sempre:
Society for Education, Music and Psychology Research

Short Research Report

Psychology of Music
1–10

DOI: 10.1177/0305735618799171
journals.sagepub.com/home/pom
SAGE

Does doubled singing increase children's accuracy? A re-examination of previous findings

Bryan E. Nichols[1] and Julie Lorah[2]

Abstract

Studies comparing solo singing to doubled singing indicate contrasting findings as to whether children evince superior solo or doubled singing. Previous findings have indicated: (a) superior solo singing; (b) superior doubled singing; or (c) no significant difference. A systematic review of studies meeting the inclusion criteria ($N = 6$) was undertaken to examine factors leading to these conflicting results. Next, a calculation of effect sizes that were unreported in previous research was based on published ANOVA tables, and expressed using Partial Eta-Squared. In direct comparisons of solo to doubled singing conditions, two studies reported that children sing more accurately in doubled singing; two studies reported more accurate solo singing; and two studies reported no significant difference by performance in the two conditions. The results indicate medium-to-large effect sizes in both directions. Several factors were enumerated to explain the contrasting findings: test administration procedures, song familiarity, vocal models, scoring methods, and teacher/researcher familiarity among the participants.

Keywords

individual differences, pitch, sensorimotor skills, singing, skill, voice

Elementary music teachers routinely ask students to sing solo as well as doubled with other children in the classroom setting. Importantly, instruction is shown to enhance singing accuracy, the degree to which a child can sing in-tune as indicated in a recent meta-analysis of instruction effects in singing accuracy (Svec, 2017). Children have most often been tested on solo singing (e.g., Demorest, Nichols & Pfordresher, 2017; Rutkowski, 1990; Salvador, 2010), defined as singing in absence of other voices/instruments playing the soloist's part (Nichols, 2016). However *doubled* singing, singing along with one or more voices or instruments on the same part (e.g., Nichols, 2016), may be a more ecologically valid assessment consistent with classroom experience than solo singing (Allen & Yen, 2001). Six previous studies comparing solo and doubled singing have reported mixed results including: (a) students perform more

[1]School of Music, The Pennsylvania State University, University Park, PA, USA
[2]School of Education, Indiana University, Bloomington, USA

Corresponding author:
Bryan E. Nichols, PhD, School of Music, The Pennsylvania State University, 233 Music 1, University Park, PA 16802, USA.
Email: bnichols@psu.edu

accurately singing solo (Goetze & Horii, 1989; Smale, 1987); (b) students perform more accurately singing doubled (Green, 1994; Nichols, 2016); (c) there is no significant difference in accuracy between solo and doubled singing (Cooper, 1995; Smith, 1973). In this paper, we will examine the possible causes for this discrepancy and provide suggestions for future research to clarify the relationship between solo and doubled singing conditions.

Doubled singing research

Authors of previous studies have defined doubled singing in different ways. Some refer to *unison* (Green, 1994) or *group* singing (Goetze & Horii, 1989), when singers are singing the same pitches together. Others use the term *doubling*, which is conceptualized from the perspective of one singer singing along with an external stimulus using the same pitches (Nichols, 2016), and the term *accompaniment* to describe doubled singing with either voices or instruments (Wise & Sloboda, 2008). Thus, doubling is a variable that can be named and operationalized in various ways. In this paper we chose the term doubling to refer to any of these situations.

Researchers have tried to address the wide variety of singing accuracy measurement factors (Nichols, 2015) by introducing more standardized approaches to the assessment of singing accuracy such as the Advanced Interdisciplinary Research in Singing (AIRS) Test Battery (Cohen, 2015) and the Seattle Singing Accuracy Protocol (SSAP; Demorest & Pfordresher, 2015; Demorest et al., 2015). Both are computer-based tests, and the SSAP includes automatic, real-time scoring methods that allow comparison across multiple studies. A doubling variable may become an important component of these emerging test protocols.

