# Singing Accuracy Development from K-Adult: A Comparative Study

Steven M. Demorest
*University of Washington & Northwestern University*

Peter Q. Pfordresher
*University at Buffalo, State University of New York*

**The development of singing accuracy, and** the relative role of training versus maturation, is a central issue for both music educators and those within music cognition. Although various studies have focused on singing accuracy in different age groups, to date we know of no data sets that maintain the consistency in recruitment, methodology, and measurement that is necessary to make direct comparisons. We report analyses of three data sets that meet these criteria: two groups of children (kindergarten, middle school), and one group of adults (college aged). The data were collected at different times, but used a similar set of tasks and identical scoring procedures. Results indicate considerable improvement in accuracy from kindergarten to late elementary that dramatically reverses such that college students perform at the level of kindergartners. It appears singing accuracy may be related to variables involving singing experience rather than general development, and singing skill could decline over time if not maintained through engagement. A secondary purpose was to explore the efficacy of acoustic scoring for some singing tasks and how well it mimics human judgments of accuracy. The acoustic scoring procedure was highly correlated with expert judgment and could provide a standard approach to scoring that is largely automated. We discuss the potential benefits of a more unified approach to measuring singing accuracy and suggest future research that includes children, adolescents and adults in the sample.

*Received: June 9, 2014, accepted September 14, 2014.*

**Key words:** music cognition, music education, singing accuracy, singing development, tone deafness

Singing ability is fundamental to developing musicianship and to the developing view of oneself as a musical being (Welch, 2006). As numerous studies have documented, the label "tone deaf" can have devastating consequences for children's views of their musicality. This poor musical self-image can shape future engagement in music, and the negative experiences of childhood are remembered vividly well into adulthood (Abril, 2007; Sloboda, Wise, & Peretz, 2005; Whidden, 2010). Consequently, it would be important for music educators to have a clear understanding of how such ability develops and can be nurtured.

Research in music education has studied children's singing development from the time just before children enter schooling through the end of their elementary years (age 11-12). This research has largely focused on how to improve singing instruction in general or how to deal with poor singers, those who don't sing on pitch, in particular (see Philips & Doneski, 2012 & Welch, 2006, for reviews). The research has yielded some important findings about children's accuracy as it relates to register development (cf. Hedden, 2012, Rutkowski, 1996; Welch, Sergeant, & White, 1997), perceptual variables (Demorest, 2001; Demorest & Clements, 2007; Geringer, 1983), and the effect of various instructional strategies (Apfelstadt, 1984; Phillips & Aitchison, 1997; Welch et al., 2009), but it has been hampered by a lack of uniformity in how singing accuracy has been defined and measured. Because each study has created its own set of tasks and scoring systems, results between studies can be very difficult to compare. Consequently, a large-scale picture of children's singing competency at various ages and stages of development has been more illusive (though see Welch et al., 2009 for data on a larger population).

A similar problem is present in the adult literature. In the last decade there has been an increased interest in poor pitch singing in cognitive psychology as a possible component of broader communication deficits such as amusia (Berkowska & Dalla Bella, 2009; Dalla Bella, Giguère, & Peretz, 2009; Loui, Guenther, Mathys, & Schlaug, 2008). This research offers some estimates of singing accuracy among adult populations. However, differences in sampling procedures, tasks, and measurements makes it difficult to compare estimates or determine the validity of any single estimate. For instance, different studies focus on group differences based on singing performance (e.g., Dalla Bella, Giguère, & Peretz, 2007; Pfordresher & Brown, 2007), pitch perception ability (e.g., Hutchins & Peretz 2012; Dalla Bella

*Music Perception*, volume 32, issue 3, pp. 293–302, issn 0730-7829, electronic issn 1533-8312. © 2015 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's rights and permissions website, http://www.ucpressjournals.com/reprintinfo.asp. DOI: 10.1525/mp.2015.32.3.293

et al., 2009), or self-identification as a "tone deaf" individual (Wise & Sloboda, 2008). Second, tasks also differ greatly and have included single pitch matching (e.g., Berkowska & Dalla Bella, 2013; Hutchins & Peretz, 2012), imitation of short unfamiliar melodies (e.g., Berkowska & Dalla Bella, 2013, Pfordresher & Brown, 2007; Wise & Sloboda, 2008), or singing of familiar songs from memory (e.g., Berkowska & Dalla Bella, 2013; Dalla Bella et al., 2007).

Finally, criterion measures used to divide participants into "accurate" versus "inaccurate" (or poor-pitch) groups have varied. Researchers that have used a criterion of +/- 100 cents around target pitches have found that 85-90% of the population sing "accurately" and that most people were being to harsh in their self-assessments of accuracy (Dalla Bella et al., 2007; Pfordresher & Brown, 2007). However, such a criterion is arguably too extreme, given that a singer attempting to sing C-E could sing C-Eb or C-F slightly sharp or flat respectively, and be deemed accurate. For the purposes of the cited research, such criteria were useful for identifying outlier populations who might be considered to have a singing "disorder" (Berkowska & Dalla Bella, 2009). By contrast, in a study by Hutchins and Peretz (2012), when the criterion of ±50 cents was used only 38% of untrained singers could match at least 90% of the pitches and 47% of the sample failed to match even 50% of the pitches. Interestingly this matches more closely with the prevalence of self-reported poor pitch singing (Pfordresher & Brown, 2007) and with measures of singing precision (Berkowska & Dalla Bella, 2013; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). When the criterion is mean deviation of ±50 cents across all tasks rather than pitch-by-pitch, then 58% of untrained singers can match pitches and intervals accurately and 78% can match patterns accurately. Along similar lines, Pfordresher and colleagues (2010) found that 31% of singers sang outside this boundary on average, compared to 13% of the sample who exceeded the 100-cent criterion used by Pfordresher and Brown (2007).

