Original Research Article

National Association for Music Education

Journal of Research in Music Education
2021, Vol. 69(1) 5–23

DOI: 10.1177/0022429420951630
jrme.sagepub.com
SAGE

# The Prevalence and Correlates of Accurate Singing

**Peter Q. Pfordresher[1] and Steven M. Demorest[2]**


## Abstract

The purpose of this study was to analyze a large sample of volunteers from the general population who were tested with an identical online measure of singing accuracy. A sample of 632 participants completed the Seattle Singing Accuracy Protocol (SSAP), a standardized measure of singing accuracy, available online, that includes a test of pitch discrimination and basic demographic questions. Analyses addressed basic questions relating to the distribution of singing accuracy as well as associations of singing accuracy with years of musical training, age, pitch discrimination ability, and musical self-perception. We addressed these issues with respect to the accuracy of pitch imitation, based on automated scoring of vocal fundamental frequency ($f_0$) in the SSAP, as well as the accuracy of singing a familiar song, based on expert ratings. Results suggest that the distribution of singing accuracy varies widely, but the modal tendency is toward accurate singing. All predictors formed unique and significant associations with singing accuracy, suggesting that multiple factors contribute to this critical musical ability. In particular, age and musical training (including instrumental training) correlate independently with singing accuracy.



## Keywords

singing accuracy, measures of musical ability, musical training effects, musical self-image, pitch discrimination


[1]University at Buffalo, State University of New York, Buffalo, NY, USA
[2]Northwestern University, Evanston, IL, USA

**Corresponding Author:**
Peter Q. Pfordresher, Department of Psychology, University at Buffalo, 206B Park Hall, Buffalo, NY 14260, USA.
Email: pqp@buffalo.edu

The basis for singing accuracy and how to remediate inaccurate or "poor pitch" singing remain both critical issues and major puzzles for music education. The importance of singing accuracy extends beyond the choral concert. Singing may open a gateway to many other important musical experiences (Welch, 2006), and withdrawal from singing can close the same door (Demorest et al., 2017). Unfortunately, there is not yet a thorough understanding of how singing develops through the life span (Loui et al., 2015). There is, however, more information about children's singing development. In 1963, Petzold began a landmark 3-year longitudinal study of children in grades 1 to 6. Although his stated goal was to document growth in auditory perception, his main measure was of singing accuracy. He found that children's singing accuracy improved as they got older and concluded that accuracy improves with age, a finding echoed by a number of subsequent investigations with children (Cooper, 1995; Green, 1990; Welch et al., 1997). In the present study, we leveraged the opportunity provided by online data collection to address developmental factors related to singing accuracy, including age and musical training.

To date, there is considerable evidence that singing accuracy improves during the elementary school years, but there are two problems with concluding that these improvements are solely related to age. The first problem is that in grades 1 through 6 (common elementary school years), students typically have mandatory general music instruction that focuses mostly on singing. As a result, most children acquire considerable singing experience until ages 11 to 12. Most studies tracking improvement of singing accuracy with age focus on elementary school children, thus making it unclear how much age versus experience in singing lessons accounts for their improvement (Loui et al., 2015). The relatively few studies that have looked at singing in middle school (Demorest, 2001; Demorest & Clements, 2007; Yarbrough et al., 1991) have concentrated on poor-pitch singers. In contrast to the music education literature, studies from cognitive psychology have focused almost exclusively on young-adult college populations (Berkowska & Dalla Bella, 2013; Hutchins & Peretz, 2012; Pfordresher & Brown, 2007; Sloboda et al., 2005). As a result, we know more about singing accuracy among children and young adults but not what goes on between these periods or singing accuracy in older adulthood.

Cross-sectional comparisons across age groups are further hindered by the fact that most studies use different measures of accuracy. A meta-analysis by Svec (2018) found that 27 of 34 studies looking at children's singing improvement between ages 5 and 11 used a different measure of accuracy. Demorest and Pfordresher (2015b) attempted to address this discrepancy by comparing results of three separate studies that used an almost identical singing task with kindergartners, sixth graders, and adults. They found that, as predicted, sixth graders were better than kindergartners in all areas of pitch matching. However, adults were more similar to kindergartners on all tasks except when matching single pitches (for which adults outperformed kindergarteners). This led the investigators to speculate that singing is not a talent but rather a "use it or lose it" skill. The decrement in performance among college students relative to middle school participants was interpreted as reflecting the fact that after the sixth grade, most students cease participation in elective music (approximately 66%;

Demorest et al., 2017). By contrast, some form of regular musical instruction is offered prior to sixth grade in most North American schools. Many adults may therefore not maintain the singing accuracy skills they obtained during regular musical instruction up to sixth grade.

A particularly important issue is the frequency of accurate or poor-pitch singing in the general population. Unfortunately, it is difficult to get a clear picture from previous research due to the differences in measures and tasks used. For a given task (e.g., imitation of four-note melodies), the frequency of singing accuracy can vary dramatically based on what measure of performance is used as well as the criterion on which division into different groups is based (Berkowska & Dalla Bella, 2013; Demorest & Pfordresher, 2015b; Hutchins & Peretz, 2012; Pfordresher et al., 2010; Pfordresher & Larrouy-Maestri, 2015). Two early studies yielded results suggesting that approximately 15% of the population may be classified as poor-pitch singers. However, even these studies differed with respect to tasks and measures. Whereas one study focused on analyses of interval errors during the singing of a familiar song from memory (Dalla Bella et al., 2007), the other focused on pitch-matching deviations during the imitative singing of novel melodies (Pfordresher & Brown, 2007).

Advances in online data collection offer a way around these issues given the possibility of collecting large samples from diverse populations. To date, such procedures have led to great advances in our understanding of how music perception abilities are distributed (e.g., Peretz & Vuvan, 2017) as well as potential genetic bases for musical deficits and ability (e.g., Mosing et al., 2014). These tests have been primarily oriented toward perception, however, which makes sense given the difficulty of collecting valid performance data outside the context of a lab. The present study represents a new trend toward collecting performance data online, focusing on singing as the most prevalent form of music-making.

