Article

MUSICÆ SCIENTIÆ

Musicae Scientiae
1–12

DOI: 10.1177/1029864916638914
msx.sagepub.com
SAGE

The effect of repeated attempts and test-retest reliability in children's singing accuracy

Bryan E. Nichols
The University of Akron, USA

Sijia Wang
The University of Akron, USA

Abstract

The purpose of this study was to examine the effect of repeated attempts at singing accuracy tasks. Test-retest reliability of singing accuracy was examined in a second administration of the test. A secondary purpose was to analyze individual variability in children's singing accuracy. Test stimuli were designed using five attempts each at a single pitch, interval, and four-note pattern, and song singing. Children aged 6–11 were given the one-on-one singing accuracy test, and an identical form of the test was administered again within 1–6 weeks. Pitch matching items were scored by measuring the deviation in Hertz from the stimuli. The song singing item was scored by singing teachers using an 8-point scale with acceptable inter-rater reliability ($r = .86$). Participants' individual best attempt out of five was equally distributed, with overall performance increasing across subsequent attempts measured in signed cent deviation. A repeated-measures ANOVA with the task type (single, interval, and pattern) and attempt (1, 2, 3, 4, and 5) as the within-subjects variables indicated no main effect for task type ($p = .129$), but a significant main effect for attempt ($p < .001$, $\eta_p^2 = .087$). Test-retest reliability was considered high ($r = .69$).

Keywords

assessment, in-tune singing, singing accuracy, singing development, test-retest reliability

Studies in singing accuracy have described populations of children and adults in terms of proficiency and deficiency in efforts to ameliorate poor pitch singing or to understand the underlying neural mechanisms for this important motor skill. A variety of measures for pitch matching or song singing have been created including scales (e.g., Brophy, 1997; Salvador, 2010) and "test batteries" such as the Montreal Battery of Evaluation of Amusia (Dalla Bella & Berkowska, 2009) or the Sung Performance Battery (Berkowska & Dalla Bella, 2013). Such measures include the use of subjective or objective methods (Larrouy-Maestri, Leveque, Schon, Giovanni,

Corresponding author:
Bryan E. Nichols, The University of Akron, 254 Guzzetta Hall, Akron, OH 44325-1002, USA.
Email: bnichols@uakron.edu

& Morsomme, 2013). Two newly-developed tests have been put forth for large-scale use by automated computer testing.

The first automated, online test was developed by the research group Advancing Interdisciplinary Research in Singing (AIRS; see Cohen, 2015). The Test Battery of Singing Skills lasts up to 30 minutes and includes many different tasks, beginning with spoken text and data collection on singing range. Next is a series of singing tasks ranging from a simple sol-mi song to *Brother John*, followed by pitch matching tasks. A following section requires participants to improvise an ending for a given song opening, then to make up a song to given picture prompts. The song concludes with singing an unfamiliar (newly-learned) song, then *Brother John* again from memory, and a closing speaking task. Data has been reported on child and adult samples using the test (Cohen, 2015; Cohen, Pan, Stevenson, & McIver, 2015).

The most recent development is the Seattle Singing Accuracy Protocol (SSAP; Demorest et al., 2015). The SSAP is a computer-automated assessment of pitch matching and song singing, preceded by an automated protocol for establishing comfortable singing range. An important feature of this test is that the program adjusts to the comfortable pitch level of the user. Another important feature of this test is the automatic scoring of the participant including the ability to immediately share results with the user or to aggregate scored data for later analysis. The two aforementioned tests reflect increasing interest in singing accuracy from the music education and psychology research communities, and they indicate the perceived importance of singing as a basic function of human social activity (e.g., Good & Russo, 2015; Welch, 2015).

Accurate singing in these assessments is defined solely in terms of pitch, whether in deviation from given pitches in pitch matching tasks or performance on song singing tasks. Initially, music educators used simple tests administered under the time constraints present in schools or according to the assumed attention span of the children sampled (Salvador, 2010). Increasingly, researchers use multiple tasks in one test, and comparisons of performance on these tasks have now been made. Preliminary findings suggest that performance can vary based on the task(s) used, the specific test items, and the characteristics of the vocal or instrumental model. Notably, individuals' pitch matching ability is not always very strongly correlated to song singing ability (Demorest, Nichols, & Pfordresher, 2016; Hutchins, Larrouy-Maestri, & Peretz, 2014; Nichols, 2015; Pfordresher & Brown, 2007; Wise & Sloboda, 2008).

The assessment of singing accuracy relies on a basic assumption that participants have the ability to actually phonate in certain ranges of their singing voices, a construct called *singing voice development*. Many children do not demonstrate a full use of the extended voice range but instead sing only in the comfortable speaking part of the voice range (Rutkowski & Miller, 2003; Welch, Sergeant, & White, 1995). Thus, researchers must determine an individual's facility in vocal range prior to testing and screen out those who cannot use the full singing range, or they must use stimuli that are restricted in range to the comfortable speaking voice range. In addition to stimuli characteristics, the response mode of singing alone or doubled by other voices is an important variable that affects the profile of an individual singer (Nichols, 2015, 2016).