The issue of what criterion best functions as a criterion for poor singing is of great importance to the field (see Dalla Bella this volume for more on this topic). However, for the present purposes what matters most is consistency in the use of any given criterion. Comparisons across studies cannot be effective when the dividing line changes from study to study. Certainly this is a problematic issue for the adult literature. But we argue that inconsistency is a more damaging issue when comparing different age groups, as one risks confounding tasks and measures used to assess accuracy with developmental trends that influence accuracy.

A simple solution for this problem, laying aside for the moment the validity of any given task or measure, is to compare the performance of different age groups on a common set of tasks and measures. Unfortunately, such comparisons in the past literature are further hampered by the lack of singing studies that include both children and adults as participants. We have a number of studies of poor pitch singing in adult populations, but few if any of those studies include participants under 18 years of age. The numerous studies of singing accuracy in music education have focused almost exclusively on participants from kindergarten to grade 5 or 6 (ages 5-11). While there are a few singing studies that include adolescent populations (Demorest, 2001; Demorest & Clements, 2007; Price, Yarbrough, Jones, & Moore, 1994; Yarbrough, Green, Benson, & Bowers, 1991) they tend to focus on children participating in music instruction or sampled *a priori* for poor-pitch singing rather than the general school population.

The common view of children's singing development is that accuracy improves with age (Petzold, 1963; Roberts & Davies, 1975; Welch et al., 1997; Yarbrough et al., 1991), but most of the participants in these developmental studies were also receiving regular music instruction through school. Consequently, it is difficult with this population to separate maturation from increased singing experience and we do have evidence that children with more musical experience at a given age can outperform those with less experience (Apfelstadt, 1984; Nichols, 2013; Petzold, 1963, Tsang, Friendly, & Trainor, 2011). At the point that music instruction becomes elective (grade 6 or 7 in most districts), we have little or no data on children's singing accuracy performance. This is significant because in the United States by eighth grade only 34% of the general population participates in elective music instruction (Keiper, Sandene, Persky, & Kuang, 2009) and that number declines as children move toward graduation from high school.

The present study is an attempt to address some of the gaps in our knowledge by comparing data from three separate studies of different age groups that used similar procedures for measuring singing accuracy. The purpose of this study was to compare directly the singing accuracy performance of two groups of children and one group of adults from three independent investigations of singing accuracy. The data for these studies were collected at different times, but using a similar methodology and set of tasks that provided a relatively direct comparison of singing performance using identical scoring procedures. A secondary purpose is to explore the efficacy of acoustic scoring for some singing tasks and how well it mimics human judgments of accuracy. Such an analysis can help us begin to reconcile the different approaches to scoring used across a variety of studies. The results should contribute to our understanding of singing development through the lifespan and may illustrate the potential value of a more standardized approach to measurement and scoring. The research questions were:

1. Are there differences in singing accuracy between children of different ages and adults on matching and song singing tasks?
2. What is the relationship between scores derived from human judgments and those generated by an acoustic analysis?

## Method

### PARTICIPANTS

The data for this study come from three separate investigations. The adult data ($n = 78$) were taken from the normal feedback condition (Experiment 1) of an investigation of poor pitch singing by Pfordresher & Brown (2007). The 6th grade data ($n = 55$) were taken from a study investigating the relationship of singing accuracy to student's views of themselves as musicians and their subsequent participation in elective music (Demorest, Pfordresher, & Kelley, 2014) and the kindergarten sample ($n = 77$) was taken from a study looking at the impact of daily singing instruction on children's singing accuracy (Demorest, Nichols, & Pfordresher, 2014).

### MATERIALS AND CONDITIONS

While the three studies investigated distinct phenomena, they were similar in how they measured singing accuracy performance. All three used procedures based on Pfordresher and Brown (2007) in which matching tasks were presented in three contexts:

1. *Single pitch:* Hearing four repetitions of a single pitch and then singing back the whole sequence (e.g., G-G-G-G)
2. *Interval pitch:* Hearing four pitches that formed a single interval (e.g., G-G-E-E) then singing back the whole sequence.
3. *Pattern pitch:* Hearing four pitches that form a pattern (e.g., G-E-C-G), and then singing back the whole sequence.

The purpose in using these three tasks was to vary sequential complexity with respect to the number of different pitches, while maintaining constant "list length." Previous research has found that singers of different skill levels respond differently to different pitch contexts. For example, while the pattern task offers a more complex pitch sequence than the single pitch, it also provides a richer tonal context that has been shown to help poor pitch singers (Demorest & Clements, 2007; Pfordresher & Brown, 2007). All three tasks involved a "call-and-response" procedure in which participants first listened to an auditory target stimulus and then repeated it immediately thereafter. For the adults, target stimuli were produced using the singing-voice synthesis software package "Vocaloid," whereas for the children we used recordings of a female singer that were checked for accuracy. All stimuli were presented at a rate of 1 note per second.

The adult sample performed all these tasks under three feedback conditions: normal auditory feedback, masked auditory feedback, or chorused feedback. For the present comparisons, only normal feedback trials were included. In addition, adults repeated each trial two times in succession, whereas other groups produced each trial once. Thus, we included only the first repetition of each adult trial in the present comparison.