In the present study, we drew on a larger sample than any existing study of singing accuracy ($N = 632$) and focused on a measure that we considered to be optimally transparent: the frequency of pitches produced accurately. Individual produced pitches were classified as accurate if they were produced within $\pm 50$ cents (i.e., one semitone) of a target pitch. This classification method was used because 50 cents marks the dividing line between proximity to the intended pitch class and an alternate pitch class within the tuning system and was used in other recent work (e.g., Demorest & Pfordresher, 2015b; Hutchins & Peretz, 2012). Singers were classified as "accurate" or "inaccurate" based on Hutchins and Peretz's (2012) criterion that a singer needs to produce at least 90% of pitches accurately to be considered an accurate singer. This is also consistent with practical criteria music educators may use when classifying accurate singers.

The purpose of this study was to ascertain the prevalence of accurate singing ability among a large online sample. We also sought to determine how participants' age, years of private lessons, years of singing experience, self-perception of musicality, and perceptual accuracy correlate with this ability. A key feature of this research is that all participants, regardless of their age, took the same standardized online measure.

The research questions were as follows: (1) How is singing accuracy distributed? (2) Is singing accuracy related to age? (3) Is singing accuracy related to singing experience and/or private music lessons? (4) Is singing accuracy related to pitch discrimination ability? (5) Is singing accuracy related to self-assessments of musicality? and (6) How do different variables jointly predict accuracy matching pitch and singing a familiar song?

## Method

### Participants

This study involves analyses of singing performances and other responses from the Seattle Singing Accuracy Protocol (SSAP; Demorest & Pfordresher, 2015a), an online measure of singing accuracy that is freely available on the web to the public, as part of a research program in September 2015. The availability of the SSAP was made known by advertisements in social media, word of mouth among personal contacts, announcements in class, and announcements in professional electronic mailing lists. Those participating as members of the "public" chose that option at the outset; participants in the present study were from this group and thus were not subject to any kind of experimental manipulations outside the structure of the SSAP. The study was conducted with the approval of and in accordance with the guidelines of the institutional review boards at Northwestern University and the University at Buffalo.

We started by analyzing 735 files from general public participants that were available from the database on April 13, 2018. First, a research assistant listened to all files to ensure that recording quality was sufficient and that audio files included singing that reflected a sufficient attempt to imitate stimuli (e.g., files with extraneous noises, coughing, and apparent joking behavior were excluded). We also eliminated any participants who self-declared a hearing or neurological deficit. After this stage, 632 participants qualified for further analysis. Participants self-reported gender as male or female in association with their singing voice: 383 participants reported as female, and 249 were male.

The sample included a wide range of ages, as shown in Figure 1a. Ages ranged from 6 to 99 years, with an average age of 30.83 ($SD = 16.31$, interquartile range = 20–41), with the most frequent ages being found within the range of traditional college education (18–22). Figure 1b displays a histogram for self-reported years of formal musical training in instrument or voice. Although many musicians contributed to our sample, the modal years of training reported was 0, with 292 (46%) reporting no training. Self-reported years of training averaged 5.37 years ($SD = 8.47$). Forty-seven participants self-identified as primarily vocalists; this constituted a small portion of the sample (7%). However, a considerably large number reported experience singing in choirs ($N = 197$, 31%).

Participants reported the country in which they completed the SSAP. The sample was global in scope, although the vast majority came from developed Western nations ($n = 605$, 97%); the most well-represented countries were the United States ($n = 411$,

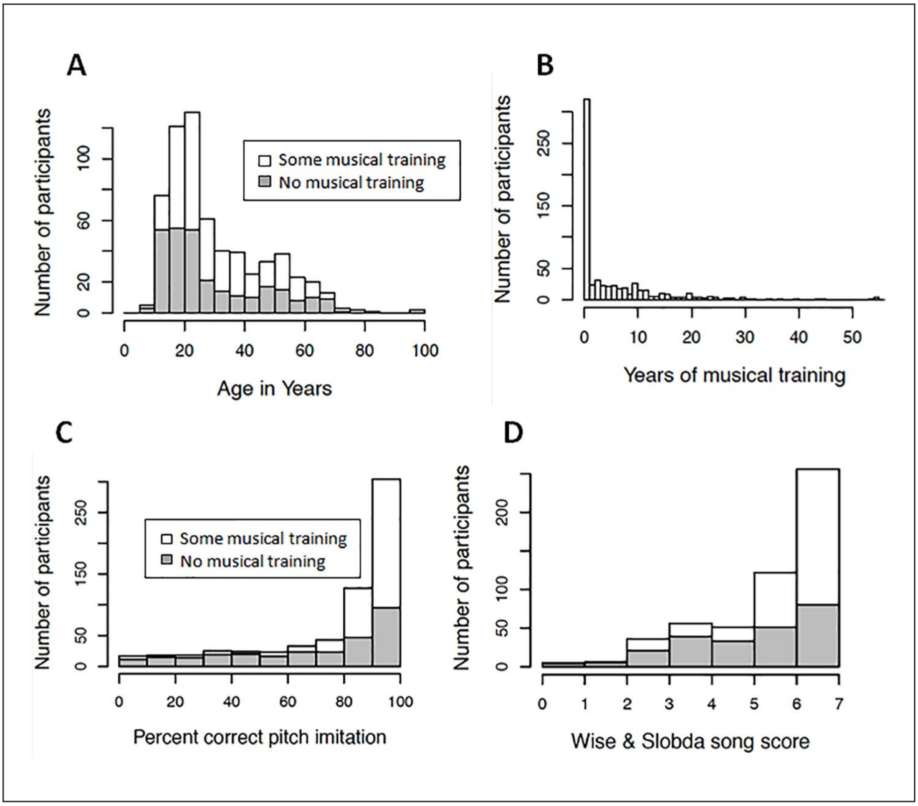


**Figure 1.** (a) Stacked histogram displaying participant ages in 5-year bins for participants indicating at least 1 year of formal musical training versus those with no years reported. (b) Years of musical training across participants (binned by year), summed across all instruments and voice. (c) Stacked histograms displaying percentage correct pitch imitation in 10% bins for participants indicating at least 1 year of formal musical training versus those with no years reported and (d) for song-singing scores using the revised Wise and Sloboda scale.

65%) and Denmark ($n$ = 127, 21%). Every other country contributed less than 3% of the total data. Eight participants (1%) were from Asia, and 4 (0.6%) were from Latin America.