Moore, Brotons, Fyk, and Castillo (1997) noted that children performed differently on three trials at a newly-learned song. Forty-five percent of their participants sang most accurately on Trial One, 27% on Trial Two, and 28% on Trial Three. The study did not take into account the effect of singing voice development (range) on assessment. The effect of repeated attempts on pitch matching items – rather than song singing – is unknown, and this is an important consideration for the construction of singing accuracy measures. Recording the number of trials-to-criterion is another measure for accuracy – how many attempts does it take to sing a task accurately? This method has been used at least once in a study of uncertain singers (Porter, 1977). That study used a constricted range to moderate the effect of singing voice development.

Another important basic testing assumption is test-retest reliability. The evaluation of test-retest reliability yields an estimate of the stability based on two administrations of a test (Allen & Yen, 2001). A secondary feature of this type of reliability evaluation is that the nature of participants' variability can indicate whether learning effects carried over from the first test to the second test. In other words, test-retest reliability sometimes indicates that some participants did better on the second test because either they learned during the period of instruction between testing, or there were practice effects that carried over from the first test. Very high test-retest reliability demonstrates that neither occurred. It important to note that high test-reliability is not reduced by the presence of a ceiling effect or as a result of low precision in measurement; these issues should generally be addressed when reporting *r*-values.

Establishing test-retest reliability is an important next step for singing accuracy research, especially considering recent discussions in the field regarding replicability in general (e.g., Open Science Collaboration, 2015). For singing accuracy research, replicability is important due to the number of factors that can vary in terms of population, assessment tasks, and protocols. Further, there are unknown effects as to the effect of using familiar or newly-learned song material and in the latter case, how much exposure to the test administrator exists prior to testing (social effects). Next, children have sometimes been tested alone and other times in small groups, which can be assumed to affect performance. Finally, the effect of singing alone or doubled by other voices is also variable (Nichols, in press).

We chose to simultaneously test for the effect of repeated attempts in singing accuracy measurement because of the possibility of learning effects over test administrations, between test administrations, and during test administrations (musicians might call this "practice effects"). That is, to what degree do participants improve – if at all – when given several successive attempts at a particular test item? The rationale for analyzing the effect of repeated attempts is to validate whether students are tested at their best when given tests of singing accuracy where the examinee has only one opportunity to sing each test item. Thus there were two separate but related research questions: First, what is the effect of repeated attempts at common singing accuracy tasks in children? Second, what is the test-retest reliability of common singing accuracy tasks in children?

Method

Sample

Participants ($N = 94$) were recruited from grades 1, 2, 3, and 5 (aged 6–11 years, $M = 94.49$ months, $SD = 15.93$) in one public ($n = 66$) and one private elementary school ($n = 28$) in an urban area of the midwestern United States. The schools were considered typical representations of a local public and a local private school, and were considered based on their proximity to the university and the willingness of gatekeepers to allow the research to occur during the school day. Students in the two schools received weekly music classes from a music specialist. Participants who completed IRB-approved consent and assent protocols were tested during their regular music class time.

Test construction

The purpose of the study warranted an assessment that would be brief enough to be completed by children as young as five years old. Thus an effort was made to keep the test duration less than five minutes. Also, the assessment had to be used for two test administrations. A

singing accuracy assessment was created using four tasks: single pitch, interval, pattern, and song singing. The first three of these were pitch matching tasks, and stimuli were chosen for each task based on moderate difficulty (Nichols, in press; Sinor, 1984; Wolf, 2005). The stimuli ranged between D4 to A4 to prevent confounding singing accuracy with singing voice development (Rutkowski & Miller, 2003); students did not have to possess a fully developed singing range for this test.

Stimuli

The stimuli were recorded by an adult female instructed to employ minimal vibrato, per recommendations in previous research (Yarbrough, Green, Benson, & Bowers, 1991). Each pitch matching task was tested by presenting the stimulus by the pre-recorded voice for five attempts. For example, the single pitch task was tested using five presentations of G4, each followed by a pause to allow for the student response time. We chose to equalize the interval and pattern tasks by presenting both tasks using four pitches. The interval task was composed of two unique pitches that created the interval but two repetitions of each pitch (F♯ F♯ D D). The pattern task was composed of four unique pitches (D E F♯ G). This design removed the possibility of a memory effect for the number of pitches for the interval and pattern tasks, but not the single pitch task.