In addition to the matching tasks, all three studies gave participants two opportunities to sing a familiar song from memory. The purpose of this task was to determine how well results based on the imitation of short, novel melodies generalize to longer well-known melodies. Also, whereas the imitation of melodies is a more controlled task than song singing (all participants are equally familiar, errors are unambiguously associated with imitation rather than recall, etc.), song singing has the value of greater ecological validity. For the grade six and adult groups the familiar song was "Happy Birthday," while for the kindergarten group it was "Twinkle, Twinkle Little Star." All studies presented stimuli in a comfortable range for the participants, all limited the range of the matching task to the range of a fifth, and all used vocal timbre as the stimulus for the matching task.[1]

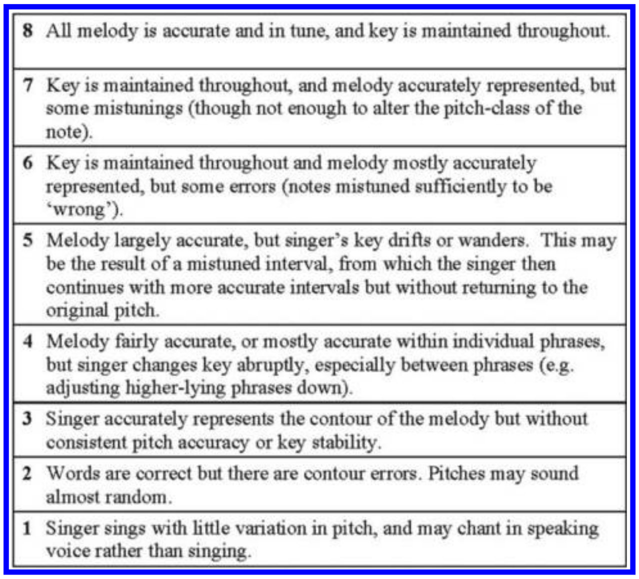
[1] Participants 1-20 in Pfordresher and Brown (2007) all matched a synthesized male voice so female participants had to transpose. The remaining participants heard the stimuli in the proper octave. For the adult and kindergarten samples the range was a fifth from C-G, which fits both the child voice and adult chest register. For the Grade 6 sample, the test register was tailored to each voice because of the presence of some changing voices. This was done by having the students sing "Happy Birthday" spontaneously and hold a tone in their comfortable register. The starting pitch of the song and the held tone were used to place them into a fifth that was most comfortable. Where there were uncertainties, the researchers had them sing more pitches to verify the test register.

### ANALYSES

In addition to comparing singing performance across ages on similar tasks, we were able to apply the same scoring criteria to all of the data. Performance on the three matching tasks was scored acoustically using a procedure adapted from Pfordresher and Brown (2007). The acoustic scoring procedure yields a mean cent deviation and an error rate for each singing task. Average values of produced F0 (extracted from the audio signal via autocorrelation, and checked for errors) are converted into cents relative to the lowest imitated pitch cross sequences (C3 for males, C4 for females). Differences between sung pitches and target pitches (also represented as cents relative to a base C) constitute *pitch deviation* scores. These scores were based on single notes rather than intervals given that the goal of matching tasks is to match absolute pitch. We further converted pitch deviation scores into acoustically derived *error rates*, by coding every pitch deviation outside a window of +/- 50 cents as an error (score of 1), with other sung pitches being coded as accurate (score of 0). Although these measures both reflect pitch deviations, they do so in different ways with pitch deviations being sensitive to the magnitude of discrepancies from target pitches and error rates measuring the frequency of discrepancies that are above threshold. For example, if a participant sings a four-note melody with deviations of -30 cents (flat), +40 cents, -60 cents, and +200 cents, the resulting absolute pitch deviation score will be 82.5 (influenced by every deviation), but the error rate will be 50% (reflecting the above threshold deviation for the final two notes). All analyses were performed using pitch extraction algorithms based on autocorrelation that were evaluated for accuracy. Note segmentation in Pfordresher and Brown (2007) was carried out manually, whereas for other data sets an automated Matlab algorithm was used to identify note onsets that were evaluated visually by the experimenter and corrected as needed.

As mentioned before, one goal of the present research was to evaluate the degree to which acoustic scores correlate with subjective evaluations of accuracy. For the matching tasks, we made these comparisons directly by having expert judges score each attempt and award one point for each correctly sung pitch. The points awarded by the human experts corresponds to the error rate measure (missed pitches) of the acoustic scoring allowing for a critical test of the similarity between the two approaches.

Singing a song from memory provides a particular challenge for acoustic scoring because each syllable must first be parsed to yield an average pitch, a challenge with the variability of text attacks and cutoffs. Then scores are calculated based not on the expected pitch, but on the expected interval between two pitches. This is done to allow for moment-to-moment tuning adjustments because less-experienced singers frequently drift or modulate during song singing (Berkowska & Dalla Bella, 2013). Parsing the song into component syllables and then extracting a mean frequency is incredibly time consuming and still doesn't completely represent concepts like "sense of tonality" for an excerpt. For this comparative study song-singing tasks were scored by expert judges using a rating scale developed by Wise and Sloboda (2008). The scale, shown in Figure 1, represents a sequential approach to accuracy that moves from chanting on text to contour to interval and then to key, a sequence that has been suggested in a variety of singing development literature (cf. Welch, 2006). One challenge of using such a scale is that within a song, a participant can sometimes execute one phrase perfectly, while missing a number of pitches on an adjacent phrase, so the rater has to try and extrapolate a kind of gestalt score for the attempt. Despite its challenges, the rating scale has been demonstrated to be reliable (Wise & Sloboda, 2008).



| | |
|---|---|
| 8 | All melody is accurate and in tune, and key is maintained throughout. |
| 7 | Key is maintained throughout, and melody accurately represented, but some mistunings (though not enough to alter the pitch-class of the note). |
| 6 | Key is maintained throughout and melody mostly accurately represented, but some errors (notes mistuned sufficiently to be 'wrong'). |
| 5 | Melody largely accurate, but singer's key drifts or wanders. This may be the result of a mistuned interval, from which the singer then continues with more accurate intervals but without returning to the original pitch. |
| 4 | Melody fairly accurate, or mostly accurate within individual phrases, but singer changes key abruptly, especially between phrases (e.g. adjusting higher-lying phrases down). |
| 3 | Singer accurately represents the contour of the melody but without consistent pitch accuracy or key stability. |
| 2 | Words are correct but there are contour errors. Pitches may sound almost random. |
| 1 | Singer sings with little variation in pitch, and may chant in speaking voice rather than singing. |

FIGURE 1. The rating scale from Wise and Sloboda (2008) used to judge song singing accuracy.