The purpose of this paper was to synthesize the literature related to singing accuracy in either solo or doubled singing conditions and to explore sources of discrepancies in six studies that compared solo and doubled singing. The research questions were, "What is the effect of doubling on singing accuracy, and what is the corresponding effect size?" and "How does a comparison of these effect sizes inform recommendations for future research and also educational practice?" First, we will review the literature regarding variables that may affect singing accuracy assessment. Next, we will calculate effect sizes and post hoc power from previous doubling studies in which no effect sizes were originally reported. The result will be a systematic comparison exploring these discrepant findings.

Systematic review

We deemed a systematic review as the appropriate method for this inquiry because the small number of studies available ($N = 6$) was not sufficient for a more rigorous statistical procedure such as meta-analysis (Cooper, 1982; Petrosino, Boruch, Soydan, Duggan & Sanchez-Meca, 2001; Rosenthal & DiMatteo, 2001). Regarding inclusion criteria, we chose to include all published research without limiting the time frame and included all studies directly comparing solo singing and doubled singing conditions in school-aged populations. A search yielded 153 papers related to pitch production. After an examination of titles and abstracts, nine publications were identified on the topic of doubled singing ranging from 1973 to 2016 (older studies did not use the term singing accuracy). Of these, three were excluded: one author reported that inaccurate students sang better alone after separating them from a group, however, there was not a test comparing solo to doubled singing (Joyner, 1969); two research articles were a publication of the original research and identical data from a dissertation study, so the dissertation studies were excluded (Goetze & Horii, 1989; Nichols, 2016). The remaining six studies included published dissertations ($n = 2$) and research articles ($n = 4$).

Table 1. Summary of article characteristics.

		Task	Stimuli	Text	Grade level	Doubling type	Scoring
Superior doubled singing	Green, 1994	song	sung from memory	text	1,2,3,5	8 peers	judges[a]
	Nichols, 2016	pitch matching (PM) plus song	recorded adult female	"doo" (PM) text (song)	4	recorded adult female	judges[b]
Superior solo singing	Cooper, 1995	pattern	recorded child	"loo"	1-5	recorded child	acoustic[c]
	Smith, 1973	song	recorded children	text	6	groups of 8	judges[d]
No significant difference	Goetze & Horii, 1989	short phrase solo song task (doubled)	researcher's live voice	both "loo" and text	K,1,3	6 peers plus researcher	acoustic
	Smale, 1987	first phrase from a song task	researcher's live voice	both "loo" and text	pre-school[e]	5 peers plus researcher	acoustic

[a]Dichotomous scoring of pitches (50% of total score) and intervals (50%). [b] Dichotomous scoring for PM, and an 8-point scale for song singing. [c] Refers to measurement of cent deviation using computer-assisted techniques. [d] 7-point scale. [e] In the USA, the first year of schooling is called Kindergarten followed by grade levels 1, 2, 3, etc. Children are 5-6 years old at the start of schooling, thus pre-school children are 4-5 years old.

Variations in test design. The six studies differed in test design, and some of these differences may have systematically impacted the results in terms of doubling (see Table 1). The mixed findings may have been due in part to the varying types of singing tasks used in these studies, such as a four-beat pattern on "loo" (Cooper, 1995), a short phrase from a song (Goetze & Horii, 1989; Smale, 1987), a newly taught children's song (Green, 1994), or a familiar song (Smith, 1973). In another study, a combination of pitch matching and song singing tasks were used (Nichols, 2016). The following sections will explore additional possible factors.