To address the second question of test-retest reliability, a song singing task was included at the end of both test occasions. The song singing task was tested using the song *Jingle Bells*, which was chosen based on its familiarity and moderate difficulty level (Demorest et al., 2016; Nichols, in press). Pitch matching tasks were conceptualized as "echo" tasks in which participants heard the pitch then sang it back (Nichols, 2015). The song task, however, was not presented in echo form as it was used in the pitch matching tasks. Instead, participants were instructed to sing Jingle Bells from memory then presented with a starting note (F♯ 4) using an electronic pitch pipe. These procedures were presented at the end of the recording playback, which allowed adequate time for the participant to perform the song at a self-selected tempo. There was no specific coaching or instruction on the pitch matching or song singing tasks used in this study, and no specific instruction related to the study occurred in the music classroom.

Procedure

The stimuli were played on a CD player during one-on-one testing, and each participant was tested in a small room near the music classroom at each school by the researcher or a research assistant. The test duration was approximately five minutes, and the same assessment was used again for the second administration of the test. The first test was administered during a one week period in the spring semester, and the second test was administered between one and six weeks later.

Scoring

Data was only included for participants who completed both test administrations. The pitch matching tasks were analyzed using Praat software (Boersma & Weenink, 2015). The data were measured in Hertz (Hz) using the middle stable portion of each vocal response (vocal "scooping" at the beginning or end of each response was not included in analysis). Next, deviation scores in Hz were created by calculating the absolute value of the difference between the given pitches and the vocal response by participants. The scores were transformed to cents for

analysis. Song singing scores were used to address the second question of test-retest reliability; those were scored on a scale of 1 to 8 using a previously-designed scale for singing accuracy (Wise & Sloboda, 2008). Two research assistants who were graduate students in a music education degree program scored a randomly-selected 20% of the participants, and the single measure intra-class correlation (ICC [2,1]) coefficient using the absolute agreement definition was calculated to equal .858 (mixed effects model, Shrout & Fleiss, 1979). Scores for these participants were averaged for inclusion in the analysis. Next, each assistant scored half the remaining participants.

Results

The gap in time between the first and the second test administration varied, with some participants taking the second test within the same week ($n = 36$) and the remaining participants taking the test five to six weeks later ($n = 58$). Therefore, a test for significant difference between the two groupings was necessary prior to examining the test-retest reliability of the total sample. There was no significant difference between the two groupings in overall pitch matching performance ($p > .05$) or song singing performance ($p > .05$), thus the following analyses were performed on the entire group as one sample. Further, there was no significant difference in performance by age ($p > .05$).

Question 1: Effect of repeated attempts

Participants made five successive attempts at each of the three pitch matching tasks. The analyses addressing Question 1 are based on the first test administration. Since the analyses were made using the absolute values of deviation scores, the distributions are first demonstrated here as signed values in cents (Figure 1).

A repeated-measures ANOVA with the task type (single, interval, pattern) and attempt (1, 2, 3, 4, 5) as the within-subjects variables indicated no main effect for task type ($p = .129$), but a significant main effect for attempt, $F(4, 372) = 6.21$, $p < .001$, $\eta_p^2 = .087$. There was a significant Bonferroni-adjusted comparison between Attempt 1 and all others ($p < .001$). A repeated-measures ANOVA indicated a significant difference in the five single pitch matching attempts, $F(2.91, 271) = 3.62$, $p = .014$, $\eta_p^2 = .037$, see Figure 2. There was a significant Bonferroni-adjusted comparison between Attempt 1 and 5, $p = .05$. A repeated-measures ANOVA indicated a significant difference in the five interval pitch matching attempts, $F(3.19, 296) = 10.01$, $p < .001$, $\eta_p^2 = .097$, see Figure 3. There was a significant Bonferroni-adjusted comparison between Attempt 1 and all others, $p < .003$ in each case. A repeated-measures ANOVA indicated no significant difference in the five pattern pitch matching attempts, $F(3.38, 314) = 9.03$, $p = .449$, $\eta_p^2 = .010$, see Figure 4. Each of the preceding analyses is reported using Greenhouse-Geisser corrections.

The individual performance of participants is plotted in a frequency distribution for each task (see Figure 5). As an example, individuals' best overall attempt averaging across tasks was on attempt 1 ($n = 17$), attempt 2 ($n = 20$), attempt 3 ($n = 13$), attempt 4 ($n = 22$), attempt 5 ($n = 22$). A chi-square test of goodness-of-fit was performed to determine whether the superior performance was demonstrated equally across the five attempts. Performance of the best attempt for single pitch matching was equally demonstrated in the population ($p > .05$).