## Results

The first research question dealt with performance differences by age. Each sample group generated two acoustic scores for each of the matching tasks, a mean

**TABLE 1.** *Error Rates (in Percent) and Mean Deviation Scores (in Cents) by Age Group Across the Three Matching Tasks.*

| | | Single Pitch | | Interval Pitch | | Pattern Pitch | | MEAN | |
|---|---|---|---|---|---|---|---|---|---|
| Age Group | *N* | %Error | Deviation | %Error | Deviation | %Error | Deviation | %Error | Deviation |
| Kindergarten | 77 | 45.75 | 109.10 | 44.09 | 88.75 | 56.13 | 130.47 | 48.65 | 109.44 |
| Grade 6 | 55 | 18.50 | 47.42 | 17.92 | 48.95 | 25.30 | 56.46 | 20.53 | 50.94 |
| Adult | 78 | 26.15 | 66.18 | 39.62 | 87.28 | 53.65 | 123.58 | 41.76 | 92.35 |
| MEAN | 210 | 31.34 | 77.00 | 35.58 | 77.78 | 47.13 | 108.53 | 38.73 | 84.24 |

deviation score and an error rate score. Both scores are shown as a function of age group (kindergarten, 6th grade, adult) and task (single pitch, interval pitch, pattern pitch) in Table 1. A 3 x 3 repeated measures ANOVA with these factors yielded a significant main effect of age for both dependent variables: error rates, $F(2, 207) = 20.97$, $p < .001$, $h^{2}_{\mathrm{p}}= 0.1$; absolute pitch deviation, $F(2, 207) =8.93$, $p < .001$, $h^{2}_{\mathrm{p}}= 0.08$. Scheffè post hoc comparisons for both measures revealed significant differences between the 6th grade participants and the other two groups, but no difference between adult and kindergartners overall.

There was a main effect of task for each dependent variable: error rates, $F(2, 206) = 70.96$, $p < .001$, $h^{2}_{\mathrm{p}} = 0.26$; absolute pitch deviation, $F(2, 206) = 29.45$, $p < .001$, $h^{2}_{\mathrm{p}} = 0.12$. Pairwise comparisons on error rates indicated a significant difference in performance among all the tasks by complexity with single pitch being the easiest, followed by interval pitch and then pattern. For absolute pitch deviations there was one exception in that the difference between single and interval pitch conditions was not significant (as can be seen in Table 1, their means were within one cent.

There was also a significant age by task interaction for both measures (Wilks Lambda): error rates, $F(4, 412) = 8.85$, $p < .001$, $h^{2}_{\mathrm{p}} = 0.08$; absolute pitch deviation, $F(4, 412) = 7.09$, $p < .001$, $h^{2}_{\mathrm{p}} = 0.06$. Figure 2 plots the interaction for the measure error rates; a similar pattern was found for pitch deviation scores. We focus on this measure given its relationship to expert ratings. Pairwise comparisons (Scheffè corrected) were run on all pairwise group differences within each condition. Whereas 6th graders exhibited better performance than kindergartners on all tasks, college student performance exceeded kindergarteners only on the single-pitch imitation task. Likewise, 6th graders performed better than college students on interval and pattern tasks, but did not differ on the single-pitch task. Thus, overall developmental gains (though not beyond 6th grade) were seen in the "simple" single-pitch task, whereas gains were found in the other tasks that, somewhat surprisingly, reversed from 6th grade to college.

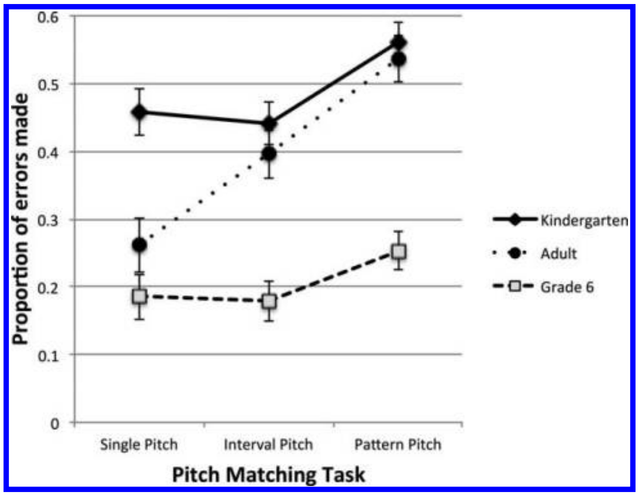


FIGURE 2. Mean proportional error rate by matching task across the three age groups. Error bars display 1 standard error of the mean.

