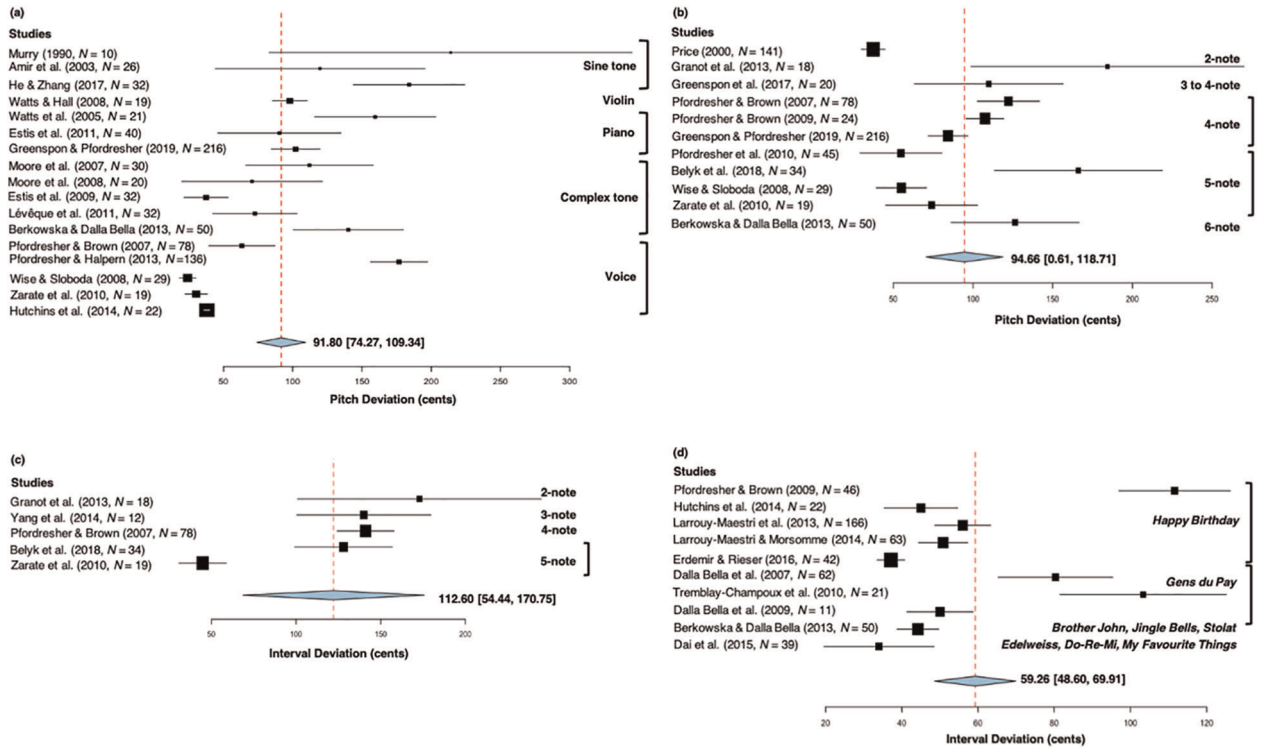

FIGURE 3. Forest plots of comparable published singing studies using similar singing task measures and populations as the current study. (a) Studies that used a single pitch matching task. (b) Studies that used tune imitation tasks (PD). (c) Studies that used tune imitation tasks (ID). (d) Studies that used familiar song singing tasks (ID). Each horizontal line on a forest plot represents an individual study with the result plotted as a box and the 95% confidence interval of the result displayed as the line. The bigger the study, the smaller the horizontal line and bigger the black box representing the point estimate. The diamond at the bottom of the forest plot shows the result when all the individual studies are combined together and averaged. The horizontal points of the diamond are the limits of the 95% confidence intervals and the vertical axis of the diamond represents the point estimate of the averaged studies.