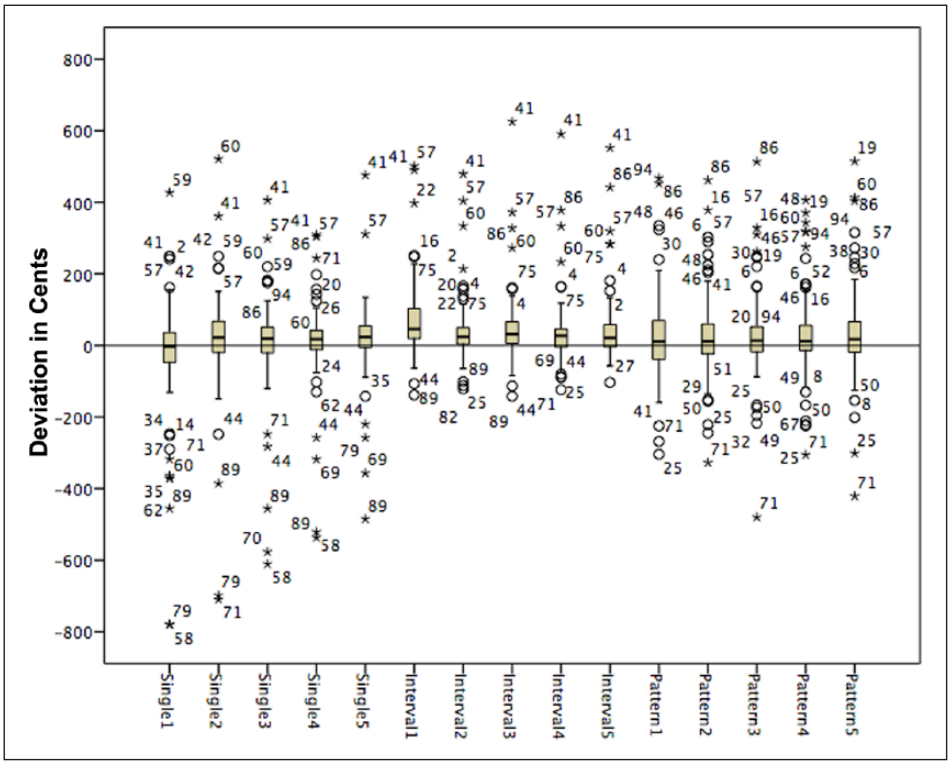


Figure 1. Deviation in cents showing distribution with outliers from first administration of the test.

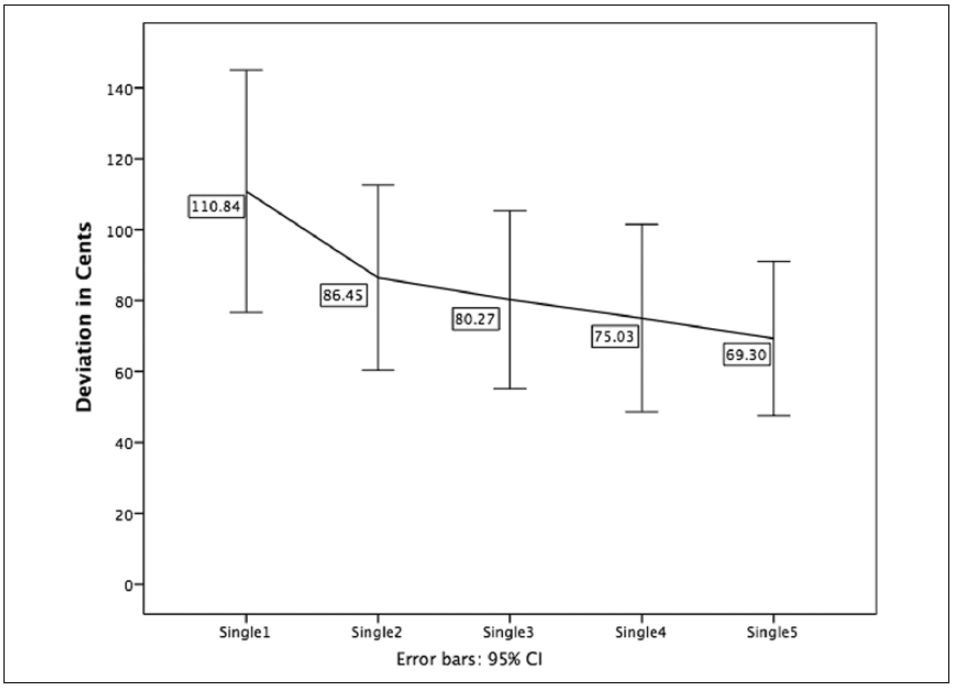


Figure 2. Signed deviation indicating mean for repeated attempts at a single pitch.