All participants were also asked to make two attempts to sing a familiar song from memory. Both song-singing attempts for each participant in the three samples were scored by two independent judges with an overall inter-judge reliability of $r = .83$. Those four scores were then averaged to yield a single song-singing score per participant. Figure 3 illustrates the difference in scores for singing a familiar song by age group. There was a significant main effect of age, $F(2, 203) = 4.11$, $p < .05$, $h^{2}_{\mathrm{p}}= 0.04$, and pairwise comparisons indicated a significant difference only between the kindergarten and adult groups. There was a significant negative correlation between the matching error rate and singing from memory score, $r(204) = -.44$, $p < .001$.

Question 2 dealt with how well acoustic scores might relate to the judgment of experts on pitch accuracy. To test this, we had two judges score the matching performance of only the kindergarten and adult samples by awarding one point for each correct pitch in a four-note response to yield a score for correct pitches per participant. The reliability of this scoring system was $r = .86$ for adult data and $r = .90$ for kindergarten data. As with the acoustic data, because the number of items between

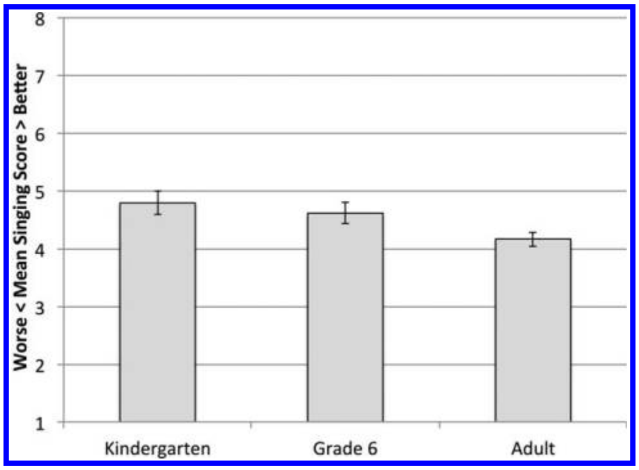


FIGURE 3. Mean rating for song-singing across the three age groups. Error bars display 1 standard error of the mean.

the two studies was not identical these scores were expressed as percent correct. To make the comparison between the two measures easier to see, we transformed proportional error rate scores from acoustic analyses into percent correct (Table 2). The overall correlation between error rate acoustic scoring and expert judgment for the matching tasks was significant and positive, $r(152) = .92$, $p < .001$ with acoustic scores consistently more conservative than expert judgment at awarding a correct pitch.[2] The correlation between expert judgments and mean cent deviation was $r(152) = .87$, $p < .001$.

## Discussion

The analyses presented here argue against the hypothesis that singing accuracy improves with age. Though considerable improvement was found among the two groups of school-aged children (kindergarten and grade 6), from grade 6 to college we observed an almost complete reversal of these gains. Moreover, grade 6 was the only group to exhibit modal performance at a level of accuracy of over 90% of notes sung correctly. What do these results mean?

One possibility, which we consider unlikely, is that the developmental trajectory of singing is nonlinear and reaches a peak just before adolescence. Though some traits show such a nonlinear trend, peak ability more often occurs during young adulthood, and this seems most plausible for singing. Rather, we interpret these data in light of environmental constraints. Singing accuracy may be related to variables involving singing experience rather than general development, and singing skill can decline over time if not maintained through engagement. In the elementary grades, singing engagement and training are often provided through mandatory music instruction and may therefore correspond with age. After grade 5 or 6 in the United States, a much smaller percentage of the general population continue to participate in singing or in music of any kind (Keiper et al., 2009) so singing accuracy may decline for those who don't continue. This would account for the lower scores of our adult sample on the matching task. Previous research has documented experience-related differences in singing accuracy in both younger (Nichols, 2013) and adult populations (Hutchins & Peretz, 2012; Larrouy-Maestri & Morsomme, 2014), which support this idea. However, the comparisons with adults are usually separated into untrained and highly trained, making it difficult to gauge the relationship of performance to singing engagement in a continuous way. Also, experience can be confounded with ability in those studies because individuals who choose to remain engaged in music report higher musical self-concepts at a young age and have likely experienced more early success (Clements, 2002; Mizener, 1993). Ideally this proposition could be tested with an experiment that trains adults who do not evidence any severe deficits in an intensive singing program and measure how much their accuracy improves over time. Programs like the "Can't Sing Choirs" in Britain (Lane, 2011) and narrative investigations with self-labeled non-singers (Whidden, 2010) have reported great success.

Although the present comparisons are more direct than those published in the past, they were nevertheless not all designed for the same purpose and were thus not identical in all respects. A particular issue we address now has to do with the recruitment strategy. Whereas child samples were drawn randomly from respective grades, the earlier study with college students adopted a more focused sampling strategy. In particular, for half the adult sample Pfordresher and Brown (2007) preselected only those participants who were self-described poor pitch singers. Somewhat surprisingly (though fortunately for our present purposes), this selection strategy had no influence on levels of accuracy in the two samples, and thus may not be a confounding factor here. Another element of the selection strategy in Pfordresher and Brown was that they sought out musically untrained participants. However, if we compare adult singing performance to only the 28 6th graders who reported no formal training there was still

[2] The one exception was the acoustic score for the adults matching a single pitch.