TABLE 10. *Results of Dependent t-tests and Descriptive Statistics for Task Performances from Matched Participants From Web-based and Laboratory-based Samples*

	Web-based		Laboratory-based					
Measure	*N*	*M (SD)*	*N**	*M (SD)*	*df*	*t*	*p*	*r*
MatchTheNote PD	41	25.51 (30.71)	41	33.85 (44.65)	40	-1.05	.30	.16
SingTheNote PD	40	102.26 (130.29)	40	65.45 (101.30)	39	1.32	.20	.21
SingTheTune PD	40	110.66 (133.87)	40	62.93 (59.72)	39	1.89	.07	.29
SingTheTune ID	40	98.52 (78.78)	40	71.14 (56.27)	39	1.59	.12	.25
HappyBirthday PD	39	81.32 (76.28)	39	53.21 (45.66)	38	1.32	.19	.21
HappyBirthday ID	39	52.01 (31.48)	39	50.92 (30.16)	38	0.18	.86	.03

**N* varied from task to task due to missing data from participants as a result of incompletion or end user technical issues.

TABLE 11. *Partial Correlations Between the Task Measures and Self-rated Singing Ability (Controlled for Sex, Age, and Years of Music Training)*

	Self-rated singing ability	
	Web-based sample ($n = 285$)	Lab-based sample ($n = 52$)
MatchTheNote PD	-.14*	-.48**
SingTheNote PD	-.47**	-.31*
SingTheTune PD	-.55**	-.59**
SingTheTune ID	-.51**	-.57**
HappyBirthday PD	-.52**	-.52**
HappyBirthday ID	-.51**	-.39**

*$p < .05$; **$p < .01$

TABLE 12. *Results of Independent t-tests and Descriptive Statistics for Task Performances from Relevant Previous Research and Current Study*

	Past studies		Current study					
Measure	***N***	***M (SD)***	***N****	***M (SD)***	*df*	*t*	*p*	*r*
Single pitch matching (PD)	812	91.80 (254.96)	279	86.31 (116.61)	1008.60	-0.48	.63	.02
Tune imitation (PD)	675	94.66 (318.81)	274	103.41 (121.95)	944.50	0.61	.54	.02
Tune imitation (ID)	161	122.14 (346.05)	274	97.33 (80.02)	170.11	-0.90	.37	.07
Familiar song (ID)	522	59.26 (124.18)	279	52.57 (39.92)	693.37	-1.13	.26	.04

**N* of web-based sample varied from task to task due to missing data from participants as a result of incompletion or end user technical issues.

and ID) and familiar song singing (ID) are shown in Figure 3, panels (a) to (d). For each study represented in the forest plot, the square box shows the mean deviation (in cents) and the horizontal line through the box illustrates the length of the confidence interval. The diamond at the bottom of each forest plot shows the pooled estimate across the studies that used similar singing tasks (Tables S9 to S12 in the Supplementary Materials at online.ucpress.edu/mp contain detailed statistics of the meta-analyses).

It is evident from the forest plots that for each singing task, there was substantial heterogeneity in task performance across studies. This is corroborated by the significant chi-squared tests ($p < .01$) and the I^2 statistic (Tables S9 - S12), which describes the percentage of variability in effect estimates that is due to heterogeneity rather than sampling error (Deeks et al., 2011). The I^2 statistics of all four measures were between 94% to 97%.

When we cross-referenced the forest plots with the sample and task characteristics from Tables 4 to 6, it became evident that the observed heterogeneity did not necessarily stem from differences in participant samples or task demands. For instance, Murry (1990) and Amir, Amir, and Kishon-Rabin (2003) both used sine tones as stimuli in the single pitch-

