THE EFFECTS OF INSTRUCTION ON THE SINGING ABILITY OF
CHILDREN AGES 5 – 11: A META-ANALYSIS

Christina L. Svec

Dissertation Proposal Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2015

APPROVED:

Debbie Rohwer, Major Professor and Chair of
the Department of Music Education
Don Taylor, Committee Member
Robin Henson, Committee Member
Benjamin Brand, Committee Member
James C. Scott, Dean of the College of Music
Costas Tsatsoulis, Interim Dean of the Toulouse
Graduate School

Svec, Christina L. *The Effects of Instruction on the Singing Ability of Children Ages 5-11: A Meta-Analysis*. Doctor of Philosophy (Music Education), August 15 2015, 192 pp., 14 tables, reference list, 280 titles.

The purpose of the meta-analysis was to address the varied and somewhat stratified study results within the area of singing ability and instruction by statistically summarizing the data of related studies. An analysis yielded a small overall mean effect size for instruction across 34 studies, 433 unique effects, and 5,497 participants ranging in age from 5- to 11-years old ($g = 0.43$). The largest overall study effect size across categorical variables included the effects of same and different discrimination techniques on mean score gains. The largest overall effect size across categorical moderator variables included research design: Pretest-posttest 1 group design. Overall mean effects by primary moderator variable ranged from trivial to moderate. Feedback yielded the largest effect regarding teaching condition, 8-year-old children yielded the largest effect regarding age, girls yielded the largest effect regarding gender, the Boardman assessment measure yielded the largest effect regarding measurement instrument, and song accuracy yielded the largest effect regarding measured task. Conclusions address implications for teaching, research pedagogy, and research practice within the field of music education.

Copyright 2015

by

Christina L. Svec

ACKNOWLEDGEMENTS

I am grateful for the friends and family members who directly and indirectly provided support that led me through my doctoral journey. My mother and father taught me to be independent and ambitious; I am proud to be their reflection. My sisters, Jonathan Evans, Allie Evans, Dheeraj Chand, Kara Saldana, Chuck Pineda, and First Presbyterian Church of McKinney friends were always quick with a word of encouragement or evening of friendship. My former students filled my mind with memories that kept me passionate about music education. My fellow doctoral students pushed me to have clear support for my thoughts, especially Tom Burlin, Jordan Edmonson, and Ben Price. Jennifer Brimhall was and will always be my dearest friend who provided endless hours of support and love. Finally, George Nassar's kindness, patience, unconditional emotional support, and statistical expertise were invaluable throughout the final months of this degree and dissertation process.

I am forever grateful to my academic mentors who molded me into a young researcher and music educator. Warren Henry inspired and encouraged my love for early childhood music education as an undergraduate. Donna Emmanuel inspired me to live in the moment, and Nate Kruse inspired me to laugh in the moment. I owe deep gratitude to my dissertation committee: Robin Henson, Don Taylor, and Debbie Rohwer. Robin Henson opened my mind to the beauty of statistics, informed decision making, and conceptual thinking. Don Taylor has been an important source of kindness from when he first observed me as a student teacher until the moment he was a part of passing my dissertation. Debbie Rohwer has been an important source of encouragement and a model of strength. She led me to see the empowerment that exists behind research methodology. Finally, I am grateful to all of the researchers whose studies made my dissertation possible, especially Joanne Rutkowski and Graham Welch.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
CHAPTER I INTRODUCTION.....	1
Prologue	1
Singing Ability Defined.....	5
Statement of the Problem.....	5
Methodological Rationale	8
Singing Ability Literature Reviews	11
Need and Purpose of the Study	15
CHAPTER II REVIEW OF LITERATURE	20
History of Singing Ability in America.....	20
Singing Voice Characteristics.....	28
Models of Vocal Development in Music Education.....	33
Empirical Research of Singing Ability	38
Instruction	38
Effect of Practice on Singing Ability.....	42
Age and Grade	42
Gender.....	43
Population	43
Measuring and Scoring Singing Ability	44
Singing Task	47
Music-specific Meta-analyses.....	48
Cognition.....	50

Music Therapy	53
Music Education	58
CHAPTER III METHOD	68
Inclusion and Exclusion Criteria.....	68
Content Validity.....	70
Literature Search Strategies	74
Reference Databases Searched	74
Coding Procedures	74
Definitions of Coding Categories	75
Intracoder Reliability	77
Study Quality	78
Statistical Methods.....	79
Effect Size Calculations	80
Pilot Study.....	86
CHAPTER IV RESULTS.....	88
Publication Bias	107
CHAPTER V CONCLUSIONS	112
Summary of Results	113
Discussion.....	114
Research Question One: Overall Effects	114
Research Question Two: Primary Moderator Variables	117
Age.....	123
Gender.....	125

Publication Bias	126
Study Quality	126
Limitations of the Current Meta-analysis	131
Research Pedagogy and Music Education	132
Implications for Future Research.....	133
APPENDIX A CONTENT VALIDITY PANEL LETTER	137
APPENDIX B GEOGRAPHIC REGIONS OF THE UNITED STATES.....	142
APPENDIX C EXCLUDED ARTICLES BY RATIONALE	144
APPENDIX D PILOT STUDY	149
Pilot Discussion and Limitations	150
Pilot Procedures	151
Pilot Study Results	153
REFERENCES	159

CHAPTER I

INTRODUCTION

Prologue

My senses are focused on the sound coming from in between their small, parted lips. My eyes are listening to their mouths, tongues, posture, evidence of breathing, and body language. My eyes are also engaged with my own body language through my peripheral vision. What am I doing to engage sound? What more could I do to influence that beautiful sound I seek, the sound that is waiting to emit from their tiny bodies? What can I do to encourage development of each child's individual singing ability?

I am listening with my ears. I am listening to the inhalation of air, hoping to hear the intake that indicates an open throat and engagement of diaphragmatic muscles. I am listening to the sound these children are creating by the interaction of their vocal folds and air. Are they singing with their head voices? Have I modeled the phrase appropriately? Is the sound I am hearing appropriately aligning with the image I perceive regarding vowels?

Someone is flat. I cut off the sound with my hands and nonverbally indicate with one finger to listen by pointing to my ear. My other outstretched, open-fingered hand holds the students still with expectation. As the stimulus source, I again model the melodic phrase on a neutral syllable with both hands near my chest, indicating that it is my turn to sing. My voice is light and childlike with only a slight shimmer of a vibrato. For them to sound like me, I must first sound like a perfect version of them. As I finish the phrase, I delicately extend my arms, providing a preparatory beat before lightly bouncing on the ictus with my arms and open hands. The action appears as if I am handing the students the phrase rather than conducting them. Truly,

that is what I am doing. I am passing the music from within myself to the singers. Their job is to receive, process, imitate, interpret feedback, and repeat.

I smile at their response. My nonverbal affirming feedback encourages smiles from their faces and voices. With their smiling comes a textured sound that passes through the hair on my arms as it soothes the surrounding air. The texture is purple, warm, regal, and pure.

The children echoed the phrase successfully, but were they successfully independent of each other? Is each student using his or her head voice? Can each student demonstrate correct melodic contour? If each student can imitate contour, can he or she accurately match pitches, intervals, phrases, and songs?

I need to know. I need to assess. Should I assess informally or formally? For the moment, I make the decision to assess informally. Instead of using precious class time to pull out one of the many measures of singing ability, such as the Singing Voice Development Measure (Rutkowski, 1996), I use my skills of perception and merely listen to evaluate. Really, I am listening not to evaluate, but to know what still needs to be taught and retaught. Most importantly, I need to teach and reteach in the moment.

Feedback is everything. To me, a small moment of feedback means making the decision as to what a student needs to hear in order to improve or excel. To the student, a small moment of feedback might mean one of the only moments during the day when he or she has all of someone's attention. The smile or word of praise in gratitude of independent singing might mean that he or she will love to sing because he or she can sing beautifully. If I say the wrong thing, however, a student could grow up thinking that he or she lacks musical abilities including a pleasant singing voice. He or she may never participate in music outside of radio interactions or shower performances. As Welch (2015) stated,

If the individual has been exposed to an appropriately nurturing environment, considerable singing skills are normally evidenced. However, the experience of negative comments during childhood, particularly from adults such as parents and teachers, can have a life-long detrimental impact on singing behaviors and the realization of musical potential. (n.p.)

While students are singing the phrase again, I quickly run to the corner of my room to retrieve the giant blue ball. I take my place in the middle of the circle and sing the phrase to my (unbeknownst to the students) favorite singer in the class. He is my favorite singer because the students respect him, and because he never fails to provide a beautiful model. As I sing the phrase, I engage him with eye contact while simultaneously bouncing the ball to him. As expected, he beautifully responds and bounces the ball back. I compliment him directly by smiling; I compliment him indirectly by requesting that the rest of the class sound exactly like he does. I then motion for him to take my place in the middle of the circle. Children may respond more confidently and, perhaps, more accurately to another child than they will to me.

I take his spot in the circle. I am now one of the students in everything except for my thoughts and for a brief moment when students look to my face after singing, hoping for positive feedback. After a few girls echo him successfully (with a smile and thank you given as my feedback), he bounces the ball to a little girl who has struggled with engaging her head voice. She has received the ball, and confidently echoes the phrase, but nearly a third interval too high. Her melodic contour is fairly accurate, and she has sung in her head voice. Her pitches are inaccurate, but I offer her praise regarding her correctness. "What a beautiful head voice you've brought to class today. Thank you for sharing!" Her success is one step forward on her path of vocal development, an assessment moment for me, and an important moment for other children

to experience what I am hoping will indicate that all children can sing, and that singing ability is not defined by accuracy alone.

All the children have had a turn, and all the children have received a moment of feedback. Solfege next. Words follow solfege. Songs follow words. More songs follow songs until the children are making music independently and independent of me.

As an educator of young children, my goal has always been to raise (a) independent music makers, and (b) tuneful singers. If a child can speak, he or she can sing. If a child can sing, he or she can make music. Making music is not specific to singing. A singing voice, however, is the only truly melodic instrument that every child possesses without cost. Therefore, it is of primary importance that each child develops his or her singing ability and has awareness of his or her unique ability to sing so that he or she may continue making music independent of me and of formal instruction.

What is singing in and of itself? Singing is an action, a behavior, an expression of emotion, a livelihood – definitions might be endless. Singing may affect language development (Fuchs et al., 2009); singing is an important part of culture and heritage (Klinger, Campbell, & Goolsby, 1998; Tsang, Friendly, & Trainor, 2012); singing might also develop “the capacity to experience one’s own emotions and produce feelings in others” (Fuchs et al., 2009, p. 182). Everyone is capable of singing (Gordon, 2003), but singing exists in children’s lives “because of their biological abilities to discern it, feel it, and express it” (Campbell, 1998, p. 169).

If one were to attempt to express the essence of this education in one word, it could only be *–singing...* Our age of mechanization leads along a road ending with man himself as a machine; only the spirit of singing can save us from this fate. (Bónis, p. 206, 1974)

Singing Ability Defined

The term singing ability is one of several used to label the development of skills and vocal characteristics children may experience through singing instruction. Other terms have included voice production (Buck, 1991), accuracy (Aaron, 1990; Doneski, 2005; Martin, 1991; Mathias, 1997; Miyamoto, 2005; Montgomery, 1988), development (Kim, 2000; McGraw, 1996; Persellin, 2006; Rutkowski, 1986; Rutkowski & Snell Miller, 2003), performance (Gault, 2002; Hedden & Baker, 2010; Jordan-DeCarbo, 1982; Phillips & Aitchison, 1997), achievement (Kane, 1994; Klinger, Campbell, & Goolsby, 1998; Levinowitz, 1987; Mang, 1995; Marshall, 2002; Newell, 2013), and ability (Feierabend, 1984; Michaud, 2014; Phillips, 1985; Stauffer, 1985). Additionally, the terms have been used synonymously across several studies and, sometimes, within the studies themselves. In the current study, the term ability was chosen due to its overarching nature and relationship with all the listed terms. Radocy and Boyle (2003) described music ability as

a broad term defying precise definition, refers to being able to do something, regardless of how a person acquired the necessary knowledge, skills, and experience... and that achievement, while “off to the side,” may be evidence of ability. (p. 384)

Singing achievement is related to singing ability, and singing development describes the growth of ability over time. Vocal production, accuracy, and performance all describe components within singing achievement and development. Singing ability was used, therefore, as an appropriate inclusive term for the current study.

Statement of the Problem

My interest in and passion for developing children’s singing ability has led me to investigate the topic further, and facets of singing ability provided context to the current study.

The research regarding singing ability is, however, extensive. Therefore, a brief overview of several relevant areas are discussed. The history of singing can provide context for the importance of singing as a content area. Empirical research can provide context for illustrating the variables associated with the effect of instruction on singing ability.

Singing was documented as an important educational component in the acculturation of children into society well before the advent of the public school systems. Singing schools were formed to improve congregational vocal performance and were in existence from 1722 until the early 1900s (Britton, 1989) when free public education included music during the normal school day (Phillips, 2014). Singing has also been documented as being universally valued in the elementary music classroom through the 1974, 1986, and 1994 versions of the National Standards as well as with the 2014 Common Core Arts Standards.

Regarding curricula, in the nineteenth century, singing schools and public school curricula addressed vocal technique, music literacy, and choral literature. As the Pestalozzian principals grew in popularity beginning in the 1830s (specifically rote style teaching and sound-before-sight learning), traditional instruction in vocal technique declined within public school curricula (Mark & Gary, 2007). Music literacy, including sight-reading development, also weighed more heavily as more weight was given to what children knew rather than how children sounded. Instead of teaching the art of singing through vocal development exercises, singing was instead taught by singing songs, referred to as the song approach (Phillips, 2014). Eventually, vocal technique found its way back into curricula through private voice instruction (Seashore, 1941) and in choral settings (Phillips, 2014).

Part of music education research has been devoted to singing voice development and teaching children to sing since the origin of public school music programs. Research regarding

the singing ability of children can be found in educational psychology journals dating back to the early 1930s (Jersild & Bienstock, 1934). Articles, dissertations, and book chapters have covered a breadth of topics including singing voice characteristics, cognitive singing mechanisms, physiology of the singing voice, timbre, and vocal range.

Regarding singing instruction, empirical research of singing ability has included research that has investigated effects of instruction on the development of singing abilities in childhood. Several studies have indicated that instruction improved singing ability (Miyamoto, 2005; Paney & Kay, 2014; Roberts & Davies, 1975). Positive effects on singing ability have been found for feedback (Porter, 1977; Welch, Howard, & Rush, 1989), literature dissemination (Apfelstadt, 1984; Klinger et al., 1998), the presence and type of singing instruction (Apfelstadt, 1984; Miyamoto, 2005; Roberts & Davies, 1975; Rutkowski, 1996), and pattern instruction (Jarjisian, 1983; Porter, 1977). Use of intervals and content familiarity has been found to have positive effects on singing ability (Fyk, 1985; Moore, 1994; Reifinger, 2009, 2012). Conflicting effects, however, on singing ability have been found for vocal model (Green, 1990; Small & McCachern, 1983; Yarbrough, Green, Benson, & Bowers, 1991; Yarbrough, Bowers, & Benson, 1992), use of singing voice (Persellin, 2006; Phillips, 2014), use of accompaniment (Atterbury & Silcox, 1993; Guibault, 2004), effect of repetitive practice (Paney & Kay, 2014; Reifinger, 2009; Small & McCachern, 1983; Welch et al., 1989), population (Moore et al., 1995/1996; Moore et al., 2004), performance context (Goetze & Horii, 1989; Moore et al., 2004; Yarbrough et al., 1991), and text (Welch et al., 1998; Gault, 2002; Guerrini, 2006; Rutkowski, 1990). Age and grade have yielded conflicting results regarding singing ability across grade levels (Fyk et al., 1995/1996; Gault, 2002; Goetze and Horii, 1989; Hornbach & Taggart, 2005; Miyamoto, 2005; Yarbrough et al., 1991; Yarbrough et al., 1992; Welch et al., 1998). Studies regarding gender differences in

singing ability have also yielded conflicting results (Apfelstadt, 1984; Goetze & Horii, 1989; Hedden & Baker, 2010; Moore, 1994; Moore et al., 1995/1996; Paney & Kay, 2014; Yarbrough et al., 1992).

Methodological Rationale

In an effort to synthesize all the singing ability research, researchers have compiled literature reviews. It has reached the point, however, when literature reviews may have become inefficient and cumbersome, given the breadth of studies. Meta-analyses have been conducted in other fields as a remedy to this concern.

Before the development of meta-analytic procedures, literature reviews were essential to synthesize bodies of research within and across various fields (Borenstein, 2009). Literature reviews, however, had functional limitations including subjectivity and researcher bias (Borenstein, 2009) as well as conflicting conclusions (Cooper, Hedges, & Valentine, 2009). Additionally, literature reviews were quickly outdated as new research studies were published.