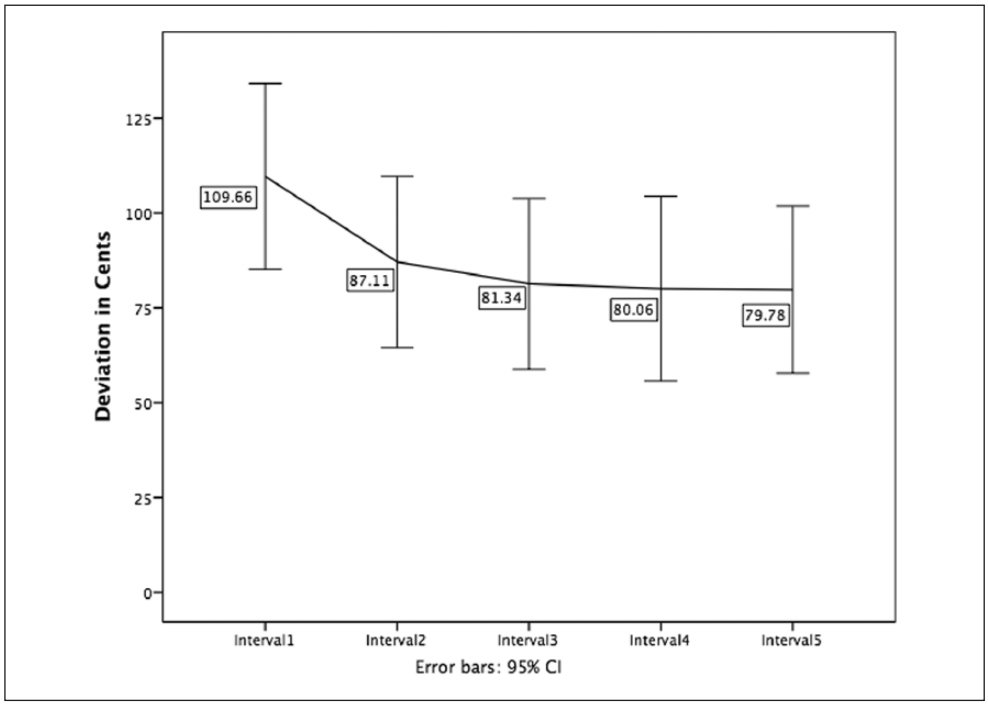


Figure 3. Signed deviation indicating mean for repeated attempts at an interval.

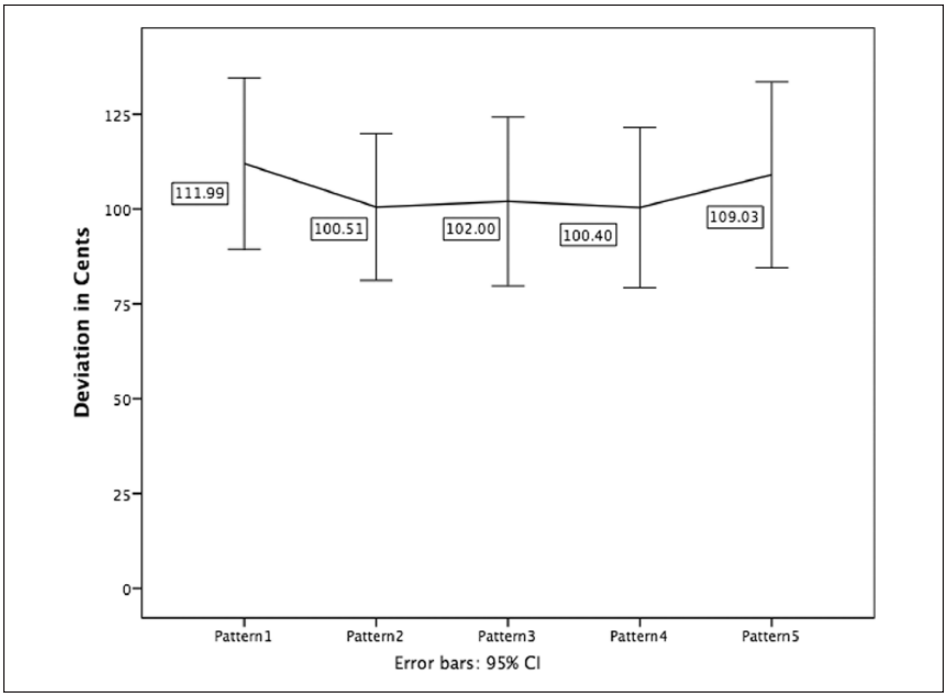


Figure 4. Signed deviation indicating mean for repeated attempts at a pattern.

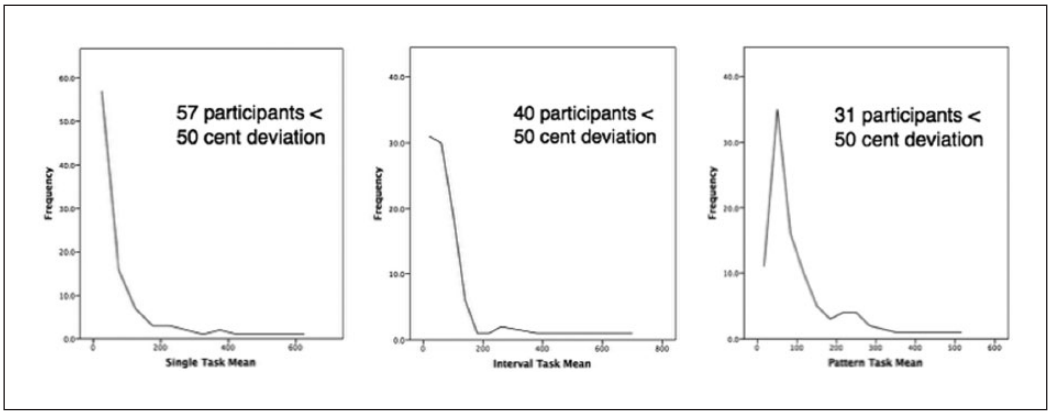


Figure 5. Frequency distribution of signed deviation in cents for mean performance on each pitch matching task (single pitch, interval, pattern).