## *Materials and Procedure*

The SSAP consists of five parts. The first part identifies the most comfortable key for a participant based on two singing tasks. First, participants chose to sing one of five familiar songs—"Jingle Bells," "Frere Jacques," "Twinkle, Twinkle Little Star," "The ABC Song," or "Mary Had a Little Lamb"—in a self-selected key. Second, they sustained a single "comfortable" pitch on "oo" for 2s. The continuous fundamental frequency ($f_0$) of the voice was extracted from both these recordings using the Matlab function *Yin* (de Cheveigné & Kawahara, 2002), and the average of these values was used to determine a key with a comfortable range of a perfect fifth centered on the mean (cf. Pfordresher & Demorest, 2020).

Once a comfortable key was determined, the second part of the measure involved imitative pitch matching based on a call–response procedure. Participants used the syllable "doo" for each individual produced pitch. There were three conditions. The first condition comprised 10 trials in which participants imitated single pitches based on a vocal timbre. Each pitch was sampled from the first 5 scale degrees from the participants' comfortable key. Model recordings were from music graduate students (one male, one female) singing without perceivable vibrato. Pitches in single-pitch trials were selected from a range of a fifth and were presented twice in random order. The second condition comprised 10 trials in which participants imitated single pitches based on a piano timbre (recorded at the Royal Conservatory of Toronto by Sean Hutchins). The same pitches were used as in the first condition but in a different order. Finally, the third condition comprised six trials in which participants imitated four-note melodies based on a vocal timbre using pitches sampled from the same sets and using vocal timbres from the first condition. The highest possible score was 44 notes correct (10 + 10 + 24 = 44).

Because the measure was conducted online, trials were occasionally skipped by participants or discarded due to poor recording quality. As such, scores were presented as a proportion of all recorded pitches (e.g., if data from one of the four-pitch patterns were not collected, the participant's proportion correct was computed out of 40 instead of 44). A previous study using a similar automated scoring method correlated at $r$ = .92 with human judgment, which was considered excellent (Demorest & Pfordresher, 2015b), and a more recent analysis suggests that the automated scoring procedure corresponds with analyses of $f_0$ carried out manually (Pfordesher & Demorest, 2020).

The third part of the measure asked participants to sing their chosen familiar song (from part one) and reproduce it first using the original text (displayed on the screen for the participant) and then to reproduce it a second time replacing each syllable with "doo." Participants were free to choose the key and tempo for these performances. We ended up scoring only the text version by hand because there were many rhythm errors for performances using "doo," possibly reflecting confusion of syllabification. As discussed in Pfordresher and Demorest (2020), automated scoring of song singing in the current version of the SSAP does not rise to the level of validity for automated scoring of imitative singing. As such, we here report results from live scorers who used a modified version of the Wise and Sloboda (2008) 8-point singing accuracy scale. The modified version used only 7 categories instead of 8. Our reviewers had difficulty reliably distinguishing between Categories 4 and 5 on the original scale, so they were combined in Category 4 on our revised scale (see Figure S1 in the supplementary files included with the online version of this article). The interrater reliability of the scale was determined by having two raters score 143 of the song examples. The resulting reliability of .84 was deemed acceptable, so the remaining 389 examples were scored by a single rater. Recordings from 100 participants were not of sufficient audio quality to be scored, leading to a lower number of participants for song-singing analyses ($N$ = 532) than for vocal imitation analyses ($N$ = 632).

The next part of the measure was an adaptive perceptual discrimination task, based on Loui et al. (2008), in which participants heard two successive computer-generated pure tones and indicated whether the second was higher or lower in pitch than the first. The order of presentation regarding higher or lower pitch changes was random, and the size of the change varied depending on whether participants responded correctly (leading to a smaller size) or incorrectly (leading to a larger size). The task ended when participant responses suggested they had reached their discrimination threshold. All participants in the sample analyzed here exhibited discrimination thresholds under 100 cents, which is important given that thresholds greater than 100 cents may indicate a pitch perceptual deficit (e.g., congenital amusia; Hyde & Peretz, 2004).

The final part of the measure was a short survey asking participants about their music experience and language background. There were questions about their age, years of private music lessons, years of group singing experience, and self-perception of musicality on a Likert scale from 1 to 7, where 1 = *entirely disagree*, 4 = *neutral*, and 7 = *entirely agree*. The items participants responded to included "I enjoy singing," "People think I am a good signer," and "I am musically talented." Once they completed the survey, they were shown their score for pitch matching, song singing (automated score only), and perceptual deviation in cents.

## Results

### *Distribution of Singing Accuracy*

We first sought to determine how accurate the sample was at matching pitch and singing songs from memory. For pitch matching (imitation), we focused on the percentage of pitches matched correctly (i.e., within a semitone) as a highly intuitive measure that is used in most educationally oriented studies of singing accuracy. Figure 1c plots a histogram of percentage correct pitch for each individual in the database. As can be seen, the distribution of accuracy scores exhibits a strong negative skew, with nearly half the scores falling into a bin spanning from 90% to 100% accuracy. About half of our sample may therefore be considered accurate pitch matchers (more than 90% correct). Note that this designation does not necessarily imply that all the remaining participants may be considered poor-pitch singers given the likelihood that at least one other category—often called "uncertain singers"—likely exists (Demorest & Clements, 2007; cf. Pfordresher & Larrouy-Maestri, 2015).[1]

Another caveat for interpreting Figure 1c is the fact that the database may be dominated by musically trained individuals given the self-selecting nature of the test. Thus, we reanalyzed the distribution of accuracy scores and included only those participants who reported no formal training in music ($n$ = 292). A comparison between trained and untrained singers can be gleaned from Figure 1c by comparing the relative heights of white (trained) versus gray (untrained) regions of each bar. Musically trained individuals dominated frequency counts for levels of accuracy greater than 90% correct, whereas musically untrained individuals tended to dominate frequencies at lower levels of accuracy, and the distribution of frequencies across bins varied significantly for trained and untrained participants, $\chi^2(9) = 86.66$, $p < .001$. Thus, musical training of any kind does appear to have an association with singing accuracy (cf. Demorest et al., 2018). Nevertheless, the modal tendency in the sample, independent of training, is toward accurate singing.