Age and pitch range. Doubling studies vary widely in the age and pitch ranges used (pitch range may be intentionally matched to participant age, or it may not). Participants in doubling studies ranged from ages 4 to 12, including ages 4 to 5 (Smale, 1987); grades Kindergarten, 1 and 3 (Goetze & Horii, 1989); grades 1, 2, 3, 5 (Green, 1994); from grades 1 to 5 (Cooper, 1995); grade 4 (Nichols, 2016), and grade 6 (Smith, 1973). The pitch ranges used were appropriate for the age of participants with one exception: Smale (1987) used a range of C4 (middle C) to G4. Other researchers including Smale (1987, p. 16) have questioned the use of pitches outside most children's comfortable range, below D4 for young children (Goetze, 1985; Wolf, 2005). Singing accuracy has been shown to increase across the elementary years (Goetze, Cooper & Brown, 1990), but no clear pattern emerges in the present set of studies.

Stimuli. In assessments, teachers and researchers can model using their own voice, a student's voice, or possibly a common instrument such as the piano. From the general singing assessment literature, children are shown to sing more accurately when the stimuli are presented in their register (Kramer, 1986; Sims, Moore, & Kuhn, 1982); a female rather than a male model is used (Yarbrough, Green, Benson, & Bowers, 1991); a child's rather than an adult's voice is

used (Green, 1990); less vibrato is used (Yarbrough, Bowers, & Benson, 1992), and when a male falsetto rather than a chest voice is used (Price, Yarbrough, Jones, & Moore, 1994; Yarbrough, Morrison, Karrick, & Dunn, 1995). Doubling studies have used a pre-recorded adult female model (Nichols, 2016), the researcher's own voice (Goetze & Horii, 1989; Smale, 1987), a recorded child's voice (Cooper, 1995), or a group of pre-recorded children (Smith, 1973). Yet another study required children to sing only one song after rehearsal with the researcher, thus no stimuli were required (Green, 1994).

Practice prior to testing. The procedures used for introducing or teaching musical material vary among studies. Unrehearsed (spontaneous) song singing from memory has been used as an assessment task (Smith, 1973), and both pitch matching items and unrehearsed song singing have been used together in a test (Nichols, 2016). In other studies, a previously unfamiliar song was taught and rehearsed by the researcher (Green, 1994; Goetze & Horii, 1989; Smale, 1987); in another study, a pattern was rehearsed by the researcher until participants were comfortable singing the pattern (Cooper, 1995). Smith (1973) used a complex design in which each participant was asked to sing in three experimental conditions in a low range and again in a high range: (a) peer group absent-unaccompanied; (b) peer group absent-accompanied; and (c) peer group-present-accompanied. The presence or absence of practice prior to testing in previous studies introduces a confounding variable because singing accuracy could be a function of practice rather than the doubling variables.

Text. Previous reports on the use of text in singing accuracy have yielded inconsistent results. In a direct comparison, Goetze (1985) found superior performance on the neutral syllable "loo" vs. song text. Smale (1987) found no significant difference between the use of a neutral syllable or text. Another study reported inconclusive findings in a comparison of text vs. a neutral syllable (Gault, 2002). For the present systematic review, some studies used a text for song tasks including *Bow Wow Wow* (Green, 1994), *Jingle Bells* (Nichols, 2016), *America* (Smith, 1973), *Pinto Pony* (Goetze & Horii, 1989), and *Funny Clown* (Smale, 1987). Only one study used song phrases on "loo" (Cooper, 1995).

Scoring. The studies in this systematic review used: (a) calculations of cent deviation (Cooper, 1995; Goetze & Horii, 1989, Smale, 1987); (b) a 50-cent judging threshold (Green, 1994; Nichols, 2016); or (c) a rating scale (Smith, 1973). The scoring of singing accuracy can be accomplished using human judgement—usually dichotomously using a 50-cent demarcation for "in-tune" (Nichols, 2016), or it can be measured in Hertz. When measured using Hertz, the absolute value of a deviation score in cents is usually tabulated for comparisons (Larrouy-Maestri, Leveque, Schon, Giovanni, & Morsomme, 2013). Regardless of the inclusion of a doubling variable, results in singing accuracy studies are highly sensitive to the scoring and analysis method chosen (Pfordresher & Larrouy-Maestri, 2015). Although the 50-cent demarcation for accuracy has been frequently used in pitch analysis, many listeners would perceive a 50-cent deviation from an intended pitch to be quite out of tune in a song context; for this reason, some reports have even used a 25-cent deviation to separate singer groups (e.g., Nichols & Wang, 2016).