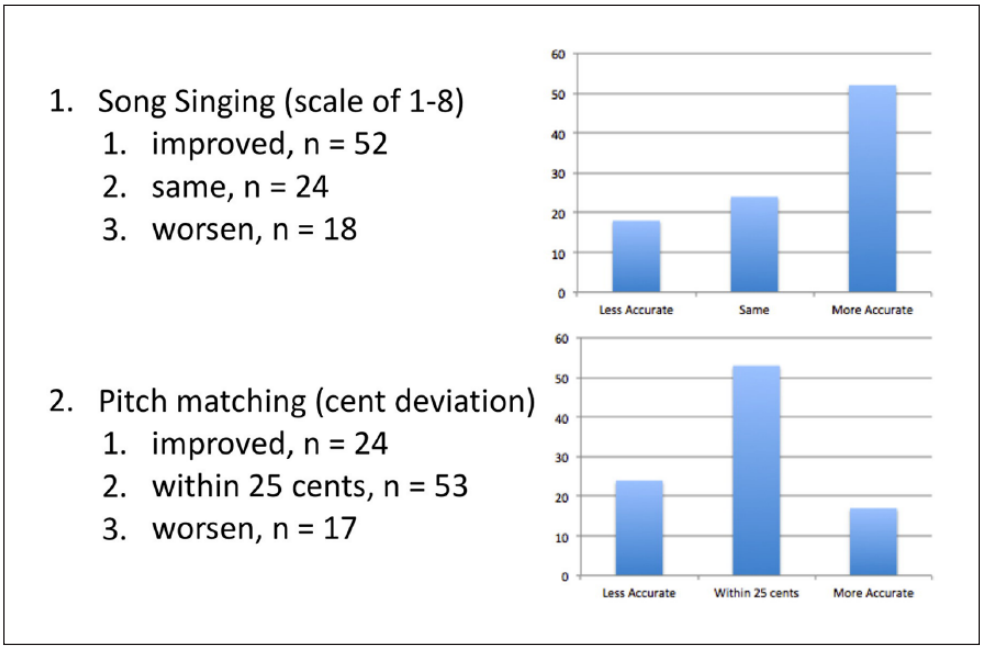
Question 2: Test-retest reliability

The assessment was administered a second time, and for these comparisons the pitch matching values were created by averaging the five attempts for each task. The song singing task was included in the first and second test administration, and those results will be examined here alongside the pitch matching results. Using the Fisher *r*-to-*z* transformation, there was no significant difference in correlations between the short- and long-retest groups on the single pitch ($p > .05$), interval ($p > .05$), or song singing tasks ($p > .05$), but there was a significant difference in correlations between the short- and long-retest groups on the pitch matching task ($p = .039$). The correlation of song singing tasks was considered to be high, $r = .758$. The pitch matching correlations were also interpreted to be high: single pitch ($r = .629$), interval ($r = .643$), and pattern ($r = .720$).

For pitch matching, both short- (mean change = +.66) *and* long-retest groups (mean change = +.35) improved significantly ($p = .014$ and $p = .019$, respectively). The higher mean change for the short-term retest group suggests a learning effect for the exam, and the positive mean change for the long-retest group suggests the possibility students may have improved accuracy overall during the year. Performance also improved for song singing, a mean increase of approximately .5 on the 8-point scale, $p < .001$. For song singing specifically, 52 participants demonstrated more accurate performance whereas others stayed the same ($n = 24$) or were less accurate ($n = 18$), see Figure 6. For pitch matching overall, 24 participants improved, whereas others sang within 25 cents ($n = 53$) or were less accurate ($n = 17$).

Discussion

The purpose of this study was to examine two basic assessment features of singing accuracy: the effect of repeated attempts at test items and the test-retest reliability of common assessment tasks. Singing accuracy was defined as the ability to sing in-tune and other singing characteristics such as tone or breath support were not evaluated. For singing accuracy, kindergarteners have been shown to not dramatically increase singing accuracy in song singing over a six month interval during the school year (Demorest et al., 2016). Over a period of one to six weeks, participants in the present study significantly improved their song singing score by



Figures 6a and 6b. Individual variability in song singing and pitch matching using mean task scores.