Distributions for scoring of song-singing accuracy, using the revised Wise and Sloboda scale (2008; see Figure 2), are shown in Figure 1d. Scores on this task were significantly correlated with percentage correct for imitative singing, $r(530) = .61$, $p < .001$. With respect to ratings of song-singing accuracy, we found that the median score was 6 out of 7 (mode = 7). This corresponded to the following rubric: "Key is maintained throughout, and accurately represented with some mistunings (though not enough to alter the pitch-class of the note)." So approximately 50% of our sample would be considered mostly accurate on both tasks. Similar to percentage correct, untrained individuals exhibited lower scores on song-singing, leading to different distributions for trained and untrained participants, $\chi^2(9) = 50.19$, $p < .001$, although the highest possible score in the Wise and Sloboda scale was the mode for both groups.

## *Musical Training*

Next, we analyzed whether musical training predicted singing accuracy.[2] Our initial measure of training included instrumental as well as vocal training given that very few participants in our sample who self-identified as some kind of a musician (approximately 57% of the sample) identified as primarily vocalist (7% of the sample). Years of training correlated significantly with percentage correct pitch matching, $r(630) = .30$, $p < .001$. Years of training accounted for a similar portion of variance in song-singing scores, $r(530) = .29$, $p < .001$ (for scatterplots, see Figure S2 in the supplementary files included with the online version of this article).

To further explore the role of training, we directly compared the 292 subjects who reported no prior private lesson training with the 340 who did on both pitch matching and song singing (the remainder did not answer the item). For pitch matching, privately trained subjects outscored untrained subjects by an average of 87% to 68% with median scores of 94% and 78%, respectively. For familiar-song scores, the analogous comparison involved means of 6.24 versus 5.38. Although few of our participants self-identified as primarily singers or having private vocal lessons, a sizable portion of our sample ($n = 197$, 31%) reported primary group music-making experience in choirs. Percentage correct pitch matching was higher among this group (87.6%) than all other participants (77.4%), although with a small effect size, $t(414.17) = 2.24$, $p = .026$, $r^2 = .01$, applying Welch's correction for unequal variances. Participants with primary group experience in choir also exceeded pitch-imitation accuracy of those participants who reported primary group experience in band (75.2%), although, again, this difference represents a small effect, $t(457.92) = 3.07$, $p = .002$, $r^2 = .02$.

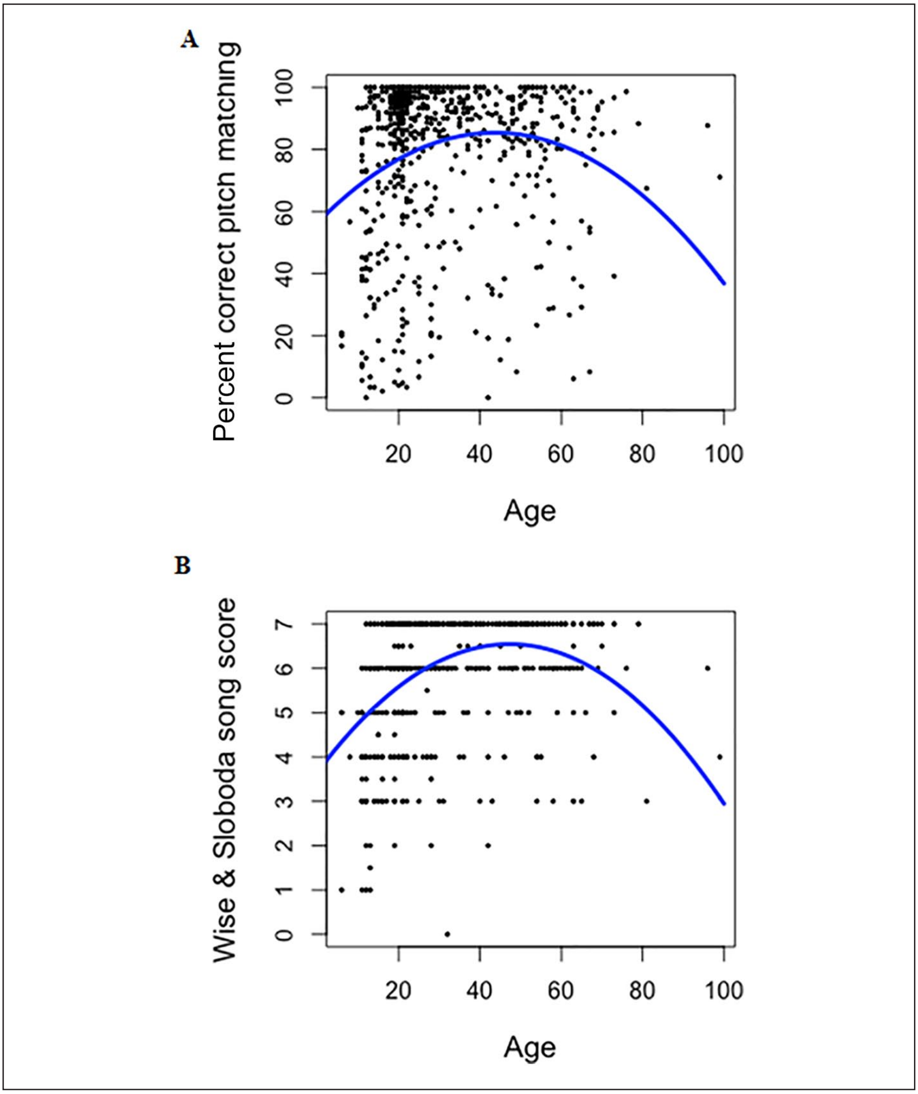


**Figure 2.** Scatterplots displaying the relationship between age in years with (a) percentage correct pitch matching and (b) singing a familiar song from memory. Lines represent best-fitting quadratic polynomial regression.

## *Age*

The correlation between age (in years) and pitch-matching accuracy on imitation tasks was statistically significant but accounted for a very small amount of the variance, $r(630) = .09$, $p = .03$. Further investigation suggested that this effect was nonlinear, and a multiple regression equation that added a quadratic (squared) term accounted for more total variance and was statistically significant, exhibiting a considerably larger (although still modest) effect size, $R(629) = .16, p < .001$. As shown in Figure 2a, the model predicts the most accurate performance for adults aged 40 to 60, with lower levels of performance for children, young adults, and older adults.