In addition to compiling reviews of literature, researchers were utilizing several other procedures to integrate results (Glass, 1976). Procedures included compiling and sorting studies based on significance, analyzing and grouping study results based on research design quality, and compiling and reanalyzing raw data from multiple studies. While these procedures included the statistical rigor that was not specific to literature reviews, they were also problematic. Compiling and sorting studies based on significance omitted the need to find meaningful effects across studies. Analyzing and grouping study results based on research design quality provided a means to compare like studies but at the cost of severely limiting comparisons. Compiling and analyzing raw data from multiple studies was effective, but time consuming and depended upon

the availability of data (Glass, 1976). Thus, more comprehensive methods of synthesizing information were required. As Glass (1976) stated,

we need more scholarly effort concentrated on the problem of finding the knowledge that lies untapped in completed research studies. We are too heavily invested in the pedestrian reviewing where verbal synopses of studies are strung out in dizzying lists. (p. 4)

Therefore, meta-analyses were systemized as an answer to the possible limitations of literature reviews. The most recent edition of the APA Manual (2010) referred to meta-analyses as literature reviews because they are “critical evaluations of material that has already been published” (p. 10). The manual further specified that meta-analyses also utilize quantitative procedures to support results. “The armchair literature review in which one cites a couple dozen studies from the obvious journals can’t do justice to the voluminous literature of educational research that we now confront” (Glass, 1976, p. 4). Glass (1976) specifically defined meta-analysis as a “statistical analysis of a large collection of analysis results for individual studies for the purpose of integrating the findings” (p. 3). Statistically synthesizing information can provide a holistic picture of a given body of research, and it can also be used to provide evidence for effective treatments, to improve behavioral science theories (Froehling, 2010), and to influence policymaking decisions (Borenstein, 2009; Froehling, 2010).

Providing an overall effect size of a body of research as well as comparing effect sizes within a body of research has been a primary component of performing meta-analyses. Statistical significance, however, is secondary to effect size. Null hypothesis statistical testing (NHST) may be used to find relationships or differences within or between dependent and/or independent variables but is not substantial information alone (APA, 2010). Unfortunately, researchers and readers alike have been relying on and reporting statistical significance alone to provide

important information (APA, 2001; APA, 2010; Henson, Hull, & Williams, 2010), because statistical significance measures the probability of a relationship between variables, but not the strength, meaning, or importance of the relationship (Cohen, Cohen, West, & Aiken, 2003). “For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in the Results section” (APA, 2001, p. 25).

Cohen, Cohen, West, and Aiken (2003) defined effect size as, a measure of the magnitude of a relationship, either in the units of the original measures such as B_{xy} or mean differences, or in standardized units such as r , r^2 , R , β , or R^2 (as contrasted to measures of the confidence of its exceeding some null value, i.e., its statistical significance). (p. 673)

Henson (2006) further elaborated,

if the null suggests that three group means are the same, for example, an effect size may quantify the degree of difference between the three means or, perhaps, the magnitude of differences between each of the three pairwise combinations of means. (p. 602)

While the p value demonstrates statistical significance, the effect size demonstrates practical significance (Huck, 2012).

Effect size indices have varied across music-specific meta-analyses due to research design (experimental or correlational) and group standard deviation (if homogeneity of variance assumptions were met or violated) (Ellis, 2010). Ellis (2010) documented 24 common effect size indices; Cohen’s d (d), Glass’s delta (d), and Hedges’ g (g), however, were the most commonly found indices in music-specific meta-analyses that compared means between groups.

Hedges' g is nearly identical to Cohen's d , with the only difference existing in the denominator of the equations. In Cohen's d , the mean difference is divided by the pooled standard deviation of both groups, whereas in Hedges' g , the mean difference is divided by the pooled standard deviation of both groups weighted by each group's sample size. Both Glass's delta and Cohen's d assume equivalent groups, whereas Hedges' g is most appropriate for groups of difference sizes. Hedges' g was used in the current meta-analysis due to inequivalent group sizes within individual studies.

Singing Ability Literature Reviews

The issue of efficiently synthesizing research is neither specific to educational psychology nor to the 1970s. In the field of music education, observations might be made regarding large-scale research integration of a continually growing pool of published research. *The Handbook of Research on Music Teaching and Learning* (1992), *The New Handbook of Research on Music Teaching and Learning* (2002), and the two volumes of *MENC Handbook of Research on Music Learning* (2011) represent publications that include large-scale literature reviews of current research related to and within music education. Content analyses of music education research have also been published on topics including string education (Kantorski, 1995), student teaching (Silveira & Diaz, 2014), article subject (Kratus, 1992), article citations (Brittin & Standley, 1997; Killian, Liu, & Reid, 2013; Sample, 1992; Schmidt & Zdzinski, 1993), and published research characteristics (Brittin & Standley, 1997; Ebie, 2002; Yarbrough, 1984; Yarbrough, 2002).

Efforts to synthesize the ever-growing body of research regarding singing ability in children have been illustrated through several prominent literature reviews (Atterbury, 1984; Goetze, Cooper, & Brown; 1990; Welch, 1979a; Welch, 1979b). The literature reviews can be

considered prominent on the basis of frequency of citations in existing singing ability literature. The reviews provided context for subsequent research studies and is revisited in the current study post-analysis for a narrative comparison of the current studies' effect size results and prominent literature review conclusions. Therefore, a brief discussion is warranted regarding categorization of variables and conclusions.

Reviews of literature have included categories regarding the factors and processes that were found to affect accuracy (Atterbury, 1984; Goetze et al., 1990; Welch, 1979a, 1979b). Singing range was the primary topic of one literature review (Welch, 1979a) and was a category or sub-category in other literature reviews (Atterbury, 1984; Goetze et al., 1990). Several literature reviews have included a category regarding the effect of training on singing (Atterbury, 1984; Goetze et al., 1990; Welch, 1979b). Goetze et al. (1990) further subdivided research based on the effect of variables on accuracy including model, singing task, text, individual versus group singing, accompaniment, and remedial training. Some articles have included a review of singing range specific to school literature (Atterbury, 1984; Welch, 1979a). Defining key terms by describing the lineage of terminology was also a common category. Terms included tone-deafness versus tone-dumbness (Welch, 1979b).

Both of Welch's (1979a, 1979b) literature reviews included conclusions that were illustrated by figures. One included a figure showing, "the pitches reported as being common for each age group by aggregating the research findings" (Welch, 1979a, p. 26). Another illustrated the "changing definition of the poor pitch singer" (Welch, 1979b, p. 56).

Contrasting conclusions were made in the reviews of literature regarding published children's literature. Welch (1979a) concluded that published children's literature used in public school settings was being pitched too high. Atterbury (1984) stated that published elementary

music textbooks included literature with inappropriate singing ranges that could be detrimental to singing voice development. Atterbury (1984) elaborated,

music books for young children have evolved from collections of songs to organized music curriculum sources, but the conceptual teaching approach has not been accompanied by the selection of organized sequential song material related to *how* children learn to sing. (Atterbury, 1984, p. 60)

In addition, Goetze et al. (1990) suggested revisiting the range of literature and viewing the topic in terms of vocal register. Goetze et al. (1990) suggested that previous studies may have reported the pitches that were most comfortable for children to sing aside from a child's use and development of head voice.

Contrasting conclusions were made in the reviews of literature regarding vocal range overall and with regard to singing tasks. The expansion of a child's vocal range has been cited as being related to age (Goetze et al., 1990; Welch, 1979a) and development of register (Goetze et al., 1990). Although Atterbury (1984) suggested that younger children should receive instruction with songs containing more descending intervals than ascending intervals, Goetze et al. (1990) cautioned against making specific generalizations without concurrent studies. Instead, Goetze et al. (1990) advocated for assessments that measured accuracy of small range tonal patterns.

Several other conclusions were made in the reviews of literature. Singing accuracy may have been more affected by environmental factors than by innate pitch discrimination (Goetze et al., 1990; Welch, 1979b). Pitch discrimination skills may be positively affected by singing instruction (Goetze et al., 1990; Welch, 1979b). Tonal memory development may be related to singing voice development (Atterbury, 1984). Other conclusions suggested that the female vocal model may encourage singing accuracy more than the male vocal model (Goetze et al., 1990),

children may sing more accurately when asked to sing independently than when singing with a group (Goetze et al., 1990), breath control training may encourage singing voice development (Goetze et al., 1990), and that the socio-economic status of a child might predict singing ability where “poorer homes produce fewer singers” (Welch, 1979a, p. 26),

Atterbury (1984) concluded in a review of literature that instruction should include opportunities to experience the speaking and singing voice (Atterbury, 1984). Atterbury (1984) suggested that instruction should omit harmonic accompaniment during melodic contour development and register development of young children. Goetze et al. (1990), however, suggested that results were inconclusive and simple harmonic accompaniment might positively affect singing ability.

Goetze et al. (1990) concluded in a review of literature that multiple complications of synthesizing singing ability research including the quantity of instruments and procedures used to measure components of singing ability. Literature reviews suggested that there was a need for alignment among studies regarding terminology (Welch, 1979a) and methods of measurement (Goetze et al., 1990). The need is reflected again in subsequent literature reviews devoted to the assessment of singing ability (Salvador, 2010; Welch, 1994).

Literature reviews on singing ability have provided important synthesis to gather and disseminate the information to improve instruction; literature reviews, however, are based on the interpretation of a result and conclusion section of a primary research study, and are therefore secondary research studies. As illustrated above, singing ability literature reviews have made both similar and contrasting conclusions regarding the same skills and characteristics of singing ability, but the conclusions have been based on different studies, therefore possibly sacrificing clarity. In contrast, meta-analyses can provide a systematic and empirical synthesis of research

that utilizes primary data in order to illustrate a picture of a body of information. “Understanding and dissemination of the existing body of research literature in music education/therapy would seem to benefit greatly from the application of meta-analysis” (Standley, 1996, p. 103). Meta-analyses have been conducted regarding music cognition and music therapy. Several meta-analyses have also been conducted within the field of music education regarding conducting (Acklin, 2010) and sight-singing (Mishra, 2013, 2014). Therefore, as the large body of singing research continues to grow, there now exists a need for statistically synthesizing past singing ability research for the purpose of informing future singing ability research and disseminating informed and transparent results and conclusions.

Need and Purpose of the Study

First and foremost, the purpose of a meta-analysis is to calculate an overall effect size for a heterogeneous pool of research. The meta-analytic process then includes recalculating effect sizes for individual studies to enable study-to-study comparison. Differences in effect sizes are predicted to exist between studies due to variables, or moderators, including sample characteristics, measurement characteristics, design characteristics, source characteristics, and study quality (Card, 2002). Moderator analyses are conducted to examine if effects can be attributed to heterogeneity. Defined by Card (2002), “moderator analyses attempt to explain heterogeneity among effect sizes by evaluating whether coded study characteristics systematically predict variation in effect sizes across studies” (p. 63). Therefore, the needs for the current study are discussed in terms of overall effect and by moderator variables.

There is a need to investigate the effects of singing instruction overall and through moderator analyses. Two parameters were important in developing the purpose of the current study. The first parameter is the importance of appropriate singing instruction. Children exposed

to music in a formal setting are provided with instruction that could encourage or discourage musical growth (Welch, in press) as evidenced by surveying adults' perceptions of their singing ability (Welch, 1979a).

Overall, the prime source of singing “failure” for an individual is often a particular moment in childhood and/or adolescence when there is a mismatch between developing singing competences and a set of singing tasks. (Welch, 2015)

Singing ability has been determined to be developmental, not innate (Atterbury, 1984; Goetze et al., 1990; Welch, 1979b, 2015). Singing ability, therefore, may have been dependent upon singing instruction and intervention for development beyond merely singing songs during the music class (Goetze et al., 1990). The quantity, types, and variations of singing instruction can be observable through the various pedagogical methods of instruction and empirical research studies on the topic.

The second parameter of the current study regards the isolation of the chosen age group of 5 to 11 years. Ages 5 through 11 have been determined important in the development of the singing voice. Bella and Berkowska (2009a) claimed that by the time a child is five years old, “brain circuitries needed for achieving adult proficient singing are already in place” (p. 280). Between the ages of 9 and 10, hormones have been shown to affect singing voice development (Doscher, 1994). While hormones may affect singing voice development, children have still been considered prepubescent in singing research through elementary school as long as the participants were in an elementary school.

The moderator analysis includes nine moderator variables. Four primary moderator variables emerged due to the breadth of research involved in the above-mentioned parameters, singing instruction and defined age group. The moderator variables include age, gender,

measurement instrument, and singing task. An additional five moderator variables comprise a secondary group of study characteristics, including publication source, publication year, population, research design, and treatment period.

There is a need to investigate the effects of age and gender through a moderator analysis due to the contradictory results found in the literature with these variables. Singing ability studies have often used age and gender as independent variables. Results from studies have reported decreased frequencies of inaccurate singers (also referred to as monotone singers or poor pitch singers) as grade level increased (Bentley, 1969; Davies & Roberts, 1975). Studies have also reported trends that have indicated lower frequencies of inaccurate singers with girls than boys (Bentley, 1969; Davies & Roberts, 1975). Studies regarding the effect of instruction on singing ability for grade and gender have been inconclusive. Several studies have found that overall singing ability improved across grade levels (Fyk, et al. 1995/1996; Goetze and Horii, 1989; Hornbach & Taggart, 2005; Miyamoto, 2005), while other studies did not find differences in singing accuracy scores across grade levels regarding vocal model (Yarbrough et al., 1991; Yarbrough et al., 1992) and response mode (Yarbrough et al., 1991). Additionally, several results have indicated a tendency for girls to have higher assessment mean scores than boys in Kindergarten through third grade (Goetze & Horii, 1989), in third and fifth grade (Moore et al., 2004), and with ages six and nine (Moore et al., 1995/1996). In contrast, other studies have found no difference in singing ability between boys and girls (Apfelstadt, 1984; Hedden & Baker, 2010; Moore, 1994; Paney & Kay, 2014; Yarbrough et al., 1992).

There is a need to investigate the effects of singing ability measurement instruments through a moderator analysis due to the variability of tests that have been used. The quantity and type of singing ability assessments has been considered problematic in singing ability research

(Goetze et al., 1990; Salvador, 2010; Welch, 1994). Researchers have used pre-existing instruments, newly created instruments, used themselves to measure singing ability, and used computer software to measure ability (Goetze et al., 1990; Salvador, 2010). The high quantity and variation of assessments has decreased the likelihood for creating a common thread of singing ability research.

There is a need to investigate the effects of singing tasks through a moderator analysis due to the variability of tasks that have been assessed. The variation of singing tasks assessed when measuring children's singing ability has been considered problematic in singing ability research (Goetze et al., 1990; Rutkowski, 1996; Salvador, 2010; Welch, 1994). Singing task has been shown to affect measurement outcomes of singing ability (Moore, 1994; Reifinger, 2009, 2012). Although research has indicated that children would be most successful in assessments that included small familiar phrases or patterns on a neutral syllable (Guerrini, 2006; Moore, 1994), several tasks have required children to sing entire songs (Rutkowski, 1996) with and without text. Singing task may deceptively influence the effectiveness of a treatment (Welch, 2015).

There is a need to investigate the effects of study characteristics including publication source, publication year, selected population, research design, and treatment period through moderator analyses (Card, 2012) due to the variability of each characteristic across singing ability studies. These specific characteristics might influence effects due to factors associated with internal and external validity. "Logic dictates that such factors do exist and will lead to variations in the magnitude of the effect" (Borenstein, 2009, p. 70). Other music-education-specific meta-analyses have also completed analyses for publication source (Mishra, 2013, 2014; Standley, 1996, 2008), publication date (Butzlaff, 2000; Standley, 1996; Standley, 2008),

population (Standley, 2008), research design (Standley, 1996), and treatment period (Standley, 1996; Standley, 2008).

The purpose of the study was to address the varied and somewhat stratified study results within the area of singing ability and instruction by statistically summarizing the data of related studies through meta-analysis. Toward this purpose, three meta-analysis research questions were developed. These included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children ages 5 through 11 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task; and (3) what are the statistical effects of study characteristics within and across the following secondary moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period.

CHAPTER II

REVIEW OF LITERATURE

Several facets of singing ability provide context to the current meta-analysis. The history of music education includes details regarding singing ability since the 1700s. The nature of the child singing voice includes physical and aural characteristics. The cognitive processes involved in singing ability development include models of vocal development. Additionally, empirical research has investigated characteristics of singing instruction. The current review of literature addresses singing ability and meta-analyses across five sections: (a) the history of singing ability in America, (b) singing voice characteristics, (c) models of vocal development in music education, (d) a literature review of empirical research categorized by moderator variables, and (e) a literature review of music-specific meta-analyses. Due to the breadth of results within the empirical research section, the included studies regarding empirical research have been categorized by the following moderator variables: (a) instruction, (b) grade, (c) gender, (d) population, (e) measuring and scoring singing ability, and (f) singing task.

History of Singing Ability in America

Singing was documented as an important educational component in the acculturation of children into society well before the advent of the public school systems. Since the early 1700s, singing has been recorded as an important component within American life (Britton, 1989; Mark & Gary, 2007). Singing served three primary purposes for the early settlers: (a) to participate in community music making, (b) to glorify God through worship, and (c) to uphold the educational practices of the wealthy. The extent of purposes tended to vary by region or religious affiliation.