Method, analysis and results: Effect size calculations

Effect sizes were not regularly reported in music education research until recently (see Morrison, 2015 for a brief discussion), thus we chose to calculate effect sizes from five of the studies that

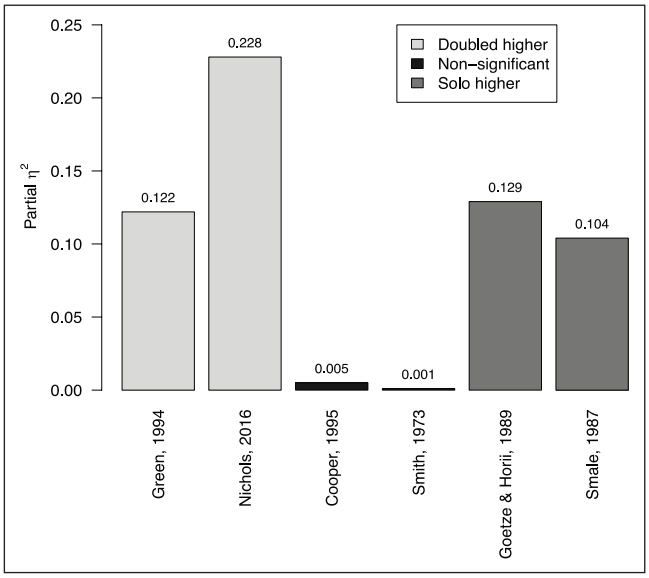

Figure 1. Effect size for doubling variable in previous studies. These effect sizes are interpreted using 0.01 (small), 0.06 (medium), and 0.14 (large).

met the inclusion criteria (one study originally reported effect sizes). Data from published ANOVA tables were used, and partial eta-squared (partial η^2) was computed for the doubling variable in the given study (see Figure 1). Partial η^2, which can be computed: SS_{effect} / (SS_{effect} + SS_{error}), is a measure of effect size representing variance attributable to the given effect out of variance attributable to that effect plus error variance (Tabachnick & Fidell, 2007). This measure is a proportion of variance and therefore can range from 0 to 1, and it is comparable across studies. This measure was chosen rather than η^2 since its value will not depend on the number and significance of the other factors in the model (Tabachnick & Fidell, 2007) and represents a recommended effect size measure for factorial ANOVA designs (Lomax, 2007; Tabachnick & Fidell, 2007). Lomax (2007) has provided guidelines for interpreting Partial η^2: 0.01 (small effect); 0.06 (medium effect); and 0.14 (large effect).

The results from the effect size calculations in the current analysis ranged from 0.0001 to 0.228. Two studies reported no statistical significance as to the effect of the doubling condition (partial η^2 0.001 and 0.005); two studies reported superior performance in the solo singing condition yielded a small effect; finally, two studies reported superior performance in the doubled singing condition yet also yielded a small effect (Lomax, 2007).

To better understand the differences in statistical significance outcomes among the studies, a post hoc power analysis was conducted for the doubling variable in each study. To conduct power analysis, the G*Power software was used (Faul, Erdfelder, Lang, & Buchner, 2007). All analyses were conducted assuming Type I error rate $\alpha = 0.05$ (see Table 2). Power can be interpreted as the probability of rejecting a false null hypothesis (King, Rosopa, & Minium, 2011) implying that the post hoc power presently computed for each of the studies under consideration represents the probability of detecting a significant doubling effect in each of these studies, assuming it exists.

As can be seen from the results, the studies that did not find a significant difference may have been underpowered, whereas the studies that did find a significant difference had high power.

Table 2. Post hoc power analysis.