A similar pattern was found for song-singing scores but with stronger associations overall. Age correlated linearly with ratings of song singing on this scale, $r(530) = .24, p < .001$, and the addition of a quadratic term, shown in Figure 2b, led to a significant increase in variance accounted for, $R(529) = .38, p < .001$.

## Perceptual Discrimination

Another question concerned the relationship between perceptual discrimination and singing abilities. For this question, we correlated perceptual discrimination scores with pitch matching and familiar song scores. The correlation between perception and pitch matching was $r(630) = -.31$, $p < .001$. Participants with smaller perceptual discrimination thresholds also tended to have better pitch-matching scores. There was a similar negative relationship between perception and singing a familiar song, $r(532) = -.21, p < .001$. These revealed statistically significant relationships between perception and production that fall within a small to moderate effect size range (Ferguson, 2009; for scatterplots, see Figure S3 in the supplementary files included with the online version of this article ).

## Self-Perception

Singers were also asked to assess their own singing and general musical ability. There were three items on the SSAP that assessed self-perception: "I enjoy singing," "People think I am a good singer," and "I am musically talented." Participants rated their agreement with each item on a 7-point Likert scale, with 7 indicating highest agreement. The item "I am musically talented" yielded the strongest association with accuracy in pitch matching, $r(629) = .41$, and song-singing, $r(532) = .38$, $p < .001$. The other items yielded weaker, although statistically significant associations: for "I enjoy singing," pitch matching $r(629) = .18$, song-singing $r(532) = .22$; for "People think I am a good singer," pitch matching $r(629) = .33$, song singing $r(532) = .35$ ($p < .01$). Thus, we focus on associations with the item "I am musically talented."

A boxplot showing how percentage correct pitch matching varies for each response is shown in Figure 3a, with corresponding data for song singing shown in Figure 3b. There is a clear association between self-assessment of "talent" and singing accuracy, although singers with poor accuracy are evident for each response category. At the same time, it is important to note that highly accurate singers and highly inaccurate singers appear in each response group. In a small but noteworthy number of instances, participants expressed poor musical self-image that conflicted with their own pitch accuracy. For instance, among the 58 participants (9%) who expressed greatest disagreement with the item "I am musically talented," 14% were categorized as accurate singers based on the criterion given previously. Even more striking, 71 participants (11%) expressed strongest disagreement with the item "People think I am a good singer," but 9 of these participants (8%) were categorized as accurate.

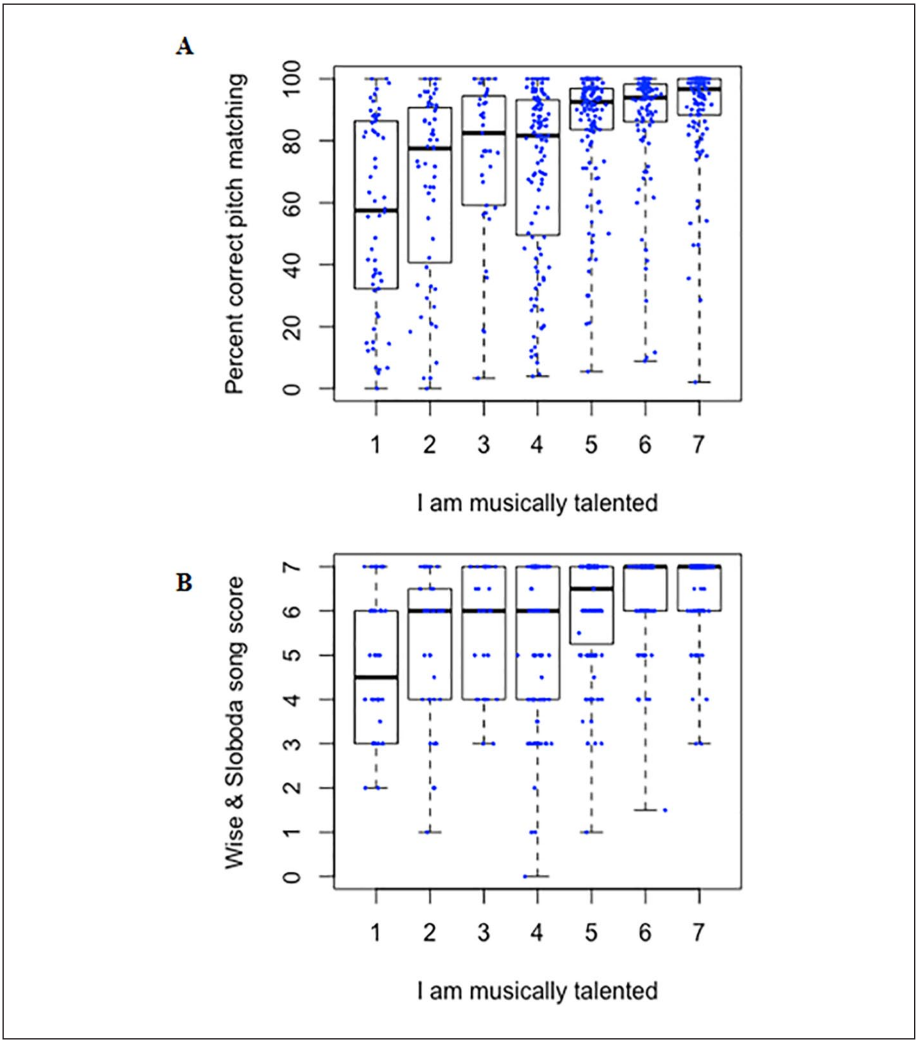


**Figure 3.** Boxplots displaying the relationship between self-evaluation responses to the item "I am musically talented," where higher numbers reflect greater agreement, with (a) pitch matching accuracy and (b) singing a familiar song from memory.

## *Hierarchical multiple regression analysis*

We addressed the relative contribution of each of these variables to singing accuracy through hierarchical multiple regression (Cohen & Cohen, 1983). At Step 1, we regressed percentage correct pitch matching on age using both the linear and quadratic terms that accounted for the most variance in the analysis reported earlier. At Step 2, we added musical training, which is somewhat influenced by one's age. At Step 3, we added pitch discrimination threshold, which is known to be influenced by musical training. At Step 4, we added self-perception of musical talent, which is likely to be influenced by all other variables and is also the predictor that we considered to be most theoretically novel. As can be seen in Table 1, the regression model at each step accounted for a significant proportion of variance, with each successive model accounting for significantly more variance than the previous step. These results suggest that singing accuracy is associated with a multicomponent model of singing accuracy, influenced by experiential, sensorimotor, and attitudinal factors. At the same time, it is worth noting that the total variance accounted for, even with all factors included, is just over 25%. Similar results were obtained for a hierarchical regression conducted on song-singing scores (also shown in Table 1).