**TABLE 2.** *Mean Percent Correct as Determined by Judges and Acoustic Analysis Across the Three Matching Tasks.*

| | | Single Pitch | | Interval Pitch | | Pattern Pitch | | MEAN | |
|---|---|---|---|---|---|---|---|---|---|
| Age Group | *N* | Judge | Acoustic | Judge | Acoustic | Judge | Acoustic | Judge | Acoustic |
| Kindergarten | 77 | 63.51 | 54.25 | 73.21 | 55.91 | 62.11 | 43.87 | 66.28 | 51.34 |
| Adult | 77 | 69.42 | 73.85 | 60.92 | 60.38 | 76.88 | 46.35 | 69.07 | 60.19 |
| MEAN | 154 | 66.46 | 64.05 | 67.07 | 58.15 | 69.50 | 45.11 | 67.68 | 55.77 |

a significant age main effect for singing accuracy across all three tasks, $F(2, 179) = 10.71$, $p < .001$, partial $h^2 = 0.11$. The musically inexperienced 6th graders, who were receiving mandatory music but nothing else, were still significantly better than their inexperienced adult counterparts (mean error rate 23.2% as opposed to 20.5% for the entire sample).

Moreover, we do think that the primary causal factor in age differences is experience, though not necessarily experience that is specific to formal training in a given instrument. Rather, we propose that the advantages for children seen here result from exposure to general musical activities that happen in the early school years. In this context, performance in the single-pitch matching task may be informative. This was the only task that suggested an improvement from kindergarten to college years (though, interestingly, no improvement from grade 6 to college). Though this task is most prevalent in the experimental literature on singing, it is arguably the most "nonmusical" singing task one can construct. It is thus perhaps not surprising that this task appeared to be less tied to involvement in general music training than the other tasks.

Another test of the typicality of our adult sample is to compare their performance to other studies of adult singing performance. For example, Hutchins and Peretz (2012) used a ±50 cent acoustic scoring criteria and report the singing performance of 53 nonmusician adult singers (less than one year of formal training) based on percent of correct pitches. They found that 38% of their sample matched 90% of the pitches correctly. This is compared to 45% of sixth graders in our study and 5% of adults. As detailed previously in this article, however, differences in methodology make it difficult to treat the findings across these studies as equivalent. In the earlier study, participants matched only single pitches, the criterion pitch sounded continuously (except when they were attempting to match), and participants could take as long as they wanted to achieve their best match, whereas our participants heard the pitch stimulus once (single, interval, or pattern) and then echoed back in a single attempt with no further feedback. If we compare only our adults' single pitch scores to those of previous studies we see more similarities. The adults in Pfordresher and Brown (2007) on average matched 74% of the single pitches correctly compared to 59% of nonmusician adults in Hutchins and Peretz (Experiment 1 & 2) and 58% of occasional singers in Berkowska and Dalla Bella (2013, Task 1). This suggests that our adult sample, while untrained, was similar to or better than other adult samples thus lending further validity to their representativeness.

Unlike the matching task, the song-singing results indicate that the youngest sample was the most accurate. This is in contrast to other researchers that have found singing a familiar song to be a more difficult task for young children (Guerrini, 2006; Welch et al., 1997). It is likely that the higher scores for kindergartners in this study were due not to differences in skill but to differences in song material. "Twinkle, Twinkle Little Star" has a total range of a sixth and predominantly stepwise motion that begins and ends on the tonic pitch. "Happy Birthday," a popular song for singing research because of its ubiquity, is actually quite challenging to sing. It begins on the dominant and spans an octave with a number of intervallic leaps of differing length. As Wise and Sloboda (2008) observe "listening to any gathering of people singing 'Happy Birthday' sometimes makes one wonder if everyone learns it accurately in the first place" (p. 20). If we compare only the grade 6 to adult scores, where the same song is used, we do see a significant decline in performance that matches what is found for the pitch matching data, $t(126) = 2.48$, $p < .05$, though the decline is smaller in magnitude. This is similar to other researchers that have found that while performance on interval and pattern matching tasks can show improvement over time, performance on singing songs from memory may not change as much (Apfelstadt, 1984; Demorest et al., 2014; Roberts & Davies, 1975; Welch et al., 1997).

Acoustic scoring was strongly correlated with human judgment using both the error rate and mean cent deviation scoring. The highest correlation was found when the acoustic data were used to determine a pitch-by-pitch error rate using a ±50 cent accuracy criterion. For research that seeks to replicate human judgment but with the precision of an acoustic scoring procedure, this would seem to be the best choice. Another benefit of this