Study	Superior Singing	N	Partial η^2	Post hoc power
Green (1994)	Doubled	241	0.122	0.99
Nichols (2016)	Doubled	120	0.245	0.99
Cooper (1995)	NS	169	0.005	0.14
Smith (1973)	NS	236	0.001	0.07
Goetze & Horii (1989)	Solo	165	0.129	0.99
Smale (1987)	Solo	93	0.104	0.99

However, it is important to note that post hoc power is a function of effect size. Therefore, it is particularly interesting to examine these results in the context of the widely differing effect sizes. Although sample size plays a role in power, the sample sizes among studies were deemed adequate (from 93 to 241 participants). However, the effect size (partial η^2) of the doubling factor varied from 0.001 to 0.245. In addition, this effect varied in direction among the different studies (i.e., some studies indicated doubled singing performed better while other studies indicated solo singing performed better). It is not clear why a factor purporting to measure the same construct would vary so much in terms of the magnitude and direction of that effect.

Finally, it is important to note that the interpretation of main effects changes in the presence of significant interactions from average effects to conditional effects (Aiken & West, 1991). Results from Green (1994) indicated a significant interaction between the doubling condition and grade, and results from Goetze and Horii (1989) indicated a significant interaction between the doubling condition and gender. Specifically, students of all grades scored higher in the doubling condition, but the difference was particularly pronounced for students in grade 5, compared with students in grades 1, 2, and 3 (Green, 1994). Additionally, for students in a study concluding solo singing is more accurate, boys particularly benefited from solo versus doubled singing compared with girls (Goetze & Horii, 1989). In both cases, the direction of the doubling effect was not modified due to the interaction, indicating conclusions regarding the main effect (the doubling condition) still seem warranted. However, it is important to interpret variables with significant interaction effects in the context of those interactions. Further, the presence of significant interaction effects in multiple studies indicates that in general, the doubling condition may impact singing accuracy differentially, depending on context.

Discussion

Studies that met the criteria for inclusion varied on many factors including: (a) doubling condition; (b) task type; (c) stimulus model; (d) procedures; (e) text; (f) participant age; (g) singing range; (h) designs for order effects; and (i) scoring/analysis. The following hypotheses are offered as possible explanations for the discrepancy in findings regarding the effect of doubling on singing accuracy (partial $\eta^2 = 0.0001–0.245$).

Singing task

Notably, each study used different tasks for the evaluation of singing accuracy (e.g., neutral syllable, pitch patterns, short phrase, songs). In Nichols (2016), some participants sang the song task more accurately in the doubled condition ($n = 60$), some performed less accurately (n

= 15), and some performed the same (n = 44). Thus, individual variability may be expected in singing accuracy results, and reporting only overall means may be problematic for describing skill development in individuals.

Specific testing procedures

The effect of the doubling variable on singing accuracy may be related to discrepancies in test administration procedures, including rehearsal before testing, the varying difficulty of pitch matching or song singing tasks, and the impact of the familiarity of a given song. Procedures for test administration varied: some researchers asked children to sing a familiar song without any priming whereas other researchers taught children an unfamiliar song until they were comfortable singing it. The effect of the researcher-as-teacher, the effect of repetition and rehearsal prior to testing, and the effect of familiar versus unfamiliar songs are unknown, and therefore, we suggest researchers explicitly model and analyze these variables in future research as well as consider these factors explicitly in standardized assessment design.

Individual differences

It is likely that individuals differ as to whether they perform more accurately singing solo versus doubled; results from studies examining this could be impacted by the proportion of singers from each group represented in the given study. For example, in one study reporting that participants on average sang significantly more accurately in the doubled condition, some participants actually sang more accurately in the solo condition (Nichols, 2016). It is unclear what underlying factors could explain this individual variability, but there is evidence that poorer singers may be overwhelmed by the presence of an external stimulus and actually perform more accurately when singing alone (Cooper, 1995; Smith, 1973). Possibly, studies reporting no significant difference in the doubling conditions were presented with equal proportions of singer types (those who performed more accurately solo or more accurately doubled). Further, whether an individual sings more accurately solo or doubled is related to other study variables, such as stimuli characteristics of item and task difficulty. Thus, future research should go beyond reporting summary statistics, such as overall mean scores. For example, future analyses could identify whether each student performs better in the solo or doubled condition and predictors of this binary outcome could be used to further understand these relationships.