### Comparison With a Lab-Based Sample

The online sample that we primarily focused on is advantageous in its diverse representation of age, experience, and regional affiliation. However, a potential limitation of this sample lies in the fact that this was a self-selecting group. A comparison sample of 414 participants was compiled of college-aged participants from the University at Buffalo who completed a console-version of the SSAP on Matlab in exchange for introduction to psychology course credit. These participants completed the procedure as part of a screening phase prior to participating in one of several studies taking place in the first author's research lab. The only exclusion criteria for their participation was the presence of a diagnosed hearing deficit or vocal phonation disorder. An additional 37 participants were excluded from an initial sample size of 451 due to having a pitch discrimination threshold greater than 100 cents.

Our primary reason for analyzing this comparison data set was to determine whether overall accuracy differs for self-selected versus randomly sampled participants. Statistical comparisons across groups are shown in Table S2 in the supplementary files included with the online version of this article. The lab-based sample included more musically untrained individuals (61% vs. 46% from the online sample) and was far more constrained with respect to age, with the vast majority of participants ranging from 18 to 22 years of age—the most typical age range of college students in North America (96% vs. 29% from the online sample). These differences support the utility of our online sample for the purposes of understanding the role of age and musical training in singing accuracy. At the same time, the lab-based sample scored significantly lower than the online sample in singing accuracy, even when limiting both samples to individuals with no training and from similar years of age ($M$ = 78% vs. 53% correct pitch matching for online and lab samples, respectively).

These differences suggest that there are distinguishing features in the self-selecting sample that lead it to represent a relatively high level of singing accuracy—quite possibly the motivation to participate voluntarily in a singing accuracy measure. It is worth noting that these differences could be interpreted as evidence of unusually high motivation among the online sample or unusually low motivation among the lab-based

**Table 1.** Hierarchical Regression Results.

| | | Step 1 | | | Step 2 | | | Step 3 | | | Step 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | Predictor | *B* | *SE* | β | *B* | *SE* | β | *B* | *SE* | β | *B* | *SE* | β |
| Imitation | Age | 1.33 | .263 | .85 | 1.10 | .256 | .69 | .90 | .247 | .57 | 1.03 | .234 | .65 |
| | Age$^2$ | −.02 | .003 | −.77 | −.01 | .003 | −.67 | −.01 | .003 | −.55 | −.01 | .003 | −.59 |
| | Music training | | | | .83 | .118 | .27 | .74 | .114 | .24 | .40 | .114 | .13 |
| | Pitch threshold | | | | | | | −.39 | .055 | −.27 | −.32 | .052 | −.21 |
| | Singing attitude | | | | | | | | | | 4.53 | .518 | .33 |
| | Adjusted $R^2$ | | .04 | | | .11 | | | .18 | | | .27 | |
| | $\Delta R^2$ | | | | | .07 | | | .07 | | | .09 | |
| | $F$ for $\Delta R^2$ | | | | | 60.53 | | | 58.63 | | | 76.31 | |
| Song | Age | .12 | .247 | 1.42 | .11 | .014 | 1.29 | .11 | .014 | 1.23 | .11 | .013 | 1.29 |
| | Age$^2$ | −.001 | .0002 | −1.21 | −.001 | .0002 | −1.12 | −.001 | .0002 | −1.06 | −.001 | .0002 | −1.09 |
| | Music training | | | | .04 | .007 | .22 | .03 | .007 | .21 | .02 | .006 | .10* |
| | Pitch threshold | | | | | | | −.01 | .003 | −.15 | −.01 | .003 | −.09* |
| | Singing attitude | | | | | | | | | | .26 | .031 | .33 |
| | Adjusted $R^2$ | | .14 | | | .19 | | | .21 | | | .30 | |
| | $\Delta R^2$ | | | | | .05 | | | .02 | | | .09 | |
| | $F$ for $\Delta R^2$ | | | | | 35.47 | | | 15.92 | | | 102.47 | |

*Note.* All beta coefficients are significant at $p < .001$, except $^*p < .05$. Imitation = accuracy across all vocal pitch imitation tasks using pitch-matching errors as the Y-variable. Song = accuracy for familiar-song singing using scores on the modified Wise and Sloboda scale as the Y-variable.

sample. It is also possible that the online group comprises individuals who chose to take the SSAP due to a preexisting sense of self-efficacy in singing. In any case, levels of accuracy in the general population likely reflect some mixture of the accuracy observed in these samples. When combining the online and lab-based samples here (total $N$ = 1,046), 37% were classified as accurate. When considering only those participants with no musical training, 24% were classified as accurate.

A related question has to do with how accuracy was distributed in the lab-based sample. The online sample exhibited a pronounced mode for overall accuracy ranging from 90% to 100% (Figure 2). The corresponding histogram of accuracy scores across participants for the lab sample likewise exhibits a mode at the highest level of accuracy (see Figure S4 in the supplementary files included with the online version of this article). Thus, even though fewer lab-based participants would be classified as accurate compared to the online sample, it is still the case in the lab-based sample that the modal tendency is toward accurate performance.

## Discussion

The present research addressed two goals: (a) to offer an updated estimate of singing proficiency in the general population and (b) to assess how singing accuracy is associated with critical measures. Both goals benefited from the use of a standardized measure of singing accuracy, the Seattle Singing Accuracy Protocol, that was designed by a team of leading researchers on singing accuracy from psychology and music education (Demorest et al., 2015). In addition, the present research benefits from the use of a considerably larger sample than other studies of singing accuracy ($N$ = 632). Moreover, unlike most studies, ours did not use a sample based on students in a school or university but rather, involves an entirely volunteer-driven pool representing a wide range of ages and musical experience from around the world.