New England settlers (including, but not limited to, the Amish, Mennonites, and Shakers) and the Moravians deeply valued community music making (Mark & Gary, 2007). Although

Moravians encouraged private music instruction, children developed musicality with or without systematic instruction; music making was an important part of daily Moravian life. Music was also an important aspect of daily life for the New Englanders, though they did not value private instruction as did the Moravians.

The Moravians, who settled in Pennsylvania, deeply valued singing for worship, singing as a performance medium based on the German choral traditions, and singing for community music making participation and culture conservation (Mark & Gary, 2007). In Moravian schools, children participated in a daily 45-minute assembly, known as Singstunde. Singstunde was a time of gathering to learn and rehearse religious church music. Moravian music educators also valued solo performance of sacred music.

Southern settlers appreciated and celebrated both secular and sacred singing as community activities. Moravian worship largely influenced Methodist sacred hymn singing (Mark & Gary, 2007), while European literature dominated the secular repertoire of the south. Vocal music training occurred primarily through private instruction for the wealthy until singing schools spread from the north.

Before the eighteenth century, early American communities emphasized survival and economic growth while valuing music and singing as a worshipful practice in the eighteenth century. However, by the eighteenth century, the quality of singing had declined. As America prospered, public desire for quality instruction increased (Mark & Gary, 2007). Instructional singing books were published and disseminated including Tuft's (1721) *An Introduction to the Singing of Psalm Tunes* and Walter's (1721) *Grounds and Rules of Musick Explained*. Singing schools were also formed to meet congregational needs by improving the quality of vocal performance (Britton, 1989).

The first singing school has been traced back to 1722. Several singing school advocates including Theodorus Frelinghuysen, Jonathan Edwards, and George Whitefield, furthered the spread of the singing schools. Not only was there a documented improvement in congregational singing, but singing schools were “also generally acknowledged to have had a positive effect on moral standards and the place of the church in the community” (Mark & Gary, 2007, p. 78). Therefore, singing schools developed an understood importance within communities. “The singing school and its music together constitute an indigenous cultural manifestation that met the religious and esthetic needs of a nation still living under pioneering conditions” (Britton, 1989, p. 38). Singing schools were privatized events in which students of all ages paid to attend; therefore, the general need for singing schools became increasingly obsolete when free public education included music during the normal school day (Phillips, 2013). Although the last singing school closed in the early 1900s, the singing schools were essential to the development of today’s school music experience (Atterbury, 1992) and exist in history as laying the foundation of music education (Britton, 1989).

Both the singing schools and the newly established public school curricula addressed several singing components including vocal technique, music literacy, and choral literature. Vocal technique instruction was implemented for improving the singing voice through technical exercises and breathing exercises based on the Italian bel canto method (Phillips, 2014). Music literacy was an encouraged component of singing schools and public schools. Also, choral literature singing was considered a goal of singing, rather than as a part of the singing instruction process. The general concept of music education was that it was thought to benefit children “physically, intellectually, socially, and morally” (Colwell & Heller, 2003, p. 234).

Several known leaders of vocal pedagogy continued the tradition of vocal technique instruction that had evolved from the singing school and public school tradition. The vocal pedagogy leaders included members of the Transcendental Generation: William Channing Woodbridge, Elam Ives, and Lowell Mason (Colwell & Heller, 2003; Mark & Gary, 2007), each of whom produced additional educational tools to develop singing instruction.

Woodbridge adapted Pfeiffer and Nageli's Pestalozzian principles for American music education (Mark & Gary, 2007). Ives first applied the Pestalozzian principles as a pedagogical experiment with Woodbridge. Ives' *American Music Singing Book* (1830) and *The Juvenile Lyre* (1831) were published as a result. Woodbridge's adapted principles were reflected in Mason's *The Song Garden* textbook series (1866) and published in Mason's (1843) *Manual of the Boston Academy of Music for Instruction in the Elements of Vocal Music on the System of Pestalozzi*. The Pestalozzian principles were unique in that they were one of the first documentations of rote style teaching (teaching a song phrase-by-phrase without written notation) and sound-before-sight learning (a teaching technique to develop music literacy through exposure to aural components of music before transferring to written notation).

Despite singing schools' prior use of texts that included a strong emphasis on vocal technique instruction, the work of Mason, Ives, and Woodbridge influenced a transition within music instruction that turned away from singing school pedagogy toward a more experiential-based use of Pestalozzian principles (Colwell & Heller, 2003; Mark & Gary, 2007). Although vocal technique instruction remained important to both pre- and post-application of Pestalozzian principles, the new Pestalozzian method provided students with a greater opportunity to be active participants in their learning where they discovered theoretical components of music through experience.

Near the end of the nineteenth century, singing for beauty, pleasure, God, and community no longer provided sufficient justification for music to remain in the public schools. Another transition in music education was initiated. Due to an increased need for displaying the importance of music, public school administrators questioned the integrity of the Pestalozzian methods, because students seemed neither to be progressing nor growing into musically literate young people (Mark & Gary, 2007). As a result, music literacy and sight-reading were given primary weight within music instruction, and music textbooks from members of the Gilded Generation (Colwell & Heller, 2003), such as Holt and Tufts' *The Normal Music Course*, were published in 1883. Other pedagogues including Luther Whiting Mason and Benjamin Jepson called for an increase in learned repertoire. On Luther Whiting Mason's teaching, Mark and Gary (2007) stated,

San Francisco second graders were also expected to sing scales with syllables and copy notes, rests, staves, and clefs on to their slates. This approach, used by advocates of the "song method," emphasized the facts of music rather than the music itself. It was probably not very different qualitatively from the note method of those who advocated drill in note reading with progressively difficult exercises. Both, evidently, felt the need to be "scientific." (p. 198)

Benjamin Jepson also advocated for the use of song material, but declared that the objective of music instruction was to develop sight-reading skills. Sight-reading development was accomplished using a balance of exercises and repertoire presented in close succession of one another. The idea was in direct opposition to previous practices where students practiced drills long before the introduction of literature.

The Progressive Music Series (1914) was a public school music text edited by the Music Supervisors National Conference's Osbourne McConathy, Edward Birge, and W. Otto Miessner and influenced by the writings of John Dewey. This text was one of the first of many that aligned with the new song approach by completely omitting exercises, scales, and other ways of building vocal technique separate from singing songs. Music literacy was taught within the context of literature, and singing was taught by singing songs (Phillips, 2014).

Several other texts were published over the course of the next few decades, each of which focused on teaching music literacy from within the context of literature – the process labeled as the song approach (Phillips, 2014). *Songs of Childhood* (1923) was a text with similar formatting as the *Progressive Music Series* (1914). *The World of Music* (Glenn, 1936) emphasized movement through Dalcroze eurhythmics. Specific people also became supporters of the song approach including William A. Hogdon and Sterrie A. Weaver (Mark & Gary, 2007).

Contextualizing music literacy was under development in the public school setting. As educators continued critiquing the music education effort, a return to the original philosophy of singing schools to teach vocal technique occurred. Educators such as William Tomlins (an English choral director) advocated that beautiful singing and tone quality, as developed through vocal technique instruction, belonged at the core of a music curriculum. Tomlins observed issues with education curricula that only focused on the song approach. Tomlins is still revered as an important historical figure in “standing for teaching children to sing” (Mark & Cary, 2007, p. 210).

In spite of vocal technique advocates, including Tomlins, the song approach was a component of elementary music curricula that continued to pervade most textbooks, methods, and philosophies in the elementary music classroom (Phillips, 2014). Seashore (1941) even

regarded the pedagogical shift from vocal technique instruction to the song approach as if vocal technique was altogether diminishing from the public school setting. Seashore (1941) also discussed how vocal technique instruction was most appropriate within a private lesson setting starting at age six: “the ideal of the school situation which I have pictured will play happily into her [private music teacher] hands by furnishing a background for a systematic study of voice or a particular instrument” (p. 22).

Sputnik motivated music education to seek another transition in music teaching in the 1960s (Mark & Cary, 2007; Phillips, 2014). American adaptations based on approaches inspired by Zoltan Kodály, Carl Orff, and Emile Jacques-Dalcroze provided what was considered to be new, inventive methods for childhood music literacy teaching (Phillips, 2014). Each method was based on the philosophy that every child was innately capable of making music. Of the three methods, the method based on Kodály’s philosophy was most closely aligned with the song approach, while the methods based on the philosophies of Dalcroze and Orff advocated for exploration above music literacy.

While the song approach had become commonplace within public school curricula in the early- to mid-twentieth century, the vocal technique instruction that was still taught within sacred choral environments could be attributed to vocal pedagogues such as Ruth Krehbiel Jacobs and Helen Kemp (Phillips, 2014). Vocal technique instruction again gained momentum in the 1980s with the popularization of secular children’s choirs by the American Choral Directors Association and singing instruction advocates such as Doreen Rao, Mary Goetze, and Jean Ashworth Bartle (Phillips, 2014). Gordon’s music learning theory also encouraged vocal technique instruction with pre-kindergarten children and, later, in the elementary music classroom (Gordon, 2003).

Singing was universally valued in the elementary music classroom as demonstrated through the 1974, 1986, and 1994 versions of the National Standards as well as with the 2014 Common Core Arts Standards. The first edition standards (Lehman et al., 1974) referred to singing with regard to application of the listed outcomes. The second edition (Hoffer et al., 1986) categorized standards into performing/reading, creating (previously labeled as organizing), listening/describing, and valuing. Valuing was an additional category that closely related to the sixth outcome in the first edition (Lehman et al., 1974). The valuing category related to how students should understand the role of music in society. The valuing category stated that children: “demonstrate an awareness of music as a part of every day life, enjoy singing, moving to music, and playing instruments along and with others, and respect music and musicians” (Hoffer et al., 1986, p. 19). In the 1986 edition, singing was not documented as an outcome.

Content standards in the 1994 National Standards were a relabeling of what had been previously referred to as outcomes. Singing was included in Content Standard 1: “Singing, alone and with others, a varied repertoire of music” (Consortium of National Arts Education Associations, 1994). This was a change from *The School Music Program* (Lehman et al., 1974) Outcome 1: “Is able to make music alone and with others,” and the second edition of *The School Music Program* (Hoffer et al., 1986) Outcome 1: “Are able to make music, alone and with others.”

In the National Core Arts Standards, Anchor Standard 1, “Generate and conceptualize artistic ideas and work” (State Education Agency Directors of Arts Education, 2014b, n.p.) included language most similar to Outcome 1 and Content Standard 1, but (similar to pre-1994 standards) did not specifically mention singing. Creative Anchor Standard 1 in Pre-K stated: “With substantial guidance, explore and experience a variety of music” (State Education Agency

Directors of Arts Education, 2014a, n.p.). For Kindergarten, the standard stated, “with guidance, explore and experience music concepts” (SEADAE, 2014a, n.p.). Subsequent standards progressively deviated from the Pre-K standard.

Although the word, singing, was omitted from the newly written Common Core Arts Standards, singing was used in several pieces of the Music Model Cornerstone Assessments for General Music Grade 2 as part of the Performance Assessment Task and Task-specific Rubrics. Under “Key vocabulary that students will use fluently throughout the assessment task” (State Education Agency Directors of Arts Education, 2013, n.p.), voice type (singing voice/speaking voice) is stated as one of several vocabulary words with which students must be fluent. Singing is also mentioned within the instructions with language similar to the previous standards regarding the inclusion of a variety of song literature and singing performance.

Singing Voice Characteristics

The child’s singing voice includes physical and aural characteristics. Physical characteristics have been identified by formant (Doscher, 1994). Aural characteristics have included formant frequencies (Doscher, 1994), timbre (Howard, Williams, & Herbst, 2014; Sergeant & Welch 2009), voice source production (Barlow & Howard, 2005), and singing range (Barlow & Howard, 2005; Cooksey, 1993; Doscher, 1994; Moore, Fyk, Frega & Brotons, 1995/1996; Pedersen, 1997; Sergeant & Welch, 2009; Wassum, 1979; Welch, 1979).

While children and adults have been cited as possessing similar singing mechanisms, the mechanisms in children have been said to continue to grow and develop through and beyond puberty, as demonstrated through voice change research (Doscher, 1994). Unfortunately, far less has been documented about specific physiological components of the child’s singing voice, because it has been more difficult to gather those data from children (Pedersen, 1997).

Understanding specifications involving children's physiological mechanisms may require invasive procedures. Therefore, most of what is known has been collected aurally.

Researchers have discussed formants in research when describing the physiology of the singing voice. Doscher (1994) defined a formant of the vocal resonating system as “a specific concentration of energy within the vocal sound wave” (p. 133). Formant frequencies, also referred to as resonance frequencies, are modified by adjustments made to physical pieces that work together to make sound, including body parts used when forming vowels (lips/tongue), raising and lowering the soft palate, and opening or closing the throat. All movements affect how sound resonates, thus affecting the formant frequencies. Formants have been categorized as either vowel or quality formants, and there are at least five: f^1 and f^2 have been associated with vowel color, while f^3 , f^4 , and f^5 have been associated with tone quality (Doscher, 1994).

A child's singing voice may sound different from an adult's singing voice, often perceived as having a ring that has been shown to be facilitated by the epilarynx and pharynx (Howard et al., 2014). By comparing the spectrum of pitches, Howard, Williams, and Herbst (2014) investigated acoustical differences between children who demonstrated a ring and children who did not demonstrate a ring. Results suggested an isolation of a ring to the “singer's formant cluster region, centered around 4 kHz” as well as “in the region around 7.5 – 11 kHz where a significant strengthening of harmonic presence is observed” (Howard et al., 2014, p. 161).

Characteristics of the singing voice may be described in terms of voice source production. Voice source may be defined as the internal physiological pieces of what works together to create sound before passing through the mouth including, but not limited to, the pharynx, larynx, epiglottis, and vocal folds (Barlow & Howard, 2005). Research regarding a

child's singing voice source production has focused on classifying voices as well as investigating differences in the produced sound across gender and training (Barlow & Howard, 2005).

Barlow and Howard (2005) found that voice source production was similar for trained and untrained female child singers in the lower register (chest voice) and different between trained and untrained female child singers in the upper range. Differences were found between trained and untrained males throughout the registers. Untrained girls tended to press their voices, while untrained boys tended to have breathier higher pitches. Fewer differences were found between trained girls and boys. The implication was that vocal training, particularly of the bel canto tradition, may decrease differences between genders and improve vocal quality.

Children have been classified as having an unchanged voice based on frequency (Hz) and closed quotient means. Cooksey (1993) documented unchanged male voices as those with a speaking voice fundamental frequency of approximately 220Hz. Pedersen (1997) documented unchanged female voices as those with a speaking voice fundamental frequency of approximately 256Hz.

Physiological and aural similarities between genders may have important implications for singing instruction. Several studies have found that few vocal developmental differences exist between genders before puberty, specifically regarding pitch range and intensity (Barlow & Howard, 2005; Doscher, 1994; Sergeant & Welch, 2009; Wassum, 1979). Regardless of physiological similarities, Sergeant and Welch (2009) found timbre differences to be audible enough for musicians to predict gender in song recordings.

Differentiating between prepubescent and pubescent children may be problematic without the measurement of specific hormones or other difficult-to-obtain-permission measures. For females, researchers have defined the prepubescent state as the time occurring prior to the

first menstruation (Howard et al., 2014). Pedersen (1997) documented that the fundamental frequency of the speaking voice was related to estrogen levels in prepubescent females. After comparing estrogen levels with females' fundamental frequencies, Pedersen (1997) predicted that unchanged female voices would lie between 256 Hz–241 Hz. Since Pedersen's (1997) publication, several researchers have used his cutoffs (Barlow & Howard, 2005; Sergeant & Welch, 2009). For males, Cooksey (2000) identified a classification system used to define the male transforming voice from prepubescent to pubescent. Cooksey (2000) labeled the unchanged male voice stage as the Premutation Stage; the stage included an average range from A3–F5, and a speaking voice fundamental frequency of approximately 220 Hz–260 Hz. Cooksey's (2000) parameters were supported by several of his earlier studies (i.e., Cooksey, 1993; Cooksey, Beckett, & Wiseman, 1984).

Vocal range in childhood has been an important topic of singing instruction research due to pedagogical implications. Appropriately pitched literature has been linked to increased pitch accuracy (Welch, 1979). Although there has tended to be no statistical differences in range between prepubescent boys and girls (Sergeant & Welch, 2009; Wassum, 1979), statistical differences have been found by grade level suggesting that a child's range progressively widens as they age (Sergeant & Welch, 2009; Wassum, 1979).

Children's range has been discussed regarding musical range and physiological range (Phillips, 2014). A child's physiological range includes all of the pitches that children are capable of reaching. A child's musical range, however, includes all the pitches that children are capable of reaching and matching with accuracy.