Table 1. Comparison of the best attempt out of five in present results of repeated attempts in pitch matching to a previous study using three repeated attempts at a newly-learned song.

	Vocal material	1st attempt best	2nd attempt best	3rd attempt best	4th attempt best	5th attempt best
Moore et al. (1997)	New song	45%	27%	38%	___	___
Current study	Pitch matching	19%	21%	14%	23%	23%

approximately half a point on the 8-point scale. However, it is unknown whether this finding may be due to testing effects or to actual skill improvement. For example, familiarity with the testing procedure, the test itself, or the test proctor is known to contribute to error in evaluation (Allen & Yen, 2001). It should be noted the test proctor was not always the same research assistant for each test administration for every participant, and the effect of this is beyond the scope of this study. Students in the short-retest group (within 1 week) demonstrated a greater average score increase than the long-retest group (up to six weeks), which further supports the interpretation that a score change was mediated by testing effects.

Moore et al. (1997) found about half of 6-to-9-year-olds to perform most accurately on the first trial of a newly-learned song (see Table 1). Participants in the present study, however, performed less variably in comparison: the distribution of individuals' superior attempt out of five was more equally distributed across attempts. However, participants performed more accurately on the fifth attempt at single pitch matching compared to the first attempt; that is, overall they sang somewhat closer to the given pitch on the fifth attempt (the only

statistically significant comparison to the first attempt). However, the fifth pitch was not more often sung most accurately than the others. This finding means it cannot be said that children sing best the first or second attempt (and so on) because individuals' best personal performance appeared to be equally distributed across the five attempts. However, participants sang closer overall to the given pitch as they progressed through attempts, which suggests that at least one practice item should be used before each task in a classroom assessment for singing accuracy.

Test-retest reliability was revealed to be very stable per the conventions in testing and measurement (Allen & Yen, 2001). In cases where the participant's test scores are perfectly linear, test-retest reliability is said to equal 1. But if the scores from the first test and the second test are not related, the reliability is said to equal 0. Caution must be used in interpreting the current results because many participants performed well on this test of a unison, interval, and pattern. Music educators can regard this finding as a positive sign of musical proficiency, but the possibility of a ceiling effect in this singing accuracy test may or may not have inflated the test-retest reliability of these items.

Teachers should make note that the retest in this study was an exact replication of the first test administration. Thus, these results are specific to measuring the assessment of performance on identical test items. The results of a retest of similar-but-different items using the same tasks are unknown, and these results should not be generalized to task-level effects. Test-retest reliability estimates are most appropriate for tests measuring traits that do not change quickly; singing accuracy skill may be one of these traits, based on previous research. In regard to task-level performance, participants performed less accurately across the three pitch matching tasks (single pitch, interval, and pattern). The current results differ from a previous study of 4th grade participants who demonstrated superior performance on five different items constituting an interval task compared to single pitch or pattern matching (Nichols, in press). The current results replicate similar findings in kindergarteners (Demorest et al., 2016).

Singing accuracy may be a normally distributed skill (Dalla Bella, 2015; Pfordresher & Larrouy-Maestri, 2015). Tests scores do not have to have a normal distribution, and many tests do not (Allen & Yen, 2001). Tests that have a bimodal or skewed distribution should be corrected for standardized testing, and for this assessment a ceiling effect did exist. The results of this study suggest that many students were capable of singing these tasks successfully. Additional test items – while not very feasible for the attention span of this age group – could provide a variety of test item difficulty to avoid a ceiling effect, and previous research has already established this for specific intervals.

Some participants sang much greater than 50 cents away from the given pitch. Some of these participants sang consistently in this way, and sometimes consistently sharp or flat. The participants in this study were not prescreened for ability, so the results may reflect the general distribution in the population. Some learning may have occurred during the test, and perhaps more so for low performers than high performers who could already successfully sing many test items. In other words, the repetition of items in the test could present a learning opportunity that affected performance on subsequent items or in the second administration of the test. This learning effect would be more or less likely depending on the use of pitch matching and song singing tasks in the music classroom. For example, in a music classroom where students are more often given instrument-playing tasks than singing tasks, students may respond differently to pitch matching tasks than students who are asked to sing frequently. Test-retest reliability in this study was shown to be high for pitch matching and song singing, and learning effects tend to lead to underestimated reliability in the population. Further, a change in participants' attitudes would also yield scores that underestimate reliability when compared. Some days

students may have been primed by activities that had just occurred in the music classroom (participants were tested during their regular music class time). Or, participants may have been more enthusiastic about participating in the first administration of the test than the second. Conversely, student performance may have been attenuated by greater anxiety levels during the first test. No measures of attention or anxiety were used for this test, but such measures could help explain variance in the student population in the future.