Measurement of singing accuracy

Differing procedures for measuring singing accuracy may relate to the discrepancy in findings. Measurement procedures for scoring singing accuracy varied widely and included the use of scales and cent deviation cut-offs as well as expert judgement versus computer scoring. Measurement of pitch deviation scores is becoming more refined in the psychology literature, and the scoring method may influence the categorization of an individual as an accurate or inaccurate singer. These scoring procedures have been applied to different types of singing tasks, which further complicates the effect of these different procedures. Previous results may be influenced by scoring method as well as item or task difficulty, especially in studies that use only one task for comparison. The literature suggests the use of multiple task types can help researchers make future comparisons to previous studies (e.g., Nichols, 2016).

Conclusions and suggestions for future research

The results presented here suggest a potential limitation in the existing literature: reports of singing accuracy apply only in specific, highly-contextual designs and do not necessarily generalize broadly. Researchers of singing effects may consider using an established protocol to make direct comparisons to previous research. Such a protocol may include multiple tasks such as pitch matching and song singing on text and neutral syllables. A doubled singing condition can be incorporated in these existing protocols. Researchers should try to include many of these design factors in future studies, and for those they are unable to control for or model in some way, clear limits to generalization should be specified.

Researchers may also want to consider computer-administered assessment for future studies. One advantage of computer-administered assessment can be automatic scoring, and one existing test has been optimized for this (SSAP, Demorest et al., 2015). Further, studies on the effect of doubled singing may require the use of varied doubling strategies such as singing with one peer, singing in groups of peers, and singing with the teacher or researcher, and perhaps also an effort to directly compare these conditions. Based on the large effect sizes both in favor of superior solo singing and in favor of superior doubled singing, researchers are cautioned to design conditions very carefully: how and with whom the participant sings may affect the results considerably. Further, established protocols can lead to greater validity and reliability in teachers' assessment of singing accuracy.

The results here imply children do not sing more accurately in all solo or doubled conditions; rather, certain conditions may elicit the most accurate performances. We suggest the use of only one context in a study requires a qualification that those results apply only to that specific environment and test conditions. It remains possible that there are one or more additional variables that moderate the relationship between doubled singing and accuracy for children. Several possible variables have been explored including scoring; effects of repetition and rehearsal; stimuli design and model type; and the use of familiar vs. unfamiliar songs. These variables should be accounted for and explicitly measured in future research to provide a more nuanced understanding of the relationship of doubled singing to singing accuracy assessment.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

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

Cohen, A. (2015). The AIRS test battery of singing skills: Rationale, item types, and lifetime scope. *Musicae Scientae*, *19*, 238–264. doi:10.1177/1029864915599599

Cooper, H. M. (1982), Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*, 291–302. doi:10.3102/00346543052002291

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. doi:10.2307/3345637

Demorest, S. M., & Pfordresher, P. Q. (2015). Singing accuracy development from K-Adult: A comparative study. *Music Perception*, *32*, 292–302. doi:10.1525/MP.2015.32.3.293

Demorest, S. M., Nichols, B. E., & Pfordresher, P. (2017). The effect of focused instruction on young children's singing accuracy. *Psychology of Music*, *46*(4), 488–499. doi:10.1177/0305735617713120

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*, 266–271. doi:10.1525/mp.2015.32.3.266

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Gault, B. (2002). Effects of pedagogical approach, presence/absence of text, and developmental music aptitude on the song performance accuracy of kindergarten and first-grade students. *Bulletin of the Council for Research in Music Education*, *152*, 54–63.