Wassum (1979) noted a general physiological vocal range by grade level based on data collected from 1,062 assessments over a five-year data collection process. Results were written

using Helmholtz pitch notation where c' is the same pitch as middle c , and c'' indicates an octave higher. A first grader has a minimum physiological range of $c' - c''$ and an extended physiological range of $a - e''$, a second grader has a minimum physiological range of $a - c''$ and an extended physiological range of $f - e''$, a third grader has a minimum physiological range of $g - e''$ and an extended physiological range of $f - a''$, a fourth grader has a minimum physiological range of $g - g''$ and an extended physiological range of $f - a''$, a fifth grader has a minimum physiological range of $g - a''$ and an extended physiological range of $e - c'''$, and finally, a sixth grader has a minimum physiological range of $f - c'''$ and an extended physiological range of $d - c'''$.

Phillips (2014) has recommended musical ranges and tessituras by grade level that have been generalized from empirical studies and pedagogical observations. A tessitura has been defined as “the general region of a vocal part within which most of the pitches fall” (Phillips, 2014, p. 97). Recommendations were also written using Helmholtz pitch notation where c' is the same pitch at middle c . A first grader has a musical range of $c' - c''$ with a tessitura falling between d' and a' , a second grader has a musical range of $b - d''$ with the tessitura falling between d' and b' , a third grader has a musical range of $b\text{-flat} - e\text{-flat}''$ with the tessitura falling between d' and c'' , a fourth grader as a musical range of $a - e''$ with the tessitura falling between d' and d'' , a fifth grader has a musical range of $a\text{-flat} - f''$ with the tessitura falling between d' and d'' , and finally, a sixth grader has a music range of $g - g''$ with the tessitura falling between d' and d'' .

Moore, Fyk, Frega, and Brotons (1995/1996) documented interval performance trends that suggested which intervals students may be more or less successful reproducing. The hierarchy of intervals, from most successful to least successful, was listed as: Unison ($C4$),

Unison (C5), descending minor third, descending perfect fourth, ascending major second, descending major third, descending major second, ascending minor third, ascending major third, descending perfect fifth, ascending perfect fifth, ascending perfect fourth, descending minor second, descending octave, ascending minor second, and most difficult, descending octave (Moore et al., 1995/1996).

Models of Vocal Development in Music Education

Swanwick (1994) suggested that in order to teach singing effectively, there must first be an understanding of the cognitive mechanisms responsible for developing intonation.

Mechanisms have been described through models of vocal development. Theoretical models represented in empirical research have described stages of singing voice development (Rutkowski, 1990; Welch, 2000). Models have also included the sequences of learning in singing voice development (Berkowska & Bella, 2009; Phillips, 2014; Phillips, Williams, & Edwin, 2012).

Two models of singing voice development have been used to categorize children based on the use of their singing voice: (a) excluding intonation and melodic contour (Rutkowski, 1996) and (b) including intonation and melodic contour (Welch, 1998; 2000). Rutkowski (1996) used her model of vocal development to establish a singing voice rating scale (Singing Voice Development Measure) that categorizes children across or in between five stages: (1) Pre-singer, (2) Speaking Range Singer, (3) Limited Range Singer, (4) Initial Range Singer, and (5) Singer. The scale also includes half points for children whose development was in between stages. A Pre-singer is described as a child who “does not sing but chants the song text” (p. 207). A child is considered a Singer if he “exhibits use of extended singing range (sings beyond the register lift: B3-flat and above)” (p. 207). Stages 2, 3, and 4 categorize the transition from Pre-singer to

Singer based on the child's ability to use consistently his or her singing voice (opposed to speaking voice) and based on the child's expansion of his or her singing range.

Welch's (1998) Model of Vocal Pitch-Matching Development includes four phases, two of which cite and expand upon Rutkowski's (1986) stages. Phase 1 includes Rutkowski's (1996) first two stages (Pre-singer and Speaking Range Singer) in describing how children use their speaking voice more than their singing voice. Phase I extends Rutkowski's (1996) description by including a child's use of song text and melodic contour,

The words of the song appear to be the initial centre of interest rather than the melody, singing is often described as "chant-like," employing a restricted pitch range and melodic phrases. In infant vocal pitch exploration, descending patterns predominate. (Welch, 1998, p. 35)

Though Welch's (1998) Phase 2 mentions the expansion of pitch range – hence the need for referencing Rutkowski's (1996) study – Welch's later phases are independently developed.

Welch (1998) asserted that by the time a child reached Phase 4, he or she has developed pitch accuracy: "No significant melodic or pitch errors in relation to relatively simple songs from the singer's musical culture" (p. 35).

Although Rutkowski's (1996) model and Welch's (1998; 2000) model have similarities, Rutkowski sought to measure only a specific piece of singing ability, use of singing voice, while Welch (1998; 2000) sought to measure development holistically in measuring use of singing voice as well as control of pitch. Rutkowski (1996) suggested that singing voice and intonation development were two separate constructs. "It seems that the use a child has of his/her singing voice may be a construct separate from and requisite to the ability to sing in-tune" (Rutkowski, 1996, p. 202). Welch (1998), on the other hand, viewed the Phases within the context of a

singing-task meaning that children were at different points of development based on whether children were singing songs or pitch-matching patterns.

If children's overall musical development were to be judged solely on the accuracy with which they could reproduce song melodies, a false picture would be created which belied their developing (and more competent) singing skills when presented with non-song tasks. (Welch, 1998, p. 37)

Both models have been used in subsequent studies (Welch, 2015).

Two models have included the learning sequences involved in singing voice development (Berkowska & Bella, 2009; Phillips et al., 2012). Both models included listening, interpreting, imitating, and evaluating. Additionally, research has investigated the learning sequences involved in singing voice development in terms of why some children may not develop accurate intonation (Lebrun, Moreau, McNally-Gagnon, Mignault-Goulet, & Peretz, 2012).

Phillips, Williams, and Edwin (2012) described a four-step sequential model referred to as the Psychomotor Process:

(1) the teacher provides a model or stimulus; (2) the student perceives and decodes the model; (3) the student imitates the model; and (4) the student analyzes teacher-provided and internal feedback regarding his effort and seeks to make improvements and/or changes in vocal production. (p. 596)

Within the second step, "the student perceives and decodes the model," two additional processes were described: "aural acuity (precise perception of pitch and memory thereof) and audiation (musical pattern discrimination)" (Phillips et al., 2012, p. 596). Berkowska and Bella's (2009) Vocal Sensorimotor Loop documented a nearly identical sequence of learning,

singing from memory of well-known melodies requires the retrieval of pitch and temporal information from long-term memory and fine motor planning/implementation. In addition, the ongoing vocal production is fed back to the system (i.e., perception), compared with the intended melody, thus eventually influencing motor planning (e.g., through error correction) for the subsequent note to be produced. Similar mechanisms are engaged in imitation. The target melody to be imitated is perceived, stored in the short-term memory, and the stored pitches mapped into motor gestures. As before, a feedback mechanism allows the singer to monitor his/her ongoing performance and to correct errors, if needed. (p. 77)

One primary difference between the two models included the origination of sound. The psychomotor process (Phillips et al., 2012) has been cited as beginning with an external source (the teacher) that provides the stimulus, therefore indicating that it was a process for teaching. The vocal sensorimotor loop (Berkowska & Bella, 2009) instead has been cited as beginning internally with the memory, indicating that the process was independent of teaching.

According to Berkowska and Bella (2009), brain damage or damage to one of the mechanisms associated with the sensorimotor vocal loop could interfere with information processing, leading to inaccurate singing in adulthood. An in-depth discussion regarding poor-pitch singing in adulthood is beyond the scope of this study. Discussions of inaccurate singing can be found in articles by Bella and Berkowska (2009a); Bella, Giguere, and Peretz (2007); Pfordresher (2007); Pfordresher and Halpern (2013); and Welch (1979).

In a review of literature, Tsang, Friendly, and Trainor (2012) described the vocal sensorimotor loop with singing development research in infancy through late childhood. Because the vocal sensorimotor loop has been documented to involve perceptual and motor systems that

develop through childhood, “experience, practice, and cognitive skills” (p. 39) may influence how information travels through the model as an adult (Tsang et al., 2012). Although each piece of the model may be supported by existing research, little is known about how all the pieces work together, or how sensorimotor interactions develop. “While the evidence is generally consistent with a complex process involving perception, production, cognition and environmental feedback, the complex interplay between these components during development remains largely unknown” (Tsang et al., 2012, p. 39).

What is suggested however, is that a disorder may prevent singing development within a small population. Lebrun, Moreau, McNally-Gagnon, Mignault-Goulet, and Peretz (2012) defined congenital amusia as “a disorder preventing afflicted individuals from perceiving musical normally” (p. 683), specifically regarding pitch perception. Most commonly detected in adults, amusia has also been diagnosed in childhood. In a case study, Lebrun et al. (2012) diagnosed an otherwise average child with amusia using a child-adapted version of the Montreal Battery for the Evaluation of Amusia (MBEA), the measurement tool most commonly used to assess adult amusia. The child came from what most would label as a musical background through having parents who were musicians, taking private lessons, and participating in a school choral program. The adapted MBEA scores were compared to three other children with similar backgrounds in addition to a larger comparison group. Lebrun et al. (2012) concluded that the child’s “musical difficulties seem to be associated with a deficit in fine-grained pitch discrimination, as observed in the amusic adult population” (p. 686). Unlike adults, her deficit was attributed to “a delay in maturation of the auditory cortex” (p. 688). Additionally, the child had “severe problems in perceiving, memorizing, and producing music” (p. 687) even though time had previously been devoted to strengthening her skills with private instruction. At the time

of publication, the participant in Lebrun et al.'s (2012) study was the only documented diagnosis of childhood amusia.

Empirical Research of Singing Ability

Instruction

Instruction has been shown to improve singing ability (Miyamoto, 2005; Paney & Kay, 2014; Roberts & Davies, 1975). There are, however, multiple variables associated with teaching children to sing. Instruction encompasses (a) the effect of teacher skills or behaviors on singing ability, (b) the effect of instructional procedures on singing ability, (c) the effect of practice on singing ability, and (d) the effect of performance context on singing ability.

Effect of Teacher Skills and Behaviors

Teacher skills and behaviors regarding singing instruction have been described through actions specific to the teacher's voice or teacher verbalizations to elicit responses from students. Research regarding the effect of teacher skills and behaviors can be grouped into three categories: (a) the vocal model, (b) feedback, and (c) use of singing voice. Conflicting results have been found for both the vocal model (Green, 1990; Small & McCachern, 1983; Yarbrough, Green, Benson, & Bowers, 1991; Yarbrough, Bowers, & Benson, 1992) and use of singing voice (Persellin, 2006; Phillips, 2014). Feedback, however, has been shown to have a positive effect on children's singing ability (Porter, 1977; Welch, Howard, & Rush, 1989).

There have been conflicting findings for the effect of vocal model. Several studies have found an overall positive effect for model, but the results have varied regarding which model timbre was most effective (Green, 1990; Yarbrough et al., 1991; Yarbrough et al., 1992). Studies have also found no effect for model for a specific population of singers included in the study such as certain singers (as opposed to uncertain singers) (Yarbrough et al., 1992) or first graders

(Small & McCachern, 1983). Yarbrough, Bowers, and Benson (1992) found that uncertain singers sang more accurately when responding to a female model with no vibrato than when responding to a child model or a female model with vibrato. Green (1990), however, found that first through sixth grade children sang most accurately when responding to a child's voice, less accurately when responding to a female voice, and the least accurate when responding to a male voice. With kindergarten through third grade children, Yarbrough et al. (1991) found a significant difference between models favoring the female voice over the male voice. Another study found no clear differences between first grade control and treatment groups when responding to a male or a female model during pretests and from pretests to posttests after a practice treatment period (Small & McCachern, 1983). Additionally, certain singers (as opposed to uncertain singers) exhibited no significant differences in singing accuracy when responding to different singing models (Yarbrough et al., 1992).

Feedback has been shown to have a positive effect on pitch-matching scores (Porter, 1977; Welch et al., 1989). Welch et al. (1989) provided weekly sessions to the Singing Assessment and Development (SINGAD) computer program for two experimental groups. SINGAD provided students with visual feedback during pitch-matching activities. Although both groups significantly improved pitch-matching scores on the posttest, the group that received teacher feedback in addition to visual feedback scored significantly higher than the group that only received SINGAD feedback. Porter (1977) also found that teacher feedback through successive approximation was an important component in developing pitch-matching skills.

The extent of use of the teacher's singing voice has been shown not to have an effect on singing ability in children. In pedagogical texts, Phillips' (2014) suggested that music specialists should only sing for students, instead of with students. Persellin (2006), however, found that

there were no significant differences between groups for Kindergarten children who received instruction where the teacher (a) sang for the students, (b) sang with the students, and (c) sang a combination of for and with the students.

Effects of Instructional Procedures on Singing Ability

Several instructional procedures have been represented in singing ability research. Literature dissemination (as defined as the process of teaching a song) was shown to have conflicting effects on singing ability (Apfelstadt, 1984; Gault, 2002; Klinger et al., 1998). The presence and type of singing instruction (Apfelstadt, 1984; Miyamoto, 2005; Roberts & Davies, 1975; Rutkowski, 1996) and pattern instruction (Jarjisian, 1983; Porter, 1977) were also shown to have positive effects on singing ability. Use of accompaniment, however, has been shown to have conflicting effects on singing ability (Atterbury & Silcox, 1993; Guibault, 2004).

Literature dissemination has been researched with conflicting findings regarding the benefits of singing songs by rote. Learning a song by rote may be described as teaching a song one phrase at a time, then linking phrases together until the entire song is taught. Another method of literature dissemination, referred to as immersion, has typically included repetitious singing where the song is repeated until children can sing independently. Research has suggested that teaching children songs by rote was significantly more effective than by immersion (Apfelstadt, 1984), as effective as by immersion (Gault, 2002), and significantly less effective than by immersion (Klinger et al., 1998).

The presence and types of singing instruction have been shown to improve singing ability (Apfelstadt, 1984; Miyamoto, 2005; Roberts & Davies, 1975; Rutkowski, 1996). The Yuba Method, which utilized technical exercises to help children transition from head voice to singing voice, was shown in one study to increase singing accuracy with elementary students

(Miyamoto, 2005). Rutkowski (1996) found that children who received instruction that included opportunities for small-group and independent singing as part of their music class earned higher singing development scores than children who received music class instruction without opportunities for small-group and independent singing. Roberts and Davies (1975) found that remedial instruction significantly improved pitch- and interval-matching scores. Apfelstadt (1984) found a significant effect for instruction where children who received singing instruction with and without melodic perception reinforcement (associated movements) received significantly higher mean singing accuracy scores than children who received traditional music instruction without singing instruction.

Pattern instruction has been shown to have a positive effect on singing accuracy (Jarjisian, 1983; Porter, 1977). Students provided with opportunities to develop discrimination skills through multiple discrimination training were shown to have increased pitch-matching scores when combined with successive approximation (Porter, 1977). Jarjisian (1983) found that children who received a combination of diatonic and pentatonic pattern instruction received higher singing accuracy mean scores than children who received either diatonic pattern instruction or pentatonic pattern instruction.

Accompaniment has been shown to have conflicting effects on singing ability. Guibault (2004) found that children who received instruction with root melody accompaniment earned similar singing achievement scores as those who received instruction without root melody accompaniment. In contrast, neither Atterbury and Silcox (1993) nor Hedden and Baker (2010) found singing accuracy differences between groups receiving music instruction with or without piano accompaniment.

Effect of Practice on Singing Ability

The effect of repetitive practice on singing ability has had conflicting results in the literature. Paney and Kay (2014) found that children's singing accuracy significantly improved with the *SingingCoach* computer game, especially when children were given between one and five practice opportunities in addition to the pretest and posttest. Welch et al. (1989) found similar results. Reifinger (2009) also found that weekly practice significantly improved the singing accuracy of tonal pattern performance. Conversely, Small and McCachern (1983) found that practice-matching practice did not improve pitch-matching skills.

Effect of Performance Context on Singing Ability

Performance context has been shown to have conflicting effects on pitch accuracy scores. Goetze and Horii (1989) found that Kinder through third graders sang more accurately when singing individually versus singing in a group. Yarbrough et al. (1991) measured pitch-matching accuracy with Kinder through third grade children regarding response mode (handsigns, pitch syllables, neutral syllables). There were no differences found between response modes. Moore et al. (2004) also measured pitch- and interval-matching accuracy with third and fifth grade children from four different countries when using response modes of echoing patterns vocally or by playing them on a xylophone. Children tended to be more successful when echoing patterns on the xylophone than when singing.

Age and Grade

Studies have yielded conflicting results regarding singing ability across grade levels. Several studies have found that overall singing ability improved across grade levels (Fyk, et al. 1995/1996; Goetze and Horii, 1989; Hornbach & Taggart, 2005; Miyamoto, 2005). When viewed through the lens of instruction, however, several studies did not find differences in

singing accuracy scores across grade levels regarding vocal model (Yarbrough et al., 1991; Yarbrough et al., 1992), and response mode (Yarbrough et al., 1991).