matching task and half of their sample comprised experienced singers or expert musicians with a similar age range (Table 4). The pooled PD estimate (214 cents) of Murry (1990), however, was substantially higher than the PD estimate (119.8 cents) from Amir et al. (2003). Similarly, Pfordresher and Brown (2007), Pfordresher and Brown (2009) and Greenspon and Pfordresher (2019) assessed the singing ability of university students who were mostly nonmusicians using a tune imitation task with four-note vocaloid stimuli (Table 5). Participants from the third study (mean PD = 84.31 cents) were comparatively more accurate than the first two studies (mean PD = 122.24 cents and 107.52 cents, respectively). Intriguingly, the sample from Pfordresher et al. (2010) who imitated five-note vocaloid stimuli were much more accurate (mean PD = 54.80 cents) than the samples from the three aforementioned studies, who imitated four-note tunes. Finally, Pfordresher and Brown (2009) and Hutchins, Larrouy-Maestri, and Peretz (2014) both tested the ability of university students without music training to sing *Happy Birthday* (Table 6) and despite the similarity in demographics, music training and singing condition, the mean ID of the two studies were vastly different (111.57 cents vs. 45 cents).

Independent *t*-tests revealed no significant differences between the pooled mean estimates from the meta-analyses and those from the current study's web-based sample for the four singing tasks, $p > .05$ for all comparisons (Table 12). Furthermore, it can be observed from Table 12 that apart from tune imitation ID, the pooled mean estimates from previous studies and the corresponding means from the current study for the remaining three tasks only differed by 5–9 cents. Notably, these three pooled mean estimates were derived from considerably more studies and participants (10–17 studies; N = 522–812) than the estimate for tune imitation ID (5 studies, $N = 161$).

Discussion

The main aim of our study was to investigate whether singing ability could be robustly assessed in a web-based environment. Overall, our results provide strong evidence for the reliability and validity of our online assessment tool in testing singing and music ability beyond a controlled laboratory environment.

The high Cronbach alpha values for all of the tasks in the web-based and laboratory samples (α =.83–.92) support the internal consistency of the online singing tool. In particular, the values for the web-based sample fell within the excellent range ($\geq$ .90) and were highest when the singing tasks were considered separately (.92). Thus, the findings indicate that the singing tasks were tapping the same underlying construct (i.e., singing ability).

The test-retest reliability of the singing tasks, which ranged from .65 to .80, is similarly high. Based on the guidelines recommended by Cicchetti (1994) for evaluating psychological assessment tools, intraclass correlations below .40 are considered poor, those between .49-.59 are fair, those between .60–.74 are good, and those between .75–1.00 are excellent. The online singing task measures thus demonstrated good to excellent test-retest reliability, which is noteworthy, especially given the average timeframe between test and retest was approximately 4.5 years. The long test-retest time interval could have resulted in real changes in a participant's singing ability, for instance from improvement due to further training or deterioration due to changes in vocal, neurological or hearing conditions. Given this, the high test-retest reliability estimates provide strong evidence of the stability of performance-based singing ability over time. Moreover, the long timeframe minimized possible carryover effects due to memory or practice effects, giving further credence to the test-retest reliability of the singing tasks.

Although GLM revealed significantly worse *Sing The Note and Sing The Tune* task performance for web-based participants compared to laboratory-based participants, the violation of the homogeneity of residual variances in these measures threw the significant findings into question. Through PSM, two subsamples of web-based and laboratory-based participants well-matched on age, sex, and years of music training were created, allowing unbiased between-group comparison. Dependent *t*-tests revealed that the two matched subsamples did not perform significantly different in all of the music and singing tasks, demonstrating that the online singing tool remains robust in a web-based setting.

The singing task measures were correlated with self-reported singing ability with moderate to large effect sizes. The significant correlations between self-report and performance-based measures potentially add weight to the convergent validity of the online singing tool (although we note that some past research has observed a dissociation between self-assessed and objectively-measured singing ability, e.g., Pfordresher & Brown, 2007; Pfordresher & Demorest, 2020b). Furthermore, in both samples, larger correlations were observed for tasks involving singing novel or familiar tunes (i.e., *Sing The Tune* and *Happy Birthday*) than for *Sing The Note*. This suggests that a single-pitch imitation task alone may be too elementary to provide adequate construct validity for singing ability and is thus

insufficient to be a representative measure of singing ability in real-world settings.