Goetze, M. (1985). *Factors affecting accuracy in children's singing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8528488)

Goetze, M., Cooper, N., & Brown, C. (1990). Recent research on singing in the general music classroom. *Bulletin of the Council for Research in Music Education, 104*, 16–37.

Goetze, M., & Horii, Y. (1989). A comparison of the pitch accuracy of group and individual singing in young children. *Bulletin of the Council for Research in Music Education*, *99*, 57–73.

Green, G. (1990). The effect of vocal modeling on pitch-matching accuracy of elementary schoolchildren. *Journal of Research in Music Education, 38*, 225-231.

Green, G. A. (1994). Unison versus individual singing and elementary students' vocal pitch accuracy. *Journal of Research in Music Education*, *42*, 105–114. doi:10.2307/3345186

Joyner, D. R. (1969). The monotone problem. *Journal of Research in Music Education, 17*, 115–124. doi:10.2307/3344198

King, B. M., Rosopa, P. J., & Minium, E. W. (2011). *Statistical reasoning in the behavioral sciences*. Hoboken, NJ: Wiley.

Kramer, S. J. (1986). *The effects of two different music programs on third and fourth grade children's ability to match pitches vocally* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8524224)

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*, 251–259. doi:10.1016/j.jvoice.2012.11.003

Lomax, R. G. (2007). *Statistical concepts: A second course*. London: Routledge Press.

Morrison, S. J. (2015). Forum. *Journal of Research in Music Education*, *63*, 143–144. doi:10.1177/0022429415595625

Nichols, B. E. (2015). Critical variables in singing accuracy test construction: A review of literature. *Update: Applications of Research in Music Education*, *34*, 39–46. doi:10.1177/8755123315576764

Nichols, B. E. (2016). Task-based variability in children's singing accuracy. *Journal of Research in Music Education*, *64*, 309–321. doi:10.1177/0022429416666054

Nichols, B. E., & Wang, S. (2016). The effect of repeated attempts and test-retest reliability in children's singing accuracy. *Musicae Scientiae*, *20*, 551–562. doi:10.1177/1029864916638914

Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The Campbell Collaboration. *The Annals of the American Academy of Political and Social Science*, *578*, 14–34. doi:10.1177/000271620157800102

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., & 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

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. doi:10.2307/3345736

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. doi:10.1146/annurev.psych.52.1.59

Rutkowski, J. (1990). The measurement and evaluation of children's singing voice development. *The Quarterly Journal of Music Teaching and Learning*, *1*(1–2), 81–95.

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. doi:10.1177/8755123310378454

Sims, W. L., Moore, R. S., & Kuhn, T. L. (1982). Effects of female and male vocal stimuli, tonal pattern length and age on vocal pitch-matching abilities of young children from England and the United States. *Psychology of Music*, Special Issue: Proceedings of the IX International Seminar on Research in Music Education, 104–108.

Smale, M. J. (1987). *An investigation of pitch accuracy of four and five-year-old singers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8723851)

Smith, R. (1973). *Factors related to children's in-tune singing abilities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 7411404)

Svec, C. L. (2017). The effects of singing instruction on the singing ability of children aged 5–11: A meta-analysis. *Psychology of Music*, *46*(3), 326–339. doi:1177/0305735617709920

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson.

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. doi:10.1177/102986490801200102

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., Bowers, J., & Benson, W. (1992). The effect of vibrato on the pitch matching accuracy of certain and uncertain singers. *Journal of Research in Music Education*, *40*, 30–38. doi:10.2307/3345772

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.

Yarbrough, C., Morrison, S. J., Karrick, B., & Dunn, D. E. (1995). The effect of male falsetto on the pitch-matching accuracy of uncertain boy singers, grades K-8. *Update: Applications of Research in Music Education*, *14*(1), 4–10.