The prevalence of singing accuracy is a critical question, although it is important to recognize that the abundance of measures and criteria that are available can alter one's conclusions considerably. Previous studies using different measures have yielded data suggesting that singing accuracy may be present in the majority of the population (e.g., Dalla Bella et al., 2007; Pfordresher & Brown, 2007) or in the minority (e.g., Hutchins & Peretz, 2012; Pfordresher et al., 2010). We chose a measure that we consider to be most useful to music educators (the percentage of pitches imitated within one semitone, ±50 cents, of a target pitch) and a criterion that struck us as having practical significance (at least 90% pitches correct to be considered accurate). Based on these standards, results from our online sample suggest that about half of the general population may be considered accurate. However, as with other studies, this estimate must be considered tentative. In particular, comparison with a random sample of college-aged participants, whose participation was motivated by fulfilling a course requirement as opposed to volunteering, exhibited considerably lower rates of accuracy. Taking both samples under consideration, a rough estimate we offer is that about one third of the population may match pitch accurately the vast majority of the time. Moreover, a clear modal tendency toward accurate singing (i.e., 90%–100% accuracy) was found in both samples. In several cases, the accuracy exhibited by participants conflicted with their self-assessments, leading to participants who self-rated as untalented or not a good singer while exhibiting highly accurate pitch matching.

Many previous studies have focused on rates of inaccurate or poor-pitch singing rather than on rates of accurate singing (e.g., Berkowska & Dalla Bella, 2013; Hutchins & Peretz, 2012; Pfordresher & Brown, 2007; Pfordresher & Larrouy-Maestri, 2015). It is important to note that participants in the present study who were not classified as accurate would not be classified as poor pitch singers based on previous criteria. For instance, if one adopted a criterion based on mean signed pitch deviations of greater than 50 cents (suggested by Dalla Bella & Berkowska, 2009) to separate poor-pitch from accurate singers, the online sample here would have 89% accurate singers (11% poor pitch). Ultimately, categorization of singing accuracy should strive toward nuance, the determination of various phenotypes for singing behavior, and the avoidance of simple dichotomies (cf. Berkowska & Dalla Bella, 2013; Demorest & Clements, 2001; Hutchins & Peretz, 2012; Pfordresher & Larrouy-Maestri, 2015).

It seems that musical training of any kind, instrumental or vocal, has a positive impact on accuracy. Singing accuracy among musically untrained individuals was considerably lower than for musically trained individuals both for the online and lab-based data sets. We also found an additional advantage for singing in choirs beyond instrumental training, although not for the small number of participants who self-identified as vocalists. Clearly, vocal training may yield other benefits (e.g., vocal timbre, articulation, expression) and may facilitate singing accuracy in a more fine-grained way than our current measures address. Nevertheless, the present data are consistent with the hypothesis that musical training in general facilitates vocal pitch imitation via the enhancement of sensorimotor associations.

The variables of age and perceptual skills were both significant in the regression but, similar to musical training, were weakly to moderately correlated with both accuracy measures. The other variable examined was self-perception of musicality. It is interesting to note how much of the variance was accounted for by self-perception, which led to the largest changes in variance accounted for within the hierarchical regression despite being entered last (i.e., with lowest power). This suggests that people associate singing with musicality and that they have a somewhat accurate sense of where they are (with some salient exceptions—as noted earlier).

When all the variables were placed in a regression, it accounted for over a quarter of the total variance in accuracy scores. Although that is helpful, it suggests that other variables have yet to be considered. Some possibilities from recent research include the ability to translate perception into action (e.g., Hutchins & Peretz, 2012; Pfordresher & Mantell, 2014), mental audiation of pitch using auditory imagery (e.g., Greenspon et al., 2017; Pfordresher & Halpern, 2013; Pruitt et al., 2019), and short-term memory for pitch (Greenspon & Pfordresher, 2019). None of these factors are included in the SSAP due to time demands, but they should be addressed in large public samples in future studies.

## Conclusion

This study represents an attempt to measure pitch accuracy in singing through a large sample using a standardized measure.[3] As such, it provides perhaps a more accurate accounting of singing accuracy in the general population and the effects of age, training, perceptual discrimination, and self-perception in assessing one's accuracy. Our results suggest that singing accuracy varies greatly in the general population, with a tendency toward accurate rather than inaccurate pitch matching, and that multiple factors contribute to singing ability. Singing accuracy thus may reflect a learned motor skill that involves the interaction among multiple perceptual, motoric, and cognitive functions (cf. Pfordresher et al., 2015). An important limitation of the present study stems from its correlational design. Future research will be designed to evaluate possible causal effects in randomized controlled trial designs.

### Acknowledgments

We thank Jacob Dubin for assistance in screening online recordings for recording quality and compliance with instructions.

### Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The development of the Seattle Singing Accuracy Protocol was made possible by a grant from the National Association for Music Education to Steven M. Demorest. The University of Washington hosted a colloquium in October 2013 that was instrumental to the conceptual development of this protocol. Peter Q. Pfordresher was supported by a Fulbright Canada Research Chair and NSF Grant BCS-1848930 during the development of this article.

### ORCID iD

Peter Q. Pfordresher https://orcid.org/0000-0002-7024-0785

### Supplemental Material

Figures S1 through S4 and Tables S1 and S2 are available in the online version of the article at https://doi.org/10.1177/0022429420951630.

### Notes

1. Although comparison of different measures is not central to the present article, it is worth noting that when data are analyzed in the same way as Pfordresher and Brown (2007), using mean signed pitch deviation scores with a criterion of ±100 cents, the frequency of singing classified as accurate goes up to 94%.

2. As shown in Table S1 in the supplementary files included with the online version of this article, Pearson correlation coefficients yielded result that were highly similar to the nonparametric Spearman's rho. Because Pearson correlations are related to multiple regression and hierarchical regression, reported later, we focus on these parametric measures in the main text.
3. Independent of our research, a similar effort to gather large-scale singing accuracy data was carried out in Australia, reported in an unpublished thesis (Tan, 2016).

## References

Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The Sung Performance Battery (SPB). *Frontiers in Psychology*, *4*, Article 714. https://doi.org/10.3389/fpsyg.2013.00714

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cooper, N. A. (1995). Children's singing accuracy as a function of grade level, gender, and individual versus unison singing. *Journal of Research in Music Education*, *43*, 222–231. https://doi.org/10.2307/3345637

Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority: Normality and "phenotypes" of poor singing. *Annals of the New York Academy of Sciences*, *1169*, 99–107. https://doi.org/10.1111/j.1749-6632.2009.04558.x

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

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*, 1917–1931.