Finally, our online singing tool showed comparable task performances between our web-based sample and those pooled from 31 past studies using similar laboratory-based singing tasks. Of note, there was considerable heterogeneity in task performance in the previous studies, even when similar stimuli were used in samples with similar characteristics (predominantly university students without formal music training). It is likely that this heterogeneity reflects inherent individual differences in singing ability that has been previously documented by Watts and colleagues (Watts, Moore, & McCaghren, 2005; Watts, Murphy, & Barnes-Burroughs, 2003). Despite this heterogeneity, the pooled mean estimates obtained from the meta-analyses for most of the task measures were highly consistent with the corresponding task measures from the web-based sample, providing further evidence of the robustness of the online singing tool.

A limitation of our study related to the differences in age, sex, and years of music training of the web-based and laboratory-based samples, which posed challenges in determining whether the testing environment has an effect on task performance. Through the use of PSM to obtain matched subsamples of web-based and laboratory-based participants to facilitate unbiased group comparison, this limitation was partly mitigated in the present study. This particular limitation also highlighted a prevalent phenomenon in psychology research that laboratory-based samples usually constitute a convenience sample of university students (often psychology) who are somewhat constrained in age and other demographic features (Henrich, Heine, & Norenzayan, 2010). In contrast, an increasing percentage of the world population now has access to the internet, allowing online samples to reflect greater diversity in demographics, with improved external validity. This supports the merits of conducting online assessments, with larger and more diverse samples to allow researchers to arrive at more robust and externally valid estimates of singing ability for a range of singing tasks.

Conclusion

In this article, we demonstrated the robustness of an online singing tool that exhibited high internal consistency, good test-retest reliability and moderate-to-high correlations with self-reported singing ability. Comparison of matched samples of participants also showed no significant difference in task performance for those who completed the online tool in a web-based as compared to a controlled laboratory setting. Comparison of the web-based sample's task performance with previously published findings of similar singing tasks in comparable populations also revealed no significant differences. Our findings therefore provide good initial evidence of the efficacy of our online singing tool in conducting robust singing assessment in a web-based environment.

Author Note

With the impending phasing out of Adobe Flash in 2020, our online singing tool has been redeveloped as an HTML5 web application which can be run on most platforms (Windows, Macintosh, Linux, Android, and iOS) and most browsers (Google Chrome, Mozilla Firefox, Internet Edge, and mobile Safari) using mobile devices or computers. To our knowledge, our online singing tool is the only web-based tool of its kind with extensive cross-platform capability. A public version of our online singing tool can be accessed from: http://go.unimelb.edu.au/h6aj. Researchers who are interested in using the online singing tool can contact Professor Sarah Wilson at sarahw@unimelb.edu.au.

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP170102479). This research was also facilitated through access to Twins Research Australia, a national resource supported by a Centre of Research Excellence Grant (ID: 1079102), from the National Health and Medical Research Council. We thank Mark Solly for his work in developing the online tool in Adobe Flash and Oscar Correa for developing the updated tool in HTML5. We thank Trisnasari Fraser for her help with recruitment, data collection and processing the singing data of the laboratory-based sample. Finally, we thank Professor Samuel F. Berkovic for his helpful comments on the manuscript.

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. None of the experiments were preregistered.

Correspondence concerning this article should be addressed to Sarah Wilson, Melbourne School of Psychological Sciences, Redmond Barry Building, University of Melbourne, Parkville VIC, Australia 3010. E-mail: sarahw@unimelb.edu.au

References

Amir, O., Amir, N., & Kishon-Rabin, L. (2003). The effect of superior auditory skills on vocal accuracy. *Journal of the Acoustical Society of America, 113*(2), 1102–1108.