Few studies have emphasized individual performance in singing accuracy. In addition to reporting significance in overall performance, this report includes descriptives of proportions of change across successive attempts and from the first test to the retest using demarcations of 50 cents for distributions and 25 cents for score changes in individuals (some individuals improved, some stayed constant, and some decreased). While a 50 cent deviation represents the difference between the given pitch and an adjacent one, most people – musicians or non-musicians – would hear a 50 cent deviation as out of tune. Still, a threshold must be chosen, though less emphasis may be given to the threshold than to the importance of individual variability in singing accuracy.

Last, much of the data on singing accuracy relies heavily on the classical test theory (CTT) framework to examine the difficulty and discrimination ability of the specific test items chosen for the assessment. More complex modeling such as item-response theory may be useful for examining individual participants' performance independent of the difficulty of the test items, which cannot be achieved by the CTT. In conclusion, not all students perform best on the first attempt, though some do. The importance of using a practice item in singing accuracy assessment cannot be overstated, especially for summative reporting. The Seattle Singing Accuracy Protocol calls for one attempt per singing test item. The results here support the use of that approach if using a practice item. Test-retest reliability is high on tests where the same test items are used for the retest. The effect of using different items or repertoire for a retest is unknown and should not be considered reliable for documenting/reporting achievement. Pitch matching accuracy may improve slowly over time; thus, teachers may want to document the development of a child's ability but avoid formally reporting such development in terms of academic achievement in a school setting. Instead, teachers may wish to use more global assessments of pitch, rhythm, and tone for reporting purposes when required.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

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

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 6.0.07) [Computer software]. Retrieved from http://www.praat.org/

Brophy, T. S. (1997). Authentic assessment of vocal pitch accuracy in first through third grade children. *Contributions to Music Education*, *24*(1), 57–70.

Cohen, A. (2015). The AIRS test battery of singing skills: Rationale, item types, and lifetime scope. *Musicae Scientae*, *19*(3), 238–264.

Cohen, A., Pan, B., Stevenson, L., & McIver, A. (2015). Does non-native language influence learning a melody? A comparison of native English and Chinese university students on the AIRS test battery of singing skills. *Musicae Scientae*, *19*(3), 301–324.

Dalla Bella, S. (2015). Defining poor-pitch singing. *Music Perception*, *32*(3), 272–282.

Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority. *Annals of the New York Academy of Sciences, 1169*(1), 99–107.

Demorest, S. M., Nichols, B. E., & Pfordresher, P. Q. (2016). *The effect of focused instruction on kindergartener's singing accuracy*. Manuscript in preparation.

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

Good, A., & Russo, F. A. (2015, August). *Singing and multi-cultural group cohesion*. Paper presented at the Biennial Meeting of the Society for Music Perception and Cognition, Vanderbilt University, Nashville, TN.

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception & Psychophysics, 76*(8), 2522–2530.

Larrouy-Maestri, P., Leveque, Y., Schon, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice, 27*(2), 251–259.

Moore, R., Brotons, M., Fyk, J., & Castillo, A. (1997). Effects of culture, age, gender, and repeated trials on rote learning skills of children 6-9 years old from England, Panama, Poland, Spain, and the United States. *Bulletin for the Council of Research in Music Education, 133*, 83–88.

Nichols, B. E. (2015). Critical variables in singing accuracy test construction: A review of literature. *Update: Applications of Research in Music Education, 34*. Advance online publication. doi:10.1177/8755123315576764

Nichols, B. E. (2016). *Doubling as an important variable in singing accuracy research*. Manuscript submitted for publication.

Nichols, B. E. (in press). Task-based variability in children's singing accuracy. *Journal of Research in Music Education*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 6251.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception, 25*(2), 95–115.

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*, 271. doi:10.3389/fnhum.2015.00271

Porter, S. Y. (1977). The effect of multiple discrimination training on pitch matching behaviors of uncertain singers. *Journal of Research in Music Education, 25*(1), 68–82.

Rutkowski, J., & Miller, M. S. (2003). The effect of teacher feedback and modeling on first graders' use of singing voice and developmental music aptitude. *Bulletin of the Council for Research in Music Education, 156*, 1–10.

Salvador, K. (2010). How can elementary teachers measure singing voice achievement? A critical review of assessments, 1994–2009. *Update: Applications of Research in Music Education, 29*(1), 40–47.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Sinor, E. (1984). *The singing of selected tonal patterns by preschool children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (AAT 8501456)

Welch, G. (2015). Singer identities and educational environments. In R. MacDonald, D. Miell, & D. Hargreaves (Eds.), *Oxford handbook of musical identities*. New York, NY: Oxford University Press.

Welch, G., Sergeant, D., & White, P. J. (1995). The singing competencies of five-year-old developing singers. *Bulletin of the Council for Research in Music Education, 127*, 155–162.

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.

Wolf, D. (2005). A hierarchy of tonal performance patterns for children ages five to eight years in kindergarten and primary grades. *Bulletin of the Council for Research in Music Education, 163*, 61–68.

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.