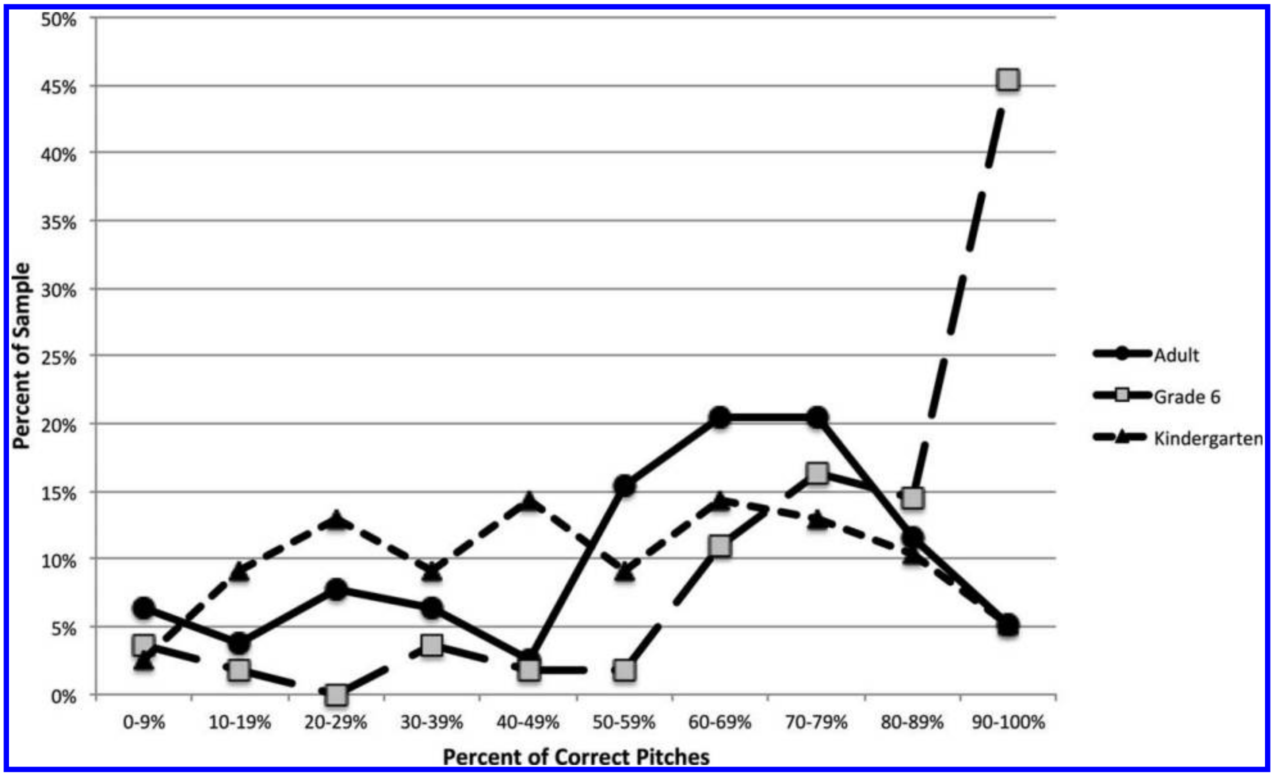


**FIGURE 4.** Singing accuracy by age expressed in total percent of correct pitches.

approach to scoring is that it can translate nicely into population norms. For example, if we wish to view these data from the standpoint of the prevalence of singing accuracy in the general population, we can transform the proportional error rate data into percent of correct pitches performed by age group. Figure 4 charts performance by age group across all three matching conditions (single, interval, and pattern). As we can see, when more complex tasks are included and scoring criteria are correlated with expert perception, the prevalence of accurate singing in the adult population drops considerably from earlier estimates. These estimates are more in line with recent studies reporting similarly strict criteria (Berkowska & Dalla Bella, 2013; Hutchins & Peretz, 2012; Pfordresher et al., 2010) and support the notion that people's self-evaluations of poor singing may be most closely related to a sense of how many notes one sings incorrectly, as opposed to one's average discrepancy from sung target notes.

## Conclusion

Singing is considered a crucial component of overall musicianship for young children (Music Educators National Conference [MENC], 1994). Consequently, music educators are concerned with helping children become accurate and confident singers by the time they leave grade school (Philips & Doneski, 2012, Welch, 2006). One tacit assumption in much of the research on singing development is that accuracy, once acquired, does not regress into inaccuracy. If we assume that our adults in this study are not systematically different from our sixth graders in amount of music training or aptitude they possessed earlier in life, then it is likely that these adults were more accurate when they were in grade school and their accuracy degraded over time with lack of use.[3] The only way to test such an hypothesis would be to perform a longitudinal or cross-sectional comparative study that traced singing performance from grade 5 (or younger) to adulthood while recording the amount and type of musical experiences (formal and informal) in which the participants engaged. Such a study would fill an important gap in our knowledge about singing development through the lifespan.

[3] For the males in the sample, it is also possible that they lost some accuracy if they did not continue singing as their voices mutated into mature male registers.

Anecdotally, there are a number of individuals who seem to possess a fine singing voice, even if they didn't participate in secondary school music or formal instruction. Such individuals illustrate the challenge of defining and recording vocal "practice," which could range from more formal training to regularly singing in the shower, in the car, with family, or in church. While few of these experiences would be deemed "training" by conservatory standards, they could have a significant impact on an individual's ability to maintain and even develop their singing skill.

The results of this study illustrate the potential power of a standard approach for understanding the development of singing performance across multiple age groups. Elsewhere in this volume we offer a suggestion for one such standard measure, which is based on combining measures previously used by authors in this volume (Berkowska & Dalla Bella, 2013; Pfordresher & Brown, 2007). Earlier studies have suggested that extreme poor pitch singing of the type that may reveal a more serious audio-motor deficit is a relatively rare phenomenon (Dalla Bella et al., 2007; Pfordresher & Brown, 2007; Wise & Sloboda, 2008), yet a majority of adults still view themselves as poor singers. Language from studies investigating self-perceptions of singing skill suggest that people tend to view singing as a fixed characteristic like "talent" rather than a temporary condition that could be improved at any time (Abril, 2007; Whidden, 2010). It may be that accurate singing is a musical skill more akin to playing the trumpet. Nobody expects adults who haven't picked up their trumpet since 5th grade to play with any skill, but we think that we either "have" or "don't have" a singing voice. Music educators need to provide singing experiences for all children K-12 that can allow them to participate and continue to engage in singing. This may point to reforming educational policies that limit musical offerings after a certain age and expanding the role of participatory singing in our culture. While it is important for teachers to help students improve their singing skill at a young age, it may be even more important to help them believe that accurate singing is attainable through practice even when they don't experience success early on.

## Author Note

This research was supported in part by NSF grant BCS-1256964, the Ruth Sutton Waters Endowed Professorship in Music, Bobbette Koon Endowment in Music Therapy, AIRS: Advancing Interdisciplinary Research in Singing and a NAfME/SRME research grant.