The age of children has had conflicting effects on text and song performance success. In one study, word performance significantly improved between ages five and six as well as six to seven (Welch et al., 1998). In the same study, children performed the pitches significantly better at age seven than at ages five and six. Another study, however, found no difference in song performance between five and six year-olds (Gault, 2002).

Gender

Studies regarding gender differences in singing ability have yielded conflicting results. Several results have indicated a tendency for girls to have higher assessment mean scores than boys in Kindergarten through third grade (Goetze & Horii, 1989), in third and fifth grade (Moore et al., 2004), and with ages six and nine (Moore et al., 1995/1996). In contrast, other studies have found no difference in singing ability between boys and girls (Apfelstadt, 1984; Hedden & Baker, 2010; Moore, 1994; Paney & Kay, 2014; Yarbrough et al., 1992). Overall, studies that have focused only on younger children have yielded no differences between genders, whereas studies including older children have shown differences, favoring girls. However, one study did find higher assessment scores for boys than for girls (Paney & Kay, 2014). Paney and Kay (2014) assessed singing ability through a pitch-matching computer game and found that boys' scores were higher than girls overall.

Population

Across populations, pitch and interval matching accuracy have yielded conflicting results. Moore et al. (1995/1996) found significant differences across populations for interval-matching accuracy. Spanish and American children tended to have higher scores than Polish children.

Moore et al. (2004), however, did not find a significant difference across populations regarding response modes (singing or xylophone) in children from England, America, Taiwan, and Spain.

Measuring and Scoring Singing Ability

Although many singing treatments have been investigated, the measurement of singing ability (specifically regarding pretest-posttest research designs) has used a shorter list of instruments. In a critical narrative review of singing ability measures represented in music education research, Salvador (2010) grouped instruments into four categories: (a) acoustic, (b) researcher-as-acoustic, (c) researcher-designed, and (d) previously published. Acoustic measures were described as assessments that require computer technology to assess pitch accuracy, resulting in little or no researcher influence on outcome scores. Salvador (2010) described researcher-as-acoustic measures as those relying on individuals making perceptual determinations of singing ability, therefore allowing the researcher or grader to influence outcome scores. In other words, the researcher, or a panel of judges, would provide scores based on what he, she, or they would hear. Researcher-designed measures and previously published measures differed in that researcher-designed measures were created for a specific study and not used thereafter. On the other hand, previously published measures had been used and cited across multiple studies. Researcher-designed measures and previously published measures also had several similarities. Researcher-designed and previously published measures were described as instruments that underwent revisions and tests of reliability. Additionally, researcher-designed and previously published measures were described, for the most part, as Likert-type scales that assessed elements of singing achievement including melodic contour and use of singing voice (Salvador, 2010).

Acoustic measures have been used when measuring pitch-matching accuracy or vocal accuracy. Several studies have reported measuring accuracy by having children sing responses into a midi-microphone connected to various computer programs (Miyamoto, 2005; Paney & Kay, 2014; Welch et al., 1989). Software programs have included the following: *SingingCoach* (Paney & Kay, 2014), Sona-Speech Model 3600 (Miyamoto, 2005), and SINGAD (Welch et al., 1989). The differences in software used across studies can make comparisons problematic in that no two programs may measure accuracy the same.

Several studies have utilized researcher-as-acoustic measures to assess pitch and/or interval accuracy (Fyk, 1985; Moore et al., 2004; Moore et al., 1995/1996; Roberts & Davies, 1975; Small & McCachern, 1983). Two studies documented participant accuracy scores by combining the frequency of correct pitches with correct intervals. By listening to participant responses, pitches and intervals were perceived as correct or incorrect (Moore et al., 2004; Moore et al., 1995/1996). Roberts and Davies (1975) constructed a test battery that included several components measuring accuracy based on individual pitch-matching, interval pitch-matching, echoing a melody, and singing a familiar song. Children were rated as accurate or inaccurate by a panel of judges.

Several studies have included researcher-designed assessments to measure singing achievement (Gault, 2002; Guilbault, 2004; Jarjisian, 1983). Researcher-designed measures have included Likert-type rating scales that have measured specific characteristics of the singing performance and have been given to multiple evaluators for the purpose of checking reliability. Gault (2002) developed a Likert-type scale that evaluated tonal and rhythmic performance elements. Jarjisian's (1983) 5-point Likert-type scale evaluated use of singing voice as well as pitch accuracy.

Previously published assessments have been used to measure singing ability or as a pre-test assessment to determine group homogeneity. Several studies have included previously published assessments, including the Singing Voice Development Measure (Atterbury & Silcox, 1993; Guerrini, 2006; Rutkowski, 1990; Rutkowski, 1996), the Vocal Accuracy Assessment Instrument (Persellin, 2006), the Vocal Pitch-matching Measure (Porter, 1977), the Tonal Rating Scale (Guilbault, 2004), and the Boardman vocal accuracy test (Apfelstadt, 1984). Tonal aptitude has been used as a pre-test measure to demonstrate homogeneity between groups. Primary Measures of Music Audiation was most often used (Apfelstadt, 1984; Atterbury & Silcox, 1993; Jarjisian, 1983; Persellin, 2006; Rutkowski, 1996). Guilbault (2004), however, used the Intermediate Measures of Music Audiation.

Assessment song material has varied across studies. Several researchers have chosen folk literature that was appropriate for a child's range at a given age (Atterbury & Silcox, 1993; Jarjisian, 1983; Rutkowski, 1996). Other researchers have chosen to create songs based on developmentally appropriate parameters (Apfelstadt, 1984; Goetze & Horii, 1989; Guilbault, 2004; Small & McCachern, 1983). Still other researchers limited the material to tonal patterns, intervals, and sing pitches (Persellin, 2006; Welch et al., 1989). Various songs and pitches have been used to measure singing voice development and achievement. Some studies have used folk songs including *All Around the Buttercup*, *Let Us Chase the Squirrel* (Gault, 2002; Klinger et al., 1998), *America, Path to the Moon* (Guerrini, 2006), and *Pinto Pony* (Atterbury & Silcox, 1993). Other studies have required children to sing patterns including the descending G5 – E5 pattern (Yarbrough et al., 1991), as well as different arrangements of C4, D4, E4, G4, and A4 (Moore et al., 2004).

Singing Task

The measurement instrument is the tool used to quantify singing ability. Singing task, however, may be described as what the children sing to demonstrate their ability. Examples of singing tasks used in studies have included (a) literature content (text-related and melodic attributes), (b) song singing, (c) tonal pattern descriptions, and (d) pitch interval descriptions.

Ease of intervals and content familiarity has been found to have positive effects on singing ability. Moore (1994) found that singers earned higher pitch-matching scores when matching the higher of two pitches as well as echoing short melodies than they did when matching the lower of two pitches. Additionally, singers earned lower scores when matching the middle pitch of a chord than all other tasks. Reifinger (2009; 2012) found that singing accuracy scores were significantly higher when singing familiar patterns than when singing unfamiliar patterns.

In one study with 10 year-olds, gender, pitch frequency, pitch volume, and pitch duration were found to influence pitch-matching ability (Fyk, 1985). When given three pitches (C4, A4, C5), boys tended to be more successful matching C4 (middle C) than girls, while girls tended to be more successful matching A4 and C5 than boys. Most children tended to have greater success reproducing pitches given at mezzo-forte (*mf*) than at piano (*p*). Children were categorized based on success frequency and type of non-success into three groups: Type I (Consonance Conformity), Type II (Active Attempt), and Type III (Immediate Success). Children categorized as Type I tended to exhibit the lowest success in pitch matching while children in Type III tended to exhibit the highest success. The mean duration threshold for children categorized into Type I was 925 milliseconds, Type II was 210 milliseconds, and Type III was 13 milliseconds.

Research results have varied regarding the effect of text on song acquisition in childhood. In a longitudinal study involving five, six, and seven year-olds, Welch et al. (1998) measured performance accuracy using two researcher-composed songs with words. They found that children at every age produced the words more accurately than the pitches. In another study involving Kinder and first graders, Gault (2002) found that students performed *Let Us Chase the Squirrel* more accurately with text than with no text, although no differences were found for the other chosen piece, *All Around the Buttercup*. Melodic construction of songs has been found to affect accuracy of singing a song with and without text. Guerrini (2006) used the Singing Voice Development Measure (Rutkowski, 1990) to measure accuracy differences when singing patterns, a well-known song (*America*), and a less well-known song (*Path to the Moon*). Results indicated higher accuracy when singing the patterns than when singing either of the songs with text.

Music-specific Meta-analyses

Before meta-analyses, researchers applied familiar statistical methods to synthesize information beyond a narrative review of literature. In the 1960s through the early 1970s, researchers synthesized and cross-tabulated studies by independent variables and “significant vs. non-significant findings” (Glass, 1976, p. 6), a process Glass (1976) labeled as the “voting method” (p. 6). Glass argued that the voting method was problematic due to sample size bias and generalization of results based solely on significance testing. While research synthesis can be recognized as an essential component of gathering and interpreting results across more than one study, it fails to account for key statistical differences among studies. Furthermore, while a literature review has been cited as an effective method for synthesizing results and implications, it may also include subjective bias from the researcher (Mishra, 2014).

In an effort to address the subjectivity within a literature review and the statistical incongruence within antiquated forms of research synthesis, the meta-analysis procedure was designed. Meta-analyses have been documented as providing an objective and quantitative means to compare research studies (Standley, 1996). Standley (1996) succinctly articulated the definition of the term, meta-analysis,

A Meta-analysis is a statistical procedure that has become an accepted method for synthesizing such an extant body of literature, describing its characteristics and providing a basis for informed generalization about the results. (p. 106)

Large-scale syntheses have integrated the growing pools of published studies in music research. *The Handbook of Research on Music Teaching and Learning* (1992), *The New Handbook of Research on Music Teaching and Learning* (2002), and the three volumes of *MENC Handbook of Research on Music Learning* (2011) represent publications that have included large-scale literature reviews of current research related to and within music education. Content analyses of music education research have also been published on topics including string education (Kantorski, 1995), student teaching (Silveira & Diaz, 2014), article subject (Kratus, 1992), article citations (Brittin & Standley, 1997; Killian, Liu, & Reid, 2013; Schmidt & Zdzinski, 1993), and published research characteristics (Brittin & Standley, 1997; Ebie, 2002; Yarbrough, 1984; Yarbrough, 2002). Multiple meta-analyses have also been found in published literature investigating topics within music cognition, music therapy, and music education.

Since there are a large number of music-specific meta-analyses this may indicate a general acceptance of meta-analytic procedures among music and music-related researchers. Of the meta-analyses found in music, three areas of research were prevalent, including (a) cognition,

(b) music therapy, and (c) music education. As a result, each area of meta-analytic music studies is examined in this section.

Cognition

The effects of music on cognition have been investigated through several meta-analyses, most finding an overall small to moderate effect size. Mozart stimuli were shown to have conflicting effects on measures of cognition including a large effect on cognitive test outcomes in one analysis (Hetland, 2000) and small to moderate effects on cognitive test outcomes in other analyses (Chabris, 1999; Hetland, 2000). Mozart stimuli were shown to have conflicting effects on spatial and spatial-temporal tasks, yielding both moderate effects (Hetland, 2000) and small effects (Pietschnig, Voracek, & Formann, 2010). Background music was shown to have no effect on behavior (Kampfe, Sedlmeier, & Renkewitz, 2010) and a small to moderate effect on arousal, shopping duration, and patronage (Garlin & Owen, 2006).

The effects of Mozart stimuli on cognitive test performance have been inconsistent in the literature. Chabris (1999) and Hetland (2000) investigated (a) the effects of the first movement in Mozart's Sonata K. 444 versus silence, and (b) the effects of the first movement in Mozart's Sonata K. 444 (and a Schubert piece) versus relaxation. To clarify, Hetland (2000) included a replication and expansion of Chabris' (1999) study. Chabris (1999) categorized variables across 16 published studies, and Hetland (2000) categorized variables across 17 published and unpublished studies researching the effect of Mozart's music on cognitive tests, as measured with components of the Stanford-Binet, Raven's Advanced Progressive Matrices, and other cognition tests. Chabris (1999) found little overall difference between groups ($d = 0.09$) regarding silence versus Mozart stimuli, favoring the group with Mozart stimulus. Hetland (2000), however, found a moderate effect ($d = 0.48$) and stated, "this result is based on a larger

sample of studies and subjects and thus is more likely to represent the true effect” (p. 57). For the second analysis, Chabris (1999), across the same 16 studies, measured cognitive test outcomes with auditory relaxation instructions using Mozart’s music and other stimuli also yielding little overall differences between groups ($d = 0.20$). Hetland (2000), however, found a larger effect ($d = 0.70$).

Several researchers have conducted meta-analyses investigating the effects of music on spatial processes, including research that examined the Mozart Effect (Chabris, 1999; Hetland, 2000a; Pietschnig et al., 2010). Overall, findings of meta-analyses of spatial processes and effects of music have been consistent. Hetland (2000) completed meta-analyses of the effects of Mozart stimuli on cognition test outcomes but also conducted two additional meta-analyses that yielded moderate effects of Mozart stimuli on spatial and spatial-temporal processes after conducting the two on the effects of Mozart stimuli on cognition test outcomes. The first meta-analysis investigated the effects of music listening on various spatial tasks across 36 experiments. Results indicated a moderate effect ($d = 0.46$). The second meta-analysis investigated the effects of music listening on spatial-temporal tasks across 13 experiments and also yielded a moderate effect ($d = 0.50$).

Pietschnig, Voracek, and Formann (2010) conducted a meta-analysis yielding small to moderate effects of Mozart stimuli on spatial ability from 40 published and unpublished studies beginning with the original research of Rauscher, Shaw, and Ky (1993). All studies were categorized by exposure to Mozart’s sonata KV 448: Mozart versus non-musical stimulus/silence, Mozart versus other music, and other music stimuli versus non-music stimuli/silence. Results yielded an overall small effect ($d = 0.37$) for Mozart versus non-musical

stimulus/silence, a small effect ($d = 0.15$) for Mozart versus other music, and a moderate effect ($d = 0.38$) for other music stimuli versus non-music stimuli/silence.

Several meta-analyses have investigated the effects of background music on multiple variables, yielding similar results of small to moderate effects (Garlin & Owen, 2006; Kampfe et al., 2010). Kampfe, Sedlmeier, and Renkewitz, (2010) investigated the effects of background music on adult listeners across two meta-analyses. Studies ($N = 66$) were categorized based on the positive and negative impact music had on behavior, specifically mundane behavior, cognition judgment, cognition achievement, and emotion. Background music was found to have little effect overall ($r = 0.03$). Garlin and Owen (2006) investigated the effect of background music in commercial settings across 32 studies with variables including affect, financial returns, attitude/perception, temporal effects, and behavior. Variables were sorted into three categories that yielded small to moderate effects: value ($d = 0.14$), affect ($d = 0.38$), and duration ($d = -0.32$). Specific findings included, “tempo has the greatest effect on arousal; slower tempo, lower volume and familiar music results in subjects staying marginally longer at a venue,” and “the mere presence of music has a positive effect on patronage” (Garlin & Owen, 2006, p. 761).

Two other meta-analyses investigated music and visual influences yielding moderate effects (Platz and Kopiez, 2012), or affective influences yielding small effects (Timmerman et al., 2008). Platz and Kopiez (2012) conducted a meta-analysis investigating visual influence on the evaluation of music performance. Specifically, the purpose of the meta-analysis was to “quantify the effect of presentation mode of a music performance on the audience’s evaluation” (p. 72). Fifteen studies yielded a moderate overall weighted effect size ($d = 0.51$). Timmerman et al. (2008) meta-analyzed 35 studies to investigate the affective response of popular music

listening. Overall results indicated that “the exposure to music increases anti-social actions and beliefs, *average r = .210*” (p. 315).

Music Therapy

The American Music Therapy Association described the goals of music therapy with seven categories including physical rehabilitation, wellness, stress management, pain management, emotional expression, memory enhancement, and communication (American Music Therapy Association, 1990-2014). The primary goals in clinical settings have been represented in meta-analysis research within the categories of physical rehabilitation, wellness, stress management, pain management, and emotional expression (Phillips, 2008). Educational goals, on the other hand, have been included in meta-analyses in terms of social and academic outcomes (Standley, 1996).

The effects of music on physical rehabilitation have been documented in numerous meta-analyses with varied results. Populations represented in music therapy physical rehabilitation studies have included premature infants (Standley, 2002), cancer patients (Zhang et al., 2012), diagnosis-non-specific patients (Loomba, Shah, Chandrasekar, Arora, & Molnar, 2012; Pelletier, 2004), older adults (Clark, Taylor, & Baker, 2012), epilepsy patients (Dastgheib, et al., 2013), mental disorder patients (Gold, Solli, Kruger, & Lie, 2009; Gold, Voracek, & Wigram; 2004; Silverman, 2003) and Parkinson’s Disease patients (de Dreu, van der Wilk, Poppe, Kwakel, & van Wegen, 2012). Meta-analysis results related to physical rehabilitation can be categorized into the following categories: physiological, physical, and psychological (Wigram, Pedersen, & Bonde, 2002). One meta-analysis indicated positive effects for all physical rehabilitation patients in terms of behavior reinforcement (Standley, 1996).