Belyk, M., Johnson, J. F., & Kotz, S. A. (2018). Poor neuromotor tuning of the human larynx: A comparison of sung and whistled pitch imitation. *Royal Society Open Science, 5*, 171544. DOI: 10.1098/rsos.171544

Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The Sung Performance Battery (SPB). *Frontiers in Psychology, 4*. DOI: 10.3389/fpsyg.2013.00714

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* (Version 5.4.01). Retrieved from http://www.praat.org/

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. DOI: 10.1037/1040-3590.6.4.284

Cogo-Moreira, H., & Lamont, A. (2017). Multidimensional measurement of exposure to music in childhood: Beyond the musician/non-musician dichotomy. *Psychology of Music, 46*(4), 459–472. DOI: 10.1177/0305735617710322

Dai, J., Mauch, M., & Dixon, S. (2015). *Analysis of intonation trajectories in solo singing.* Paper presented at the Proceedings of the 16th International Society for Music Information Retrieval Conference. Malaga, Spain: ISMIR.

Dalla Bella, S. (2015). Defining poor-pitch singing: A problem of measurement and sensitivity. *Music Perception, 32*(3), 272–282. DOI: 10.1525/mp.2015.32.3.272

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*(2), 1182–1189. DOI: 10.1121/1.2427111

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2009). Singing in congenital amusia. *Journal of the Acoustical Society of America*, *126*(1), 414–424. doi.org/10.1121/1.3132504

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (Eds.). (2011). Chapter 9: Analysing data and undertaking meta-analyses. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011). The Cochrane Collaboration. Available from https://training.cochrane.org/handbook/archive/v5.1/

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*, 266–271. DOI: 1525/mp.2015.32.3.266

Demorest, S. M., Kelley, J., & Pfordresher, P. (2017). Singing ability, musical self-concept, and future musical participation. *Journal of Research in Music Education*, *64*(4), 405–420. DOI: 10.1177/0022429416680096

Erdemir, A., & Rieser, J. J. (2016). Singing without hearing: The use of auditory and motor information when singers, instrumentalists, and nonmusicians sing a familiar tune. *Music Perception, 33*(5), 546–560. DOI: 10.1525/mp.2016.33.5.546

Estis, J. M., Coblentz, J. K., & Moore, R. E. (2009). Effects of increasing time delays on pitch-matching accuracy in trained singers and untrained individuals. *Journal of Voice, 23*(4), 439–445. DOI: 10.1016/j.jvoice.2007.10.001

Estis, J. M., Dean-Claytor, A., Moore, R. E., & Rowell, T. L. (2011). Pitch-matching accuracy in trained singers and untrained individuals: The impact of musical interference and noise. *Journal of Voice, 25*(2), 173–180. DOI: 10.1016/j.jvoice.2009.10.010

George, D., & Mallery, P. (2010). *SPSS for Windows step by step: A simple guide and reference 17.0 update* (10th ed.). Boston, MA: Pearson.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. DOI: 10.1146/annurev-psych-010814-015321

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104. DOI: 10.1037/0003-066x.59.2.93

Granot, R. Y., Israel-Kolatt, R., Gilboa, A., & Kolatt, T. (2013). Accuracy of pitch matching significantly improved by live voice model. *Journal of Voice, 27*(3), 390.e13–390.e3.9E20. DOI: 10.1016/j.jvoice.2013.01.001

Greenberg, D. M., Kosinski, M., Stillwell, D. J., Monteiro, B. L., Levitin, D. J., & Rentfrow, P. J. (2016). The song is you: Preferences for musical attribute dimensions reflect personality. *Social Psychological and Personality Science, 7*(6), 597–605. DOI: 10.1177/1948550616641473

Greenspon, E. B., & Pfordresher, P. Q. (2019). Pitch-specific contributions of auditory imagery and auditory memory in vocal pitch imitation. *Attention, Perception, and Psychophysics, 81*(7), 2473–2481. DOI: 10.3758/s13414-019-01799-0