*Correspondence concerning this article should be addressed to* Steven M. Demorest at Bienen School of Music, Northwestern University, 711 Elgin Road, Evanston, IL 60208 or to Peter Q. Pfordresher, Department of Psychology, 355 Park Hall, University at Buffalo, Buffalo, NY 14260. E-mail: sdemorest@northwestern.edu or pqp@buffalo.edu

## References

Abril, C. R. (2007). I have a voice but I just can't sing: A narrative investigation of singing and social anxiety. *Music Education Research, 9,* 1-15. doi:10.1080/14613800601127494

Appelstadt, H. (1984). Effects of melodic perception instruction on pitch discrimination and vocal accuracy of kindergarten children. *Journal of Research in Music Education, 32,* 15-24.

Berkowska, M., & Dalla Bella, S. (2009). Acquired and congenital disorders of sung performance: A review. *Advances in Cognitive Psychology, 5,* 69-83. doi:10.2478/v10053-008-0068-2

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

Clements, A. (2002). *The importance of selected variables in predicting student participation in junior high choir* (Unpublished doctoral dissertation). University of Washington, Seattle, WA. Retrieved from ProQuest Digital Dissertations. (AAT 3062930)

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.

Dalla Bella, S., Giguère, J-F., & Peretz, I. (2009) Singing in congenital amusia. *Journal of the Acoustical Society of America, 126,* 414-424.

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

Demorest, S. M., Nichols, B., & Pfordresher, P.Q. (2014). *The effect of focused instruction on kindergartener's singing accuracy.* Manuscript in preparation.

Demorest, S. M., Pfordresher, P. Q., & Kelley, J. F. (2014, April). *School music participation: Exploring the role of students' self-concept and singing ability.* Paper presented at the National Association for Music Education National Conference. St. Louis, MO.

GERINGER, J. M. (1983). The relationship of pitch-matching and pitch discrimination abilities of preschool and fourth grade students. *Journal of Research in Music Education, 31*, 93-99.

GUERRINI. (2006). The developing singer: Comparing the singing accuracy of elementary students on three selected vocal tasks. *Bulletin of the Council for Research in Music Education,* 167, 21-31.

HEDDEN, D. (2012). An overview of existing research about children's singing and the implications for teaching children to sing. *Update: Applications of Research in Music Education, 30*(2), 52-62. doi: 10.1177/8755123312438516

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.

KEIPER, S., SANDENE, B. A., PERSKY, H. R., & KUANG, M. (2009). *The nation's report card: Arts 2008 music and visual arts* (NCES 2009–488). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

LANE, M. (2011). Can the tone deaf learn to sing? *BBC News Magazine* (January 10). Retrieved from http://www.bbc.com/news/magazine-12127843

LARROUY-MAESTRI, P., & MORSOMME, D. (2014). Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logopedics Phoniatrics Vocology, 39*, 11-18.

LOUI, P., GUENTHER, F. H., MATHYS, C., & SCHLAUG, G. (2008). Action–perception mismatch in tone-deafness. *Current Biology, 18*(8), R331-R332.

MUSIC EDUCATORS NATIONAL CONFERENCE. (1994). *National Standards for Arts Education,* Reston, VA: Author.

MIZENER, C. (1993). Attitudes of children toward singing and choir participation and assessed singing skill. *Journal of Research in Music Education. 41,* 233-245.

NICHOLS, B. E. (2013). *Task-based variability in children's singing accuracy* (Unpublished doctoral dissertation). University of Washington, Seattle, WA.

PETZOLD, R. (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.

PFORDRESHER, P. Q., & BROWN, S. (2007). Poor-pitch singing in the absence of "tone deafness." *Music Perception, 25,* 95-115. doi:10.1525/mp.2007.25.2.95

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*, 2182-2190.

PHILLIPS, K. H., & AITCHISON, R. E. (1997). Effects of psychomotor instruction on elementary general music students' singing performance. *Journal of Research in Music Education, 45*, 185-196.

PHILIPS, K. H., & DONESKI, S. M. (2012). Research on elementary and secondary school singing. In R. Colwell & P. Webster (Eds.), *MENC handbook of research on music learning: Volume 2 applications* (pp. 176-232) New York: Oxford University Press.

PRICE, H. E., YARBROUGH, C., JONES, M., & MOORE, R. S. (1994). Effects of male timbre, falsetto, and sine-wave models on interval matching by inaccurate singers. *Journal of Research in Music Education, 42,* 269-284.

ROBERTS, E., & DAVIES, A. D. (1975). Poor pitch singing: Response of monotone singers to a program of remedial training. *Journal of Research in Music Education, 23*, 227-239.

RUTKOWSKI, J. (1996). The effectiveness of individual/small-group singing activities on kindergartners' use of singing voice and developmental music aptitude. *Journal of Research in Music Education, 44,* 353-368.

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*(1), 255-261.

TSANG, C. D., FRIENDLY, R. H., & TRAINOR, L. J. (2011). Singing development as a sensorimotor interaction problem. *Psychomusicology: Music, Mind and Brain*, *21*(1-2), 31-43. doi: 10.1037/h0094002

WELCH, G. F. (2006). Singing and vocal development. In G. McPherson (Ed.), *The child as musician* (pp.311-329). New York: Oxford University Press.

WELCH, G. F., HIMONIDES, E., PAPAGEORGI, I., SAUNDERS, J., RINTA, T., STEWART, C., ET AL. (2009). The national singing programme for primary schools in England: An initial baseline study. *Music Education Research, 11,* 1-22.

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.

WHIDDEN, C. (2010). Hearing the voice of non-singers: Culture, context & connection. In L. K. Thompson & ?M. R. Campbell (Eds.), *Issues of identity in music education: Narratives and practices* (pp. 83-107). Charlotte, NC: Information Age Publishing.

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(1),* 3–26. doi: 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.