Physiological topics of rehabilitation in music therapy have measured the effects of music on a patient's "heart rate, blood pressure, respiration, skin temperature, electrodermal activity, and brainwaves" (Wigram et al., 2002, p. 137). Overall effects of music on physiological topics have included a large effect on oxygen saturation (Standley, 2002), large effects on heart rate (Clark et al., 2012; Loomba et al., 2012; Pelletier, 2004; Zhang et al., 2012), conflicting effects on respiratory rate (Zhang et al., 2012), small to medium effects on blood pressure (Loomba et al., 2012; Zhang et al., 2012), and large effects on interictal epileptic discharge (Dastgheib et al., 2013).

Music was shown to decrease heart rate in cancer patients ($z = 2.32$) (Zhang et al., 2012) and non-cancer patients (standard mean difference of -2.63) (Loomba et al., 2012) but not older adults during exercise routines ($z = 0.80$) (Clark et al., 2012). In a meta-analysis, Zhang et al. (2012) found that recorded music had a large effect across studies on the decrease of respiratory rates in cancer patients ($z = 5.16$). In the same meta-analysis, recorded music had a nominal effect on systolic blood pressure ($z = 0.44$) and diastolic blood pressure ($z = 0.98$) across studies (Zhang et al., 2012). In another meta-analysis, however, music therapy was shown to have a small to moderate effect on decreasing systolic and diastolic blood pressure (standard mean difference of -2.63) across studies (Loomba et al., 2012). Finally, music therapy and exposure to Mozart's music (specifically K.488 sonata) decreased interictal epileptic discharge in 84% of epilepsy patients with a 95% CI [77.0, 89.2] (Dastgheib et al., 2013).

Music and music therapy have had conflicting results regarding the effects of music on NICU patients (Standley, 2002), older adults and exercise (Clark et al., 2012), and patients with Parkinson's disease (de Dreu et al., 2012). Music was shown to be effective in developing sucking and feeding abilities in NICU premature infants ($d = 0.73$) (Standley, 2002). For older

adults, listening to music was not shown to have short-term effects regarding walking distance ($z = 0.40$) and time ($z = 1.10$) during exercise; however, music positively affected endurance over time ($g = 1.58$) (Clark et al., 2012). Finally, music-based movement therapy had a positive effect on gait-related interventions of Parkinson's disease patients ($z = 2.31$) (de Dreu et al., 2012).

The effects of music and music therapy on the psychological wellness of patients with mental disorders have been researched through several meta-analyses, all with an overall positive effect size. Music listening and music therapy interventions were shown to have a positive effect on the symptoms of patients with schizophrenia ($d = 0.71$) (Silverman, 2003), and psychotic and non-psychotic mental disorders ($R^2 = 0.78$) (Gold et al., 2009). In another meta-analysis, music therapy was shown to have a positive effect on children and adolescents with psychopathological disorders ($d = 0.61$) (Gold et al., 2004).

A goal of music therapy has been to help clients and patients develop an overall sense of wellness and improved quality of life using music (American Music Therapy Association, 1990-2014). Wellness topics measured in meta-analyses have included relaxation promotion (de Niet, Tiemens, Lendemeijer, & Hutschemaekers, 2009), healthy sleep habits (de Niet et al., 2009; Wang, Sun, & Zang, 2013), restfulness (Zhang et al., 2012), and anti-depression (Gold et al., 2009; Ueda, Suzukamo, Sato, & Izumi, 2013; Vasjonyte & Madison, 2013; Zhang et al., 2012). The effectiveness of music on behavior reinforcement has also been conducted through meta-analyses for wellness patients (Standley, 1996).

Regarding wellness variables, relevant meta-analyses have indicated that, for the most part, music has tended to have a positive effect on patient outcomes reflected by effect size. Music-assisted relaxation with recordings “had a moderate effect on the sleep quality of patients with sleep complaints” and “can be used... to improve sleep quality” (standard mean difference

of -0.74) (de Niet et al., 2009, p. 1356). Another sleep study also found positive effects for music with patients with acute and chronic sleep disorders (standard mean difference of -0.63) (Wang et al., 2013), although it was shown to have negative effects on fatigue in cancer patients ($z = 0.39$) (Zhang et al., 2012). Music was shown to have a positive effect on the quality of life as well as alleviating symptoms of depression for patients with cancer ($z = 4.65$) (Zhang et al., 2012) and mental disorders (Gold et al., 2009). In alleviating symptoms of depression for patients with dementia, meta-analyses reported both large effects ($ES = 1.04$) (Vasionyte & Madison, 2013) and small effects (standard mean difference of -0.49) (Ueda et al., 2013).

Another goal of music therapy has been to aid in stress management and anxiety (American Music Therapy Association, 1990-2014). Stress management topics in music therapy-specific meta-analyses have included treatments within patient care and non-patient care settings. Populations represented in patient care settings have included cancer patients (Zhang et al., 2012), dementia patients (Ueda et al., 2013), pre-mature infants (Standley, 2002), and colonoscopy patients (Tam, Wong, & Twinn, 2008).

Relevant meta-analyses have indicated that music stimuli and music therapy techniques had overall positive effects on stress management and alleviating anxiety in both patient care settings and general populations. In patient care settings, music was shown to reduce anxiety overall in cancer patients before, during, and after cancer-related treatments ($z = 2.30-3.77$) (Zhang et al., 2012), and in dementia patients (standard mean difference of -0.64) (Ueda et al., 2013). Music was also shown to assist with patient stress management before and during colonoscopies by reducing the duration time of the procedure (standard mean difference of -2.84) as well as marginally decreasing sedation (standard mean difference of -0.46) (Tam et al., 2008). Regarding non-patient care populations, music therapy was shown to have a positive effect in

settings that induced short-term situational stress ($d = 0.67$) (Pelletier, 2004). Pelletier (2004) found that “music combined with relaxation techniques does have a significant impact on decreasing arousal due to stress,” (p. 202) especially with third-trimester adolescents ($d = 2.02$), experienced musicians ($d = 1.49$), and with individualized (versus group) treatment ($d = 0.79$).

Music therapy-specific meta-analyses have yielded conflicting results regarding pain management effectiveness. Zhang et al. (2012) documented an overall large effect of music treatment on pain management in cancer patients ($z = 3.10$). Bechtold et al. (2008), however, documented a small effect on pain management for colonoscopy patients across studies ($z = 1.71$).

Meta-analyses have documented positive effects for the effect of music on emotional expression. Recorded lullabies were shown to have a large effect on NICU infants regarding behavioral state ($d = 0.73$ – 1.95) (Standley, 2002). Patients who listened to music before a colonoscopy reported an overall improved treatment experience ($z = 3.44$) (Bechtold et al., 2008). Music did not, however, influence a colonoscopy patient’s “willingness to repeat the same procedure in the future” across studies ($R = 3.89$) (Bechtold et al., 2008, p. 19).

Music therapy has also yielded primarily positive effects when used in educational settings for both social (Standley, 1996) and academic objectives (Butzlaff, 2000; Standley, 1996; 2008). Standley (1996) meta-analyzed 98 published and unpublished studies on the effectiveness of contingent music in reinforcing behavior in educational settings. Overall, contingent music was effective for children regarding social ($ES = 2.04$) and academic ($ES = 2.18$) objectives, with the exception of one independent variable. Music was measured as less effective than candy in one study ($ES = -3.12$). Additionally, music as a contingency was found

to have been more effective in increasing behavior ($ES = 2.97$) than decreasing behavior ($ES = 2.77$).

Butzlaff (2000) and Standley (2008) found overall positive effects of music on literacy. In two separate meta-analyses, Butzlaff (2000) investigated associations between literacy and music instruction. The meta-analysis examined 25 studies of relational research that included correlations between student reading scores and formal music instruction. Findings included a strong effect ($r_{range} = -.0.19 - 0.65$; $r_{mean} = 0.17$) between participation and standardized reading/verbal test scores yielding a positive effect in approximately 80% of the analyzed studies. Students who participated in music classes also tended to have higher reading scores than students who did not participate in music classes.

Standley (2008) gathered research regarding effects of music interventions on visual decoding ability. Across the 30 studies analyzed, music interventions had a moderate positive overall mean effect on reading skills ($d = 0.32$). The strongest effects occurred for students with special needs ($d = 0.81$), at-risk students ($d = 0.66$), and Pre-Kindergarten children ($d = 0.62$). The strongest intervention effects included the use of contingent music as a reward ($d = 0.66$), using music to cue student attention toward a reading task ($d = 0.54$), and musically delivering reading skills or information ($d = 0.44$).

Music Education

As meta-analyses in the field of music education has developed, it has been applied in various contexts with its first appearance lacking direct applications to music education specifically. Standley (1996) was one of the first to publish a meta-analysis research study in a music education journal. Although published in *Journal of Research in Music Education*, the meta-analysis included implications for general education teachers and music therapists, not for music educators (see the

music therapy section for meta-analysis details). Since then, however, several more meta-analyses have been published with implications for music educators. Hetland's (2000) dissertation utilized meta-analytic procedures to answer several research questions, one of which was music education specific. Acklin's (2010) dissertation meta-analyzed research on the topic of choral and instrumental conducting. Several other meta-analyses have since been published in music education research journals (Standley, 2008; Mishra 2013, 2014). Other meta-analyses may exist, but were inaccessible to the primary researcher, or, as in the case of the Henke (2010), used the term meta-analysis to describe a synthesis of a large body of literature without using statistical analysis.

Across three meta-analyses, Hetland (2000) analyzed 19 published ($n = 10$) and unpublished ($n = 9$) quasi-experimental studies to address the research question: "Does active instruction in music enhance preschool and elementary students' performance on spatial tasks" (p. 104). Theoretical underpinnings included Shaw's and Parsons's neural connections theories as well as Gardner's near transfer theories. Studies were coded according to data factors, sample characteristics, experimental design factors, sample characteristics, and program factors. All analyzed studies took place within elementary music environments.

The first meta-analysis included 15 studies to determine the overall effect of music instruction on spatial task performance. There was a large effect across studies measuring spatial task performance pre- to posttest treatment ($r = 0.37$). Hetland (2000) stated that, "active instruction in music does appear to enhance spatial-temporal performance for preschool and elementary aged children, at least while instruction is occurring and at least up through two years of instruction" (p. 157). Seventeen contrast analyses of moderator variables determined the effect size was due to music instruction.

The contrast analyses were also used to investigate specific questions of the 15 studies from the first meta-analysis. Contrast analysis variables included (a) age, (b) student socio-economic status, (c) length of program, (d) lesson format, (e) parental involvement, (f) keyboard instruction, (g) piano versus xylophone instruction, (h) standard music notation, (i) expressive movement, (j) composition or improvisation inclusion, (k) Object Assembly Subtest, (l) article author, (m) Hawthorne Effect, (n) experimenter bias, (o) pre-existing differences, (p) publication bias, and (q) study quality. Meaningful effects were found for two variables: lesson format favoring students in private (versus group) music instruction environments, and notation favoring programs that included standard music notation versus no notational instruction.

The second meta-analysis combined, “five independent effect sizes from three studies (694 subjects) that used Raven’s Standard Progressive Matrices as outcome measures” (Hetland, 2000, p. 114). While the purpose of the first meta-analysis was to measure the overall effect of music instruction on spatial task performance, the purpose of the second meta-analysis was to measure the effect of music instruction on non-spatial/logical tasks based on a single measure. Analyses yielded no effect for Raven’s Standard Progressive Matrices indicating that music instruction may not “enhance measures that require mainly logical ability” (Hetland, 2000, p. 153).

The third meta-analysis included eight studies to test “whether music instruction might enhance spatial abilities beyond the spatial-temporal dimension” (Hetland, 2000, p. 115). Spatial memory measures included the Stanford-Binet Bead Memory task, the Children’s Embedded Figures Test, the Woodcock-Johnson Visual Closure test, and subtests of the WPSSI-R. Results indicated that music instruction might enhance other “spatial memory, spatial recognition, mental rotation, and/or spatial visualization tasks” (Hetland, 2000, p. 156).

The use of meta-analysis for drawing overarching statistical conclusions regarding the effect of music instruction on spatial reasoning allowed Hetland (2000) to combine the results of studies. The results showed that while music instruction exhibited an overall effect on improving spatial reasoning, it did not present the same effect for other areas of reasoning. Non-spatial reasoning, such as logic, did not exhibit improved test results due to music instruction. Furthermore, music instruction may have had an effect on other facets of spatial ability, which suggested that while spatial reasoning benefited from music instruction, other areas such as spatial memory may or may not yield similar benefits.

Acklin (2010) meta-analyzed the effects of conducting on ensemble performance. The “purpose of this study is to methodically investigate one of the research lines in-depth using the meta-analytic techniques of a best-evidence synthesis” (p. 77). The researcher labeled the study as a best-evidence synthesis. Best-evidence syntheses discuss the effectiveness of a body of research in terms of the effect sizes reported by individual studies. If effect sizes were not reported, the researcher computed the effect size if the data is available. Neither overall effect sizes nor moderator variable effect sizes were computed.

Where data was available, Acklin (2010) computed study-specific effect sizes to determine Cohen’s *d*. Twenty-three experimental and descriptive studies were categorized as: “1) Expressive versus Nonexpressive, 2) Conductor-gesture Instruction, 3) Specific Conducting Gestures, 4), Nonverbal and Verbal method, and 5) Synchronization” (Acklin, 2010, p. 82). For expressive versus non-expressive conducting ($n = 7$), two studies provided adequate information to compute effect sizes that favored expressive conducting ($d = 0.24, 0.24$). For conducting-gesture instruction ($n = 3$), large effects were found for young ensembles ($d = 1.73, 0.90, 1.45, 1.61$). For specific conducting gestures ($n = 4$), three studies yielded effect sizes indicating gestures may have an effect on

increased vocal tension, specifically gestures that included the fist ($d = 1.32$), stabbing ($d = 1.10$), palm facing up ($d = 0.99$), palm facing down ($d = 0.04$) and palms facing sideways ($d = 0.22$). Preparatory gestures ($d = 0.89$) also had an effect on performance. For nonverbal and verbal methods ($n = 5$), effects were found across two studies favoring verbal instruction ($d = 0.37$) and nonverbal instruction ($d = 0.99$).

Predictive variables of ability and skill development have been investigated in sight-reading research and yielded contradictory results. Across two published articles, Mishra (2013, 2014) meta-analyzed studies to provide a holistic picture of current sight-reading research and provide supported implications for future teaching. The purpose of the first article was three-fold: (a) to investigate various sight-reading treatments represented in music education research, (b) to investigate trends in the influence of study design on treatment outcome, and (c) to investigate trends in pre- to posttest scores within comparison groups (Mishra, 2013).

The meta-analysis included 92 studies. The population of studies included published and unpublished dissertations and theses, published journal articles, and published paper presentations (via published conference proceedings). Analyses were limited to experimental studies that included pretest-posttest control group designs and equivalent-group posttest-only designs.

One hundred and twenty-four individual analyses were categorized by moderator variable. Research design-related moderator variables included publication type, age, experience level, sight-reading test, performance mode (instrumental/vocal), control group task, if a same or different test was used pre- to posttest, and treatment (Mishra, 2013). Four judges sorted the 124 treatment analyses into the following sub-categories: aural training ($n = 15$), collaboration ($n = 14$), controlled reading ($n = 11$), creative activities ($n = 9$), instrumental training ($n = 3$), interval drill ($n = 3$), movement ($n = 2$), notation ($n = 5$), rhythmic drill ($n = 15$), and singing/solfege ($n = 26$) (Mishra,

2013). All other studies that were not agreed upon by the judges were sorted into non-categorized ($n = 21$).

Although there was a significant effect for treatment ($d = -0.22$), treatment groups were heterogeneous. The extraction of 18 outliers resulted in homogeneous categories with a significant effect ($d = -0.18$) for treatment on sight-reading. Mishra (2013) noted that the negative directionality of the effect size indicated that “the results favor the treatment over the control group” (p. 20).

A Q statistic moderator analysis was used to measure differences between moderator variables and levels within variables. Overall, treatment was the only statistically significant moderator variable ($p = 0.03$). Significant effect sizes within treatment levels included non-categorized ($d = -0.34$), aural training ($d = -0.28$), controlled reading ($d = -0.26$), creative activities ($d = -0.24$), and singing/solfege ($d = -0.20$).

Sight-reading treatments were shown to have a meaningful effect on sight-reading scores. Seventy-eight pre- to posttest homogeneous research design studies were isolated and analyzed to investigate trends in pre- to posttest scores within comparison groups. The analysis yielded a moderately strong effect size ($d = -0.48$) for improvement pre- to post-test scores. Mishra (2013) noted that the negative directionality of the effect size indicated that “the groups improved pre- to posttest” (p. 20).