Greenspon, E. B., Pfordresher, P. Q., & Halpern, A. R. (2017). Pitch imitation ability in mental transformations of melodies. *Music Perception, 34*(5), 585–604. DOI: 10.1525/mp.2017.34.5.585

Harrison, P. M. C., & Müllensiefen, D. (2018). Development and validation of the Computerised Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports, 8*(1), 12395. DOI: 10.1038/s41598-018-30318-8

He, H., & Zhang, W. (2017). Sensorimotor mismapping in poor-pitch singing. *Journal of Voice, 31*(5), 645.e623–645.e632. DOI: 10.1016/j.jvoice.2017.02.018

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83. DOI: 10.1017/S0140525X0999152X

Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011). The Cochrane Collaboration. Available from https://training.cochrane.org/handbook/archive/v5.1/

Honing, H., & Ladinig, 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. DOI: 10.18061/1811/31692

Honing, H., & Reips, U.-D. (2008). Web-based versus lab-based studies: A response to Kendall (2008). *Empirical Musicology Review, 3*(2), 73–77. DOI: 10.18061/1811/31943

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception and Psychophysics*, *76*(8), 2522–2530. DOI: 10.3758/s13414-014-0732-1

Hutchins, S., & Peretz, I. (2012). A frog in your throat or in your ear? Searching for the causes of poor singing. *Journal of Experimental Psychology: General, 141*(1), 76–97. DOI: 10.1037/a0025064

IBM Corp. (2010). *IBM SPSS Statistics for Windows* (Version 19.0). Armonk, NY: IBM Corp.

Internet World Stats (2019). *World internet users statistics and world population stats* [website]. Retrieved from https://www.internetworldstats.com/stats.htm

Kendall, R. A. (2008). Commentary on "The potential of the internet for music perception research: A comment on lab-based versus web-based studies" by Honing & Ladinig. *Empir ical Musicology Review, 3*(1), 8–10. DOI: 10.18061/1811/31693

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. DOI: 10.1016/j.jcm.2016.02.012

Krantz, J. H., & Dalal, R. S. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.

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

Larrouy-Maestri, P., Lévêque, Y., Schön, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice, 27*(2), 259.e251–259.e255. DOI: 10.1016/j.jvoice.2012.11.003

Larrouy-Maestri, P., & Morsomme, D. (2014). Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logopedics Phoniatrics Vocology, 39*(1), 11–18. DOI: 10.3109/14015439.2012.696139

Lévêque, Y., Giovanni, A., & Schön, D. (2012). Pitch-matching in poor singers: Human model advantage. *Journal of Voice*, *26*(3), 293–298. DOI: 10.1016/j.jvoice.2011.04.001

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., et al. (2015). *Computer-aided melody note transcription using the Tony software: Accuracy and efficiency.* Paper presented at the Proceedings of the First International Conference on Technologies for Music Notation and Representation. Paris, France.

Mauch, M., & Dixon, S. (2014, May). *PYIN: A fundamental frequency estimator using probabilistic threshold distributions.* Paper presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy.

Moore, R. E., Estis, J., Gordon-Hickey, S., & Watts, C. (2008). Pitch discrimination and pitch matching abilities with vocal and nonvocal stimuli. *Journal of Voice, 22*(4), 399–407. DOI: 10.1016/j.jvoice.2006.10.013

Moore, R. E., Keaton, C., & Watts, C. (2007). The role of pitch memory in pitch discrimination and pitch matching. *Journal of Voice*, *21*(5), 560–567. DOI: 10.1016/j.jvoice.2006.04.004

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE, 9*(2), e89642. DOI: 10.1371/journal.pone.0089642

Murry, T. (1990). Pitch-matching accuracy in singers and non-singers. *Journal of Voice, 4*(4), 317–321. DOI: 10.1016/S0892-1997(05)80048-7

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–88). San Diego, CA: Academic Press.