Demorest, S. M. (2001). Pitch matching performance of junior high boys: A comparison of perception and production. *Bulletin of the Council for Research in Music Education*, *151*, 63–70.

Demorest, S. M., & Clements, A. (2007). Factors influencing the pitch matching of junior high boys. *Journal of Research in Music Education*, *55*, 190–203. https://doi.org/10.1177/002242940705500302

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

Demorest, S. D., Nichols, B. E., & Pfordresher, P. Q. (2018). The effect of focused instruction on young children's singing accuracy. *Psychology of Music*, *46*, 488–499. https://doi.org/10.1177/0305735617713120

Demorest, S. M., & Pfordresher, P. Q. (2015a). *Seattle Singing Accuracy Protocol – SSAP* [Measurement instrument]. https://ssap.music.northwestern.edu/

Demorest, S. M., & Pfordresher, P. Q. (2015b). Singing accuracy development from K–adult: A comparative study. *Music Perception*, *32*, 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*, 266–271. https://doi.org/10.1525/MP.2015.32.3.266

Hutchins, S. M., & 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*, 76–97. https://doi.org/10.1037/a0025064

Hyde, K. L., & Peretz, I. (2004). Brains that are out of tune but in time. *Psychological Science*, *15*, 356–360.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538.

Green, G. A. (1990). The effect of vocal modeling on pitch-matching accuracy of elementary schoolchildren. *Journal of Research in Music Education*, *38*, 225–231. https://doi.org/10.2307/3345186

Greenspon, E. B., & Pfordresher, P. Q. (2019). Pitch specific contributions of auditory imagery and auditory memory in vocal pitch imitation. *Attention, Perception & Psychophysics*, *81*, 2473–2481. https://doi.org/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*, *35*, 585–604. https://doi.org/10.1525/mp.2017.34.5.585

Loui, P., Demorest, S. M., Pfordresher, P. Q., & Iyer, J. (2015). Neurological and developmental approaches to poor pitch perception and production. *Annals of the New York Academy of Sciences*, *1337*, 263–271. https://doi.org/10.1111/nyas.12623

Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action-perception mismatch in tone-deafness. *Current Biology*, *18*, R331–R332.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*, *25*, 1795–1803. https://doi.org/10.1177/0956797614541990

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

Petzold, R. G. (1963). The development of auditory perception of musical sounds by children in the first six grades. *Journal of Research in Music Education*, *11*, 21–43. https://doi.org/10.2307/3344529

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness." *Music Perception*, *25*, 95–115. https://doi.org/10.1525/MP.2007.25.2.95

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

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. A. Russo, B. Ilari, & A. J. Cohen (Eds.), *Routledge Companion to Interdisciplinary Studies in Singing, Volume 1: Development* (pp. 322–333). Routledge.

Pfordresher, P. Q., Demorest, S. M., Dalla Bella, S., Hutchins, S., Loui, P., Rutkowski, J., & Welch, G. F. (2015). Theoretical perspectives on singing accuracy: An introduction to the special issue on singing accuracy (Part 1). *Music Perception*, *32*, 227–231. https://doi.org/10.1525/MP.2015.32.3.227

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

Pfordresher, P. Q., & Larrouy-Maestri, P. (2015). On drawing a line through the spectrogram: How do we understand deficits of vocal pitch imitation? *Frontiers in Human Neuroscience*, *9*, Article 271. https://doi.org/10.3389/fnhum.2015.00271

Pfordresher, P. Q., & Mantell, J. T. (2014). Singing with yourself: Evidence for an inverse modeling account of poor-pitch singing. *Cognitive Psychology*, *70*, 31–57. https://doi.org/10.1016/j.cogpsych.2013.12.005

Pruitt, T. A., Halpern, A. R., & Pfordresher, P. Q. (2019). Covert singing in anticipatory auditory imagery. *Psychophysiology*, *56*, Article e13297. https://doi.org/10.1111/psyp.13297

Sloboda, J. A., Wise, K. J., & Peretz, I. (2005). Quantifying tone deafness in the general population. *Annals of the New York Academy of Sciences*, *1060*, 255–261. https://doi.org/10.1196/annals.1360.018

Svec, C. L. (2018). The effects of instruction on the singing ability of children ages 5 to 11: A meta-analysis. *Psychology of Music*, *46*, 326–339. https://doi.org/10.1177/0305735617709920

Tan, Y. T. (2016). *The genetic basis of singing ability: A twin study* [Unpublished doctoral dissertation]. University of Melbourne.

Welch, G. F. (2006). Singing and vocal development. In G. McPherson (Ed.), *The child as musician* (pp. 311–329). Oxford University Press.

Welch, G. F., Sergeant, D. C., & White, P. J. (1997). Age, sex, and vocal task as factors in singing "in tune" during the first years of schooling. *Bulletin of the Council for Research in Music Education*, *133*, 153–160.

Wise, K. J., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined "tone deafness": Perception, singing performance and self-assessment. *Musicae Scientiae*, *12*, 3–26. https://doi.org/10.1177/102986490801200102

Yarbrough, C., Green, G., Benson, W., & Bowers, J. (1991). Inaccurate singers: An exploratory study of variables affecting pitch matching. *Bulletin of the Council for Research in Music Education*, *107*, 23–34.

## Author Biographies

**Peter Q. Pfordresher** is professor and chair in the Department of Psychology, University at Buffalo, State University of New York. His research focuses on the relationship between perception and action in the context of music, including the sensorimotor bases of poor-pitch singing and the interplay between melody and rhythm during perception and production.

**Steven M. Demorest** was professor of music in the Bienen School of Music, Northwestern University. He studied singing development throughout the life span, choral musicianship, music participation, and cross-cultural music cognition. Professor Demorest tragically passed away shortly after the initial submission of this article. The first author did his best to address comments in the review process in a way that he hopes would have been acceptable to his collaborator. Professor Demorest made great contributions to music education, including the study of singing accuracy, and this article is dedicated to his memory.

Submitted September 4, 2019; accepted July 16, 2020.