Treatments yielded positive and negative sight-reading effects. Mishra (2013) found that several specific treatments improved sight-reading: “Training aural skills, eye movements (Controlled Reading), composition and improvisation (Creative Activities), and the use of singing and solfege” (p. 16). Contrarily, “collaborative activities, training in movement or on an instrument did not improve sight-reading, nor did interval or rhythmic drill” (Mishra, 2013, p. 16).

Mishra's (2013) results may support the theory of developing the musician's ear and building a musical vocabulary before applying that vocabulary to written notation, otherwise referred to as sound before sight. Meaningful treatments involved skills associated with music construction, predictive abilities, and pulling from expectations built from previous experiences as opposed to merely the "simple visual decoding of notational symbols" (Mishra, 2013, p. 17). In spite of the results, Mishra (2013) noted more research may still be required to identify specific strategies for developing sight-reading.

Mishra published a second study on sight-reading. The purpose of the second article was "to determine the extent of the overall relationship between previously tested variables and sight-reading" (Mishra, 2014, p. 455). Specifically, the article investigated the relationship between variable constructs and sight-reading in addition to examining moderator variables within the constructs. The meta-analysis included 92 studies. The population of studies included published and unpublished dissertations and theses, published journal articles, and published paper presentations. Analyses were limited to studies that included sight-reading variables correlated with interval-level performance measures. Unpublished journal articles were omitted from analyses.

One hundred and fifty-four variables and moderator variables that spread across 597 correlations were coded and categorized into 17 constructs. Variables included (a) years of music experience, (b) mode of experience, (c) music achievement measures, (d) theory class grades, (e) academic achievement measures, (f) GPA scores, (g) ear training measures and class grades, (h) measures of sing/play by ear ability, (i) IQ/memory measures, (j) perceptual speed measures, (k) reaction time measures, (l) sight-reading ability/instruction/practice, (m) personality measures, (n) leadership ability, (o) anxiety measures, (p) practice time, (q) technical ability, (r) music interest, (s) music preference, (t) music commitment, and (u) parental involvement. Moderator variables

included publication type, experience level, sight-reading test, and performance tool (instrumental/vocal). Five hundred and ninety-seven correlations between sight-singing and a given variable were categorized into 17 constructs including music aptitude ($n = 158$), music study ($n = 111$), music knowledge ($n = 64$), academic achievement ($n = 49$), ear training ability ($n = 45$), IQ ($n = 28$), perception ($n = 21$), psychomotor ($n = 20$), sight-reading ability ($n = 18$), personality ($n = 17$), practice ($n = 17$), technical ability ($n = 12$), age ($n = 11$), attitude ($n = 8$), early exposure ($n = 8$), memorization ($n = 7$), and improvisation ($n = 4$) (Mishra, 2014).

Constructs and sight-reading analyses yielded both strong and weak positive correlations. Constructs of improvisation ($r = 0.65$), ear training ($r = 0.54$), technical ability ($r = 0.48$), music knowledge ($r = 0.48$), and age ($r = 0.41$) yielded the strongest statistically significant correlations within the homogeneous sample. Constructs of practice ($r = 0.30$), IQ ($r = 0.28$), perception ($r = 0.19$), memorization ($r = 0.18$), and early exposure ($r = 0.11$) yielded the weakest statistically significant correlations within the homogeneous sample. Constructs of personality ($r = 0.12$) and attitude ($r = 0.11$) both yielded weak statistically insignificant effect sizes, which the investigator concluded as being unrelated to sight-reading (Mishra, 2014).

A Q statistic moderator analysis was used to measure differences between levels within constructs. Effect sizes for moderator variables were reported with outlier influence that may have accounted for the significance found between levels (Mishra, 2014). For publication type, effect sizes were significantly different across levels within the attitude construct where the effect size for published studies ($d = -0.59$) was lower than for unpublished theses ($d = 0.12$) and unpublished dissertations ($d = 0.22$). For sight-reading mode (instrumental sight-reading and vocal sight-singing), effect sizes were significantly different across levels within the practice (sight-reading $d = 0.26$, sight-singing $d = 0.49$) and technical ability constructs (sight-reading $d = 0.76$, sight-singing $d =$

0.38). Within the practice construct, the effect size for the sight-reading mode ($d = 0.26$) was significantly lower than for the sight-singing mode ($d = 0.49$); however, within the technical ability construct, the effect size for the sight-reading mode ($d = 0.76$) was significantly higher than for the sight-singing mode ($d = 0.38$).

For experience level, effect sizes were significantly different across levels within the age, ear training, IQ, music aptitude, music study, personality, and technical ability constructs. Mishra (2014) attributed the significant differences across levels as a consequence of a disproportionate number of studies that included secondary and college musician populations and excluded elementary and college non-musician populations. Additionally, Mishra (2014) added that there was “no discernible pattern to the results, but for the constructs where level was a significant factor, most correlations were higher for secondary students, with lower correlations between the given construct and sight-reading for the college musicians” (p. 459). Effect sizes were stable across levels within the constructs of academic achievement, attitude, early explore, improvisation, memorization, music knowledge, perception, practice, psychomotor, and sight-reading ability.

Specific sight-reading tests cited in Mishra’s (2014) meta-analysis included researcher-composition tests, the Watkins-Farnum Performance Scale, grades, ratings, and published repertoire. Effect sizes correlating sight-reading tests and variables were significantly different across levels within the attitude, music aptitude, music study, and technical ability constructs. Where reported, ratings and grades tended to have lower effect sizes for constructs while the Watkins-Farnum Performance Scale tended to have higher effect sizes. The constructs of improvisation, ear training, technical ability, and music knowledge had the strongest correlations with sight-reading. Mishra (2014) concluded that the correlations might support the theory that sight-reading is a developed skill, as opposed to stabilized independent variables, and may be improved with practice (Mishra,

2014). “Attempts to investigate the causes of this variability may begin to explain the correlations, which seems to be a more productive approach than continuing to report correlations” (Mishra, 2014, p. 463).

CHAPTER III

METHOD

The purpose of the study was to address the varied and somewhat stratified study results within the area of singing ability and instruction by statistically summarizing the data of related studies through meta-analysis. Toward this purpose, three meta-analysis research questions were developed. These included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children ages 5 through 11 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task; and (3) what are the statistical effects of study characteristics within and across the following secondary moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period.

The population of the meta-analysis included 34 studies that measured the effects of instruction on singing ability. An initial search through literature yielded 161 possible studies. Application of the inclusion and exclusion criteria resulted in the exclusion of 127 studies. See Appendix Tables C.1, C.2, C.3, and C.4 for excluded or unused article citations and rationale.

Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were outlined to define and narrow the literature search. Initial inclusion criteria included eight statements, and initial exclusion criteria included seven statements (See Table 1). The justifications for inclusion and exclusion statements were based upon the purpose of the meta-analysis, existing singing ability literature, and recommendations from published meta-analyses and texts regarding meta-analytic procedures.

One component of the meta-analysis was generalizing singing ability information for elementary music teachers in an average elementary music classroom with elementary-aged children. Three inclusion and exclusion statements were written to ensure generalizability. First, to be included, studies must have investigated singing ability of children within an age range typically found in elementary schools (5 through 11 or 12), excluding children younger than kindergarten and older than sixth grade. Second, because classrooms generally have a mixture of neurotypical and neuro-atypical children, the meta-analysis only included studies reflecting that population; it excluded special populations such as auditioned choirs (where a choir would equal one participant) and non-neurotypical children. Third, singing analysis must have required participants to sing independently for at least part of the data collection, because children may require opportunities to sing independently for singing ability development (Hedden, 2012); studies without independent singing as part of data collection were excluded.

Because the meta-analysis sought to address singing ability and instruction, included studies were limited to the effects of instruction on singing ability. Most other constructs were permitted if instruction was also involved and if singing ability was measured through singing performance with a few exceptions. Studies with children who spoke tonal languages were excluded due to variables that may attribute to singing development having more to do with language acquisition than with singing instruction. Studies regarding language acquisition were also excluded.

Studies could only be included if they reported necessary descriptive statistics. Means, standard deviations, and sample sizes are needed to calculate effect sizes (Card, 2012). F-test or t-test statistics are also acceptable when means, standard deviations, and sample sizes are not

readily available. If the study satisfied all other criteria, correspondence with the study researcher was anticipated to gather required information.

Meta-analyses, like most quantitative analyses, should be carried out (and reported) in a manner that would enable approximate replication. For replication purposes, and due to the researcher's proximity to interpreters, only studies written in English (or having already been translated to English) were included in the meta-analysis. For replication purposes, studies within the meta-analysis were also limited to those published in peer-reviewed journals as well as unpublished doctoral dissertations.

Limiting research designs may enable effect sizes to be visibly comparable for those who are unfamiliar with meta-analytic procedures, and it may minimize variability of effects accounted for by research design. Therefore, the audience could focus on the effects of instruction, specifically. Three statements were written to indicate included or excluded research designs. First, the study pool was limited to those that utilized a pretest-posttest control group design or a posttest-only research design. Second, studies needed to be quantitative and experimental or quasi-experimental. Third, longitudinal studies were excluded as results may have been more indicative of singing voice development rather than overall singing ability.

Content Validity

Three experts in the field were contacted to establish content validity of the inclusion and exclusion criteria list and initial list of studies. Experts A and B were senior researchers who had published studies that were represented in the meta-analysis. Expert C was chosen based on high representation of research on singing ability. Invitation letters and initial list letters were emailed to each member of the content validity panel. A copy of the letter is located in Appendix A. See Table 2 for the revised version of the inclusion and exclusion criteria statements.

Members of the content validity panel provided suggestions regarding the overall content of the analysis, specific suggestions regarding inclusion criteria, and specific suggestions regarding exclusion criteria. Suggestions regarding the overall content of the analysis included clarifying the age range and clarifying the term *special needs*. Revisions were made accordingly.

The panel members questioned limiting the analysis to quantitative studies and limiting the analysis to the age range of 5 to 11 years, which began a dialogue regarding future meta-analytic research on singing ability. It was the researcher's stance that the meta-analysis be as narrowly defined as possible regarding the research design with the intention of later expanding. The content validity panel suggested that the expanded work include qualitative research, longitudinal research, and case studies.

The panel members suggested including studies located in book chapters and Master's theses. The suggestion regarding book chapters was taken under advisement, however unsuccessful due to lack of reported data. Book chapters were therefore left off of the inclusion and exclusion list. Master's theses were also not added to the inclusion list due to the tendency to be excluded from other comparable meta-analyses as well as general inaccessibility.

The panel members also commented on the exclusion list criteria with suggestions for including tonal-language studies, language acquisition studies, and speaking voice studies. Revisions were not made. It was the researcher's decision to continue to exclude the above research topics due to the goal of narrowing research to the effects of music instruction on singing ability.

Panel members suggested including publication date range criteria. Publication date range criteria was not added because of the limited number of studies regarding singing ability,

and because there was not a need. The oldest qualifying study was from the 1960s, and did not warrant a date restriction.

The content validity panel also suggested 12 additional studies to include in the meta-analysis. A total of six studies fit the updated inclusion and exclusion criteria. The remaining six studies did not fit the updated inclusion and exclusion criteria and were therefore excluded from the current study.

Table 1

Initial Inclusion/Exclusion Criteria Statements

Inclusion Criteria	Exclusion Criteria
(a) Studies must have been written in English.	(a) Studies must not have been longitudinal.
(b) Studies must have included effect size information or sufficient data used to compute and effect size.	(b) Studies must not have investigated relationships between singing ability and other constructs (absent of instruction).
(c) Singing ability must have been represented by measuring a characteristic or characteristics (i.e., pitch accuracy, use of singing voice) of singing performance.	(c) Studies must not have included an investigation regarding the singing ability of children who speak a tonal language (i.e., Cantonese, Taiwanese).
(d) Singing performances must have required students to sing independently for at least part of the data collection.	(d) Studies must not have included preschool, middle school, and junior high school participants.
(e) Studies must have been quantitative with experimental or quasi-experimental research designs.	(e) Studies must not have measured the singing ability of groups of children (as in a choral setting) where the group was measured $n = 1$.
(f) Studies must have investigated singing ability, also referred to in research as singing achievement or singing development.	(f) Studies must not have included results that are only generalizable to specific populations (e.g., special needs children, children's choirs, and private voice students).
(g) Studies must have been published within a peer-reviewed journal or doctoral dissertation.	(g) Studies must not have included investigations of singing ability with language acquisition.
(h) Study sample population must have been limited to children between the ages of 5 and 11 or kindergarten through fifth grade (as labeled in the American school system) as well as sixth grade under the circumstance that sixth grade was part of the elementary school.	

Table 2

Revised Inclusion/Exclusion Criteria Statements

Inclusion Criteria	Exclusion Criteria
(a) Studies must have been written in English.	(a) Studies must not have been longitudinal.
(b) Studies must have included a combination of sample size, group means and standard deviations, and f-test/t-test statistics to compute effect size information where effect sizes were not already available.	(b) Studies must not have only investigated relationships between singing ability and other constructs (absent of instruction).
(c) Singing ability must have been represented by measuring a characteristic or characteristics (i.e., pitch accuracy, use of singing voice) of singing performance.	(c) Studies must not have included an investigation regarding the singing ability of children who speak a tonal language (i.e., Cantonese, Taiwanese).
(d) Singing performances must have required students to sing independently for at least part of the data collection.	(d) Studies must not have included adult, preschool-, middle school-, junior high school-, and college-aged participants (as defined by the American school system scale).
(e) Studies must have been quantitative with experimental or quasi-experimental research designs.	(e) Studies must not have measured the singing ability of groups of children (as in a choral setting) where the group was measured $n = 1$.
(f) Studies must have investigated singing ability, also referred to in research as singing achievement or singing development.	(f) Studies must not have included results that are only generalizable to specific populations (e.g., special needs children, children's choirs, and private voice students) beyond music-specific populations (e.g., certain/uncertain singers, remedial singers, and talented singers).
(g) Studies must have been published within a peer-reviewed journal or doctoral dissertation.	(g) Studies must not have included investigations of singing ability with language acquisition.
(h) Study sample population must have been limited to children between the ages of 5 and 11 or kindergarten through fifth grade (as labeled in the American school system) as well as sixth grade under the circumstance that sixth grade was part of the elementary school.	
(i) Study sample population must be generalizable to neurotypical children.	

Literature Search Strategies

Reference Databases Searched

An exhaustive literature search was performed in order to pool studies related to instructional effects of singing ability. Key search phrases included *singing ability child*, *singing achievement child*, *vocal development child*, *teaching singing child*, and *singing development child*. Initial searches utilized Summon, which is an online service that compiles lists of relevant peer-reviewed journal articles from journal publishers. Searches were also conducted within the following electronic databases: EBSCOhost, ERIC, PsycInfo, Sage journals online, ScienceDirect, Wiley-Blackwell Journals, JSTOR, and ProQuest dissertations and theses. Additional articles were found through backward searches (gleaning articles from existing reference lists), forward searches (searching for articles that have cited seminal works or seminal researchers), and by contacting several seminal researchers in the field.

White (2009) described four modes of research dissemination: “informal oral, informal written, formal oral, and formal written” (p. 52). For the purpose of the meta-analysis, only formal written was included, as “Its typical products – papers, articles, monographs, and reports – constitute the primary literatures by which scientists lay open their work to permanent public scrutiny in hopes that their claims to new knowledge will be validated and esteemed” (White, 2009, p. 52).

Coding Procedures

Coding procedures provide transparency in grouping and labeling studies, should the meta-analysis be replicated. Definitions of coding categories, intracoder reliability, and study quality explain coding procedures. The coding was performed by the primary researcher.

Definitions of Coding Categories

Fourteen coding categories were defined for the current meta-analysis: (a) citation, (b) population sample size, (c) subgroup sample sizes, (d) age, (e) dependent variable (measurement instrument), (f) teaching condition, (g) population, (h) research design, (i) publication source, (j) task measured, (k) treatment period, (l) random sample, and (m) random assignment. Data within the categories were used for the moderator analyses and the Chapter V study quality discussion.

Categories were coded with numbers, words, or both. The citation category included the author(s) and the date of the study. The sample size category included the overall sample size, and the subgroup sample sizes included sample sizes with regard to treatment, grade, age, and/or gender. The age category included information regarding the mean age of participants or general age ranges of participants. A formula was used to convert grades to ages where needed: $\text{Age} = \text{Grade} + 5$ (Card, 2012, p. 80).

The dependent variable (measurement instrument) was used to code and group the singing tests by like-tests and commonalities. Two categories were labeled by name due to frequent representation within the research: (a) SVDM, and (b) Boardman. The subsequent categories included the following characteristics: (c) acoustic measures measuring pitch accuracy, (d) tests that were similar to the SVDM, measured pitch accuracy, and were researcher developed, (e) tests that measured pattern or interval accuracy and scored by ear (researcher as acoustic), (f) tests that measured song or tonality accuracy and scores by ear, and (g) tests that primarily measured children's use of their head voice. Tests were also coded based on by whom they were designed: (a) researched designed, researcher as acoustic, (b) previously designed, researcher as acoustic, (c) researcher designed, acoustic, and (d) previously designed, acoustic.