Peretz, I., & Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics, 25*(5), 625–630. DOI: 10.1038/ejhg.2017.15

Pfeifer, J., & Hamann, S. (2015). Revising the diagnosis of congenital amusia with the Montreal Battery of Evaluation of Amusia. *Frontiers in Human Neuroscience, 9*, 161. DOI: 10.3389/fnhum.2015.00161

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of 'tone deafness.' *Music Perception, 25*(2), 95–115. DOI: 10.1525/mp.2007.25.2.95

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception and Psychophysics, 71*(6), 1385–1398. DOI: 10.3758/app.71.6.1385

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128*(4), 2182–2190. DOI: 10.1121/1.3478782

Pfordresher, P. Q., & Demorest, S. M. (2020a). 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: Development* (Vol. 1, pp. 322–333). London, UK: Routledge.

Pfordresher, P. Q., & Demorest, S. M. (2020b). The prevalence and correlates of accurate singing. *Journal of Research in Music Education,* DOI: 10.1177/0022429420951630

Pfordresher, P. Q., & Halpern, A. R. (2013). Auditory imagery and the poor-pitch singer. *Psychonomic Bulletin and Review, 20*(4), 747–753. DOI: 10.3758/s13423-013-0401-8

Price, H. E. (2000). Interval matching by undergraduate nonmusic majors. *Journal of Research in Music Education, 48*(4), 360–372. DOI: 10.2307/3345369

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49*(4), 243–256. DOI: 10.1026//1618-3169.49.4.243

Thoemmes, F. (2012). *Propensity score matching in SPSS*. Retrieved from http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf

Tremblay-Champoux, A., Dalla Bella, S., Phillips-Silver, J., Lebrun, M. A., & Peretz, I. (2010). Singing proficiency in congenital amusia: Imitation helps. *Cognitive Neuropsychology, 27*(6), 463–476. DOI: 10.1080/02643294.2011.567258

Ullén, F., Mosing, M. A., Holm, L., Eriksson, H., & Madison, G. (2014). Psychometric properties and heritability of a new online test for musicality, the Swedish Musical Discrimination Test. *Personality and Individual Differences, 63*, 87–93. DOI: 10.1016/j.paid.2014.01.057

Wallace, B. C., Schmid, C. H., Lau, J., & Trikalinos, T. A. (2009). Meta-Analyst: Software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology, 9*(1), 80. DOI: 10.1186/1471-2288-9-80

Watts, C. R., & Hall, M. D. (2008). Timbral influences on vocal pitch-matching accuracy. *Logopedics, Phoniatrics, Vocology, 33*(2), 74–82. DOI: 10.1080/14015430802028434

Watts, C., Moore, R., & McCaghren, K. (2005). The relationship between vocal pitch-matching skills and pitch discrimination skills in untrained accurate and inaccurate singers. *Journal of Voice, 19*(4), 534–543. DOI: 10.1016/j.jvoice.2004.09.001

Watts, C., Murphy, J., & Barnes-Burroughs, K. (2003). Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. *Journal of Voice, 17*(2), 185–194. DOI: 10.1016/S0892-1997(03)00023-7

Wise, K., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined 'tone deafness': Perception, singing performance, and self-assessment. *Musicae Scientiae, 12*, 3–23. DOI: 10.1177/102986490801200102

Yang, W., Feng, J., Huang, W., Zhang, C., & Nan, Y. (2014). Perceptual pitch deficits coexist with pitch production difficulties in music but not Mandarin speech. *Frontiers in Psychology, 4*(1024). DOI: 10.3389/fpsyg.2013.01024

Zarate, J. M., Delhommeau, K., Wood, S., & Zatorre, R. J. (2010). Vocal accuracy and neural plasticity following micromelody-discrimination training. *PLOS ONE, 5*(6), e11181. DOI: 10.1371/journal.pone.0011181

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences, 1400*(1), 33–45. DOI: 10.1111/nyas.13410