The instruction moderator variable included the specific information regarding how two or more groups were compared (e.g., music instruction vs. no music instruction) and was coded into four categories with subsequent subgroups. The first category, the effect of teacher skills or behaviors on singing ability, included the following subgroups: (a) vocal model, (b) feedback, and (c) use of singing (singing for/with students). The second category, the effect of instructional procedures on singing ability, included the following subgroups: (a) literature dissemination (process of teaching a song), (b) presence and type of instruction, (c) pattern instruction, (d) use of accompaniment, and (e) the effect of practice on singing ability. The third category, the effect of performance context on singing ability, included the following subgroups: (a) individual verses group singing, (b) response modes (handsigns, pitch syllables, neutral syllables, text), (c) movement, and (d) song range.

The population category was coded based on the region of the sample population: (a) Northeast, (b) Southeast, (c) Southwest, (d) West, (e) Midwest, (f) Hawaii, (g) Canada, and (h) England. A list of regions and their corresponding states can be found in Appendix B. The research design category was coded as: (a) pretest-posttest control group design, (b) multiple treatment pretest-posttest control group design, (c) multiple treatment pretest-posttest design, (d) posttest-only control group design, and (e) multiple treatment posttest-only control group design, and (f) posttest-only control group design. The publication source category was coded as (a) Journal of Research in Music Education (JRME), (b) The Bulletin of the Council for Research in Music Education (CRME), (c) Psychology of Music (POM), (d) Update: Applications of Research in Music Education (Update), (e) dissertations, and (f) Research Issues in Music Education (RIME). The singing task measured category was initially coded as (a) pitch accuracy, (b) pitch accuracy, use of singing voice, and range, (c) pitch accuracy, range, (d) pitch accuracy,

pattern accuracy, melodic contour, (e) pitch accuracy, melodic contour, tonality accuracy, use of singing voice, (f) pattern accuracy, (g) pattern accuracy, use of singing voice, (h) pattern accuracy, tonality accuracy, use of singing voice, (i) interval accuracy, melodic contour, use of singing voice, (j) song accuracy, (k) tonality accuracy, and (l) use of singing voice. Treatment period category included the amount of time that the treatment was given as measured in weeks. Categories random assignment and random sample were coded as (a) yes, or (b) no.

Intracoder Reliability

In a meta-analysis, intracoder reliability has been cited as the extent to which study characteristics were coded consistently (Card, 2012). The Agreement Rate (AR) formula has been generally accepted as the most commonly used formula where $AR = (\text{number of observations agreed upon}) / (\text{total number of observations})$ (Card, 2012; Orwin & Vevea, 2009). AR does not control for coding agreement happening by chance (as does Cohen's Kappa, for example). Controlling for chance, however, was not necessary in the current study, because cells were either in agreement or disagreement.

To check for intracoder reliability of coding procedures in the main study, the researcher recoded all of the studies two months after initial coding. Nine categories across 32 common studies yielded 288 total common observations. Of the 288 total common observations, there were 263 agreements and 25 disagreements, yielding a tenable AR of 91.32%. The AR of 70% or higher has been commonly cited as acceptable across studies (Card, 2012).

Although the intrarater reliability was determined tenable, it only measured common cells between the coding and recoding. Initial coding yielded 592 cells (16 cells per study x 37 studies). Recoding yielded 490 cells (14 cells per study x 35 studies). Differences in cells were attributed to the changes resulting from the pilot study, studies added after the initial coding, and

studies excluded from recoding. Changes in collected information resulted from pilot procedures (specifically redefining *task measured* category).

Common cell disagreement was attributed to differences in defining the singing task category and differences in sample size. *Task measured* was rectified after the pilot study, and differences between initial and recoding were expected due to redefining the variable between coding and recoding. The initial coding definition of singing task included information regarding what children sang as part of the singing test. The variable was redefined to reflect what specifically was measured regarding the singing test, thereby also reflecting what was sung. For example, instead of coding that the child sang *All Around the Butterscup* for the singing task, song accuracy may have been listed if that was how the study researcher chose to measure accuracy. Song accuracy reflects that the child sang a song, and that the entire song was important to the score. Contrarily, the child could have been expected to sing an entire song, but only specific pitches were scored as a measure of accuracy. The singing task measured would therefore be coded as pitch accuracy instead of song accuracy. In the context of the current study, how children were scored was of greater consequence than what children sang.

Sample size differences were attributed to the extraction of data for initial coding from study abstracts. All recoded information was extracted from the body of the study. The body of the study tended to provide a more detailed description of the sample.

Study Quality

Study quality refers to “the fit between a study’s goals and the study’s design and implementation characteristics” (Valentine, 2009, p. 130). Addressing study quality in a meta-analysis provides a transparent picture of results regarding validity and generalizability. The

variation and rigor of individual study characteristics “can have an impact on the validity of conclusions arising from that synthesis” (Valentine, 2009, p. 130).

Three methods have been cited for measuring study quality within a meta-analysis (Card, 2012; Cooper et al., 2009). The first method is measuring study quality by using a pre-existing scale that determines an individual study’s quality score and by excluding studies that do not meet a specific cutoff score. The second method includes study quality as a moderator variable that is represented continuously by a pre-existing scale score or that uses cut-offs to categorize scores dichotomously (e.g., by high, medium, and low quality). For the purpose of the current meta-analysis, study quality was instead accounted for by the third method that uses study characteristics as reflective indicators of quality. The third was most strongly recommended by Card (2012).

Study quality was written into inclusion and exclusion criteria, and study quality was written into moderator variables that represented study characteristics sensitive to internal and external validity. Study quality measured as such, “requires conceptual and empirical co-occurrence among coded study features,” as opposed to the former two methods, which use study characteristics as “formative indicators of study quality” (Card, 2012, p. 70). Therefore, study quality is discussed through components of the meta-analysis designed to control threats to external validity, internal validity, and construct validity. Study quality is discussed in the following section and revisited in Chapters IV and V as reflections of effect size results.

Statistical Methods

The current meta-analysis has three consecutive parts that encompass multiple calculations. The first part yielded an overall weighted mean effect size across all studies within

the current meta-analysis. The second part investigated heterogeneity among effect sizes. The third part investigated moderator variables.

Several options were available to analyze current meta-analytic processes. Card (2012) recommended Excel, SPSS, or SAS for young researchers over meta-analysis computer programs. Card (2012) elaborated,

The danger of relying on them [meta-analysis programs] exclusively – especially when you are first learning to conduct meta-analyses – is that they may encourage erroneous use when you are not adequately familiar with the techniques. (p. 14)

Although SPSS macros were located to aid in the use of SPSS for meta-analytic analyses, the syntax commands were out of date and did not align with SPSS 22. SPSS 22 was used, however, to compute group standard deviations when data were not provided within a study. Excel was used to calculate equations in the current meta-analysis where warranted, specifically regarding calculating p-values and sums. All other calculations were done on a TI-83 graphing calculator.

Effect Size Calculations

Hedges' g was chosen to represent the effect sizes in the current meta-analysis. Hedges' g is nearly identical to Cohen's d , with the only difference existing in the denominator of the equations. In Cohen's d , the mean difference is divided by the pooled standard deviation of both groups, whereas in Hedges' g , the mean difference is divided by the pooled standard deviation of both groups weighted by each group's sample size. Cohen's d assumes groups are the same size, whereas Hedges' g is most appropriate for groups of difference sizes. Hedges' g is used in the current main study and pilot study meta-analyses due to inequivalent group sizes within individual studies.

Cohen (1988) provided cutoffs for small, medium, and large d effect sizes of 0.2, 0.5, and 0.8, respectively. Effect sizes lower than 0.2 have been labeled as trivial (Ellis, 2010). The cutoffs have also been commonly generalized to Hedges' g effect sizes (Card, 2012; Ellis, 2010). According to Cohen (1988), small effects have been relatively common, "when phenomena are studied which cannot be brought into the laboratory, the influence of uncontrollable extraneous variables ("noise") makes the size of the effect small relative to these" (p. 25). A medium effect, or any effect larger than 0.5, has been defined as, "one large enough to be visible to the naked eye" (Cohen, 1988, p. 26).

Caution is requested when referring to effect sizes in terms of a cutoff index. Interpreting effects using cutoffs may provide the audience with an accessible context for the effects; cutoffs, however, may be misleading and overgeneralize or minimize importance. Two rationale for caution include sample size issues and judgment issues. The overall sample sizes in a meta-analysis tend to exhibit a large range of samples and a large range of group sizes. As Rosnow, Rosenthal, and Rubin (2000) stated,

Although sample sizes smaller than 30 have often been considered acceptable in psychology, it would be difficult (power $\sim .12$) for effects that are commonly characterized as "small" ($g = 0.20$) or "medium" ($g = 0.50$; power = .46) to be found significant at the .05 level when the smaller of the two samples is less than 30. (p. 446)

The work that went into developing teaching strategies, implementing teaching strategies, and researching the strategies requires skill, knowledge, and passion. It may be important to value each method of instruction within the current meta-analysis with caution when using size-specific labels of small, medium, and large. After all, the effect sizes exist within a broader context.

In order to obtain an overall weighted mean effect size for the main study: $\overline{ES} = \Sigma(w_i ES_i) / \Sigma w_i$ (Card, 2012), the researcher first calculated effect sizes based on study subgroups. Hedges' g effect sizes were calculated from the reported means and standard deviations of instruction effect for the following designs: change from pretest to posttest: $g_{\text{change}} = M_2 - M_1 / [(s_1^2 + s_2^2) / 2]^{1/2}$ (Card, 2012), and between-group effects: $ES = (M_1 - M_2) / s_{\text{pooled}}$ (Card, 2012). Cutoff numbers for small sample adjustments to effect sizes vary, therefore the equation was applied to all effect sizes calculated with groups smaller than 20 participants: $g_{\text{adjusted}} = g - \{3g / [4(n_1 + n_2) - 9]\}$ (Lipsey & Wilson, 2001).

The variance of the effects of change from within-group pretest to posttest mean scores was calculated: $\{[2(1 - r) / N] + (g_{\text{change}}^2 / 2N)\}$ (Card, 2012; Lipsey & Wilson, 2001). The variance of effects between-groups was also calculated: $v = (n_1 + n_2) / n_1 n_2 + ES_{sm}^2 / 2(n_1 + n_2)$ (Borenstein, 2009). Of the 35 studies, six studies included correlations (or raw scores to find correlations) between pretest and posttest scores needed to calculate v_{change} (Feierabend, 1984; Kim, 2000; McGraw, 1996; Montgomery, 1988; Richner, 1976; Stauffer, 1985). Card (personal communication, May 7, 2015) verified that the r in the equation could be substituted with an average of all correlations made within and across the six studies. Available raw data yielded a total number 39 correlations that ranged from -0.31–0.98 ($r = 0.75$, $SD = 0.25$). As an aside, Borenstein's (2009) v equation was intended for Cohen's d effect sizes. The researcher, however, chose the equation for two reasons: (1) Borenstein's d effect size equation was identical to g equations found elsewhere with the addition of his $J(df)$ correction for Hedges' g (i.e. Card, 2012), and (2) because Borenstein (2009) stated,

The expression given here for variable of g is J^2 times the variance of d , but many authors ignore the J^2 term because it is so close to unity in most cases. Again, though it is

preferable to include this correction [$J(df)$] factor, including it is likely to make little practical difference. (p. 227)

Effect size variances and inverse variance weights were calculated: $SE_g = v_g^{1/2}$ (Borenstein, 2009) and $w_i = 1/SE^2$ (Card, 2012). Weights were important in calculating mean effect sizes as they accounted for precision issues when estimating population effects due to sample sizes. “The standard error of an effect size, which is inversely related to sample size, quantifies the amount of imprecision in a particular study’s estimate of the population effect size” (Card, 2012, p. 176). Each individual’s study mean effect size and weight was then used to calculate the overall weighted mean effect size for the study: $\overline{ES} = \Sigma(w_i ES_i) / \Sigma w_i$ (Card, 2012). Effect size calculations were limited to effects of change (pretest to posttest within a group) and effects for between-groups, because all studies included the element of instruction over time.

Alternate equations were used to determine effect sizes in studies where means and/or standard deviations were not reported in studies. In several studies, effect size calculations utilized the F statistic (Newell, 2013; Roberts & Davies, 1975; Rutkowski, 1996): (a) between-groups with unequal sample sizes $g = [F_{(1, df)}(n_1 + n_2)/n_1 n_2]^{1/2}$ (Card, 2012), (b) between-groups with equal sample sizes $g = 2(F_{(1, df)}/N)$ (Card, 2012), and (c) repeated measures $g = [F_{\text{repeated } (1, df)}/N]^{1/2}$ (Card, 2012). In one study, effect size calculations utilized the X^2 statistic (Hedden & Baker, 2010): $2\{[X^2_{(1)} / (N - X^2_{(1)})]^{1/2}\}$ (Card, 2012). In one study, effect size calculations utilized p values (Hedden & Baker, 2010): $g = 2Z/(N)^{1/2}$ (Card, 2012).

Wilson’s Practical Meta-Analysis Effect Size Calculator, an online tool designed to accompany a text (Lipsey & Wilson, 2001), was used to calculate effects for a two-way ANOVA (Klinger, Campbell, & Goolsby, 1998). In one study, effect sizes were calculated from independent t -scores (Paney & Kay, 2014): $g = t_{\text{dependent}}/N^{1/2}$ (Card, 2012). Lastly, two studies

provided the mean change between pretest and posttest scores and the standard deviation of the corresponding mean change (Buck, 1991; Paney & Kay, 2014). The researcher calculated the effects based on the following logical formula: $g_{\text{change}} = M_{\text{change}}/SD_{\text{change}}$.

Attempts were made to obtain raw scores and standard deviations when sufficient data were not available. In one study, the primary researcher had passed away, and the original data were no longer available (Klinger, Campbell, & Goolsby, 1998). Another study reported a three-way ANOVA table (Small & McCachern, 1983), and the primary researcher no longer had original data from which to gather descriptive statistics. The study was therefore excluded from the current meta-analysis. One study reported partial MANCOVA tables (Phillips, 1985); attempts to retrieve more data were untenable. One study included means and values generated from Wilcoxin Matched Pairs Signed-Ranks Tests (Mathais, 1997). Results from Wilcoxin Matched Pairs Signed-Ranks Tests are based on median scores, and the primary researcher could not locate raw accuracy scores; the study was therefore excluded from the current meta-analysis. Attempts to gather additional data from primary researchers were unsuccessful due to initial non-response (Holmes, 2009, Wyatt, 1993), follow-up non-response (Paney & Kay, 2014), and lost data (Klinger et al., 1998; Phillips & Aitchison, 1997).

Once an overall effect size was determined across the studies, standard errors of the mean effect sizes were calculated: $SE_{ES} = (1/\Sigma w_i)^{1/2}$ (Card, 2012). SE_{ES} 's were then used to compute a Z score: $Z = \overline{ES} / SE_{ES}$ (Card, 2012). Z scores were used to determine if the effect size was significantly greater than zero. Standard errors of the mean effects were also used to compute confidence intervals: $ES_{LB} = ES - Z_{1-\alpha} SE_{ES}$ and $ES_{UB} = ES + Z_{1-\alpha} SE_{ES}$ (Card, 2012).

Next, a fixed effect Q statistic was calculated to determine if study effects were homogeneous or heterogeneous: $Q = \Sigma(w_i ES_i^2) - [\Sigma(w_i ES_i)]^2 / \Sigma w_i$ (Card, 2012). Effect size

heterogeneity would indicate that effect size deviation across studies was due to factors in addition to factors beyond sampling error expectations (Card, 2012). If a Q test indicates that study effects are homogeneous, then the meta-analysis could be considered empirically complete and further investigation into moderators would be theoretically driven. If a Q test indicates that studies effects are heterogeneous, however, moderator analyses can assist in determining the source of variation (Card, 2012).

In addition to calculating a Q statistic, an I^2 statistic was calculated to determine the magnitude of heterogeneity: $I^2 = \{[Q - (k - 1)]/Q\} \times 100\%$ when $Q > (k - 1)$ or the equation $\times 0$ when $Q < (k - 1)$ (Card, 2012). “The I^2 index can be interpreted as the percentage of the total variability in a set of effect sizes due to true heterogeneity, that is, to between-studies variability” (Huedo-Medina, Sanchez-Meca, & Marin-Martinez, 2006, p. 194). Huedo-Medina et al. (2006) listed associated small, medium, and large magnitude cutoffs as around 25%, 50%, and 75%, respectively.

Upon determining heterogeneity, the current meta-analysis investigated categorical and continuous moderator variables within subgroups. The process for determining weight mean effect sizes within categorical subgroups reflected the process mentioned above: calculate mean effect size, standard error of the mean effect size, variance, and mean effect size weights before computing a Q statistic for each applicable subgroup. Subgroup Q statistics were then used to determine if subgroups had mean effect sizes that were significantly different from the overall weighted mean effect size: $Q_{\text{total}} = Q_{\text{between}} + Q_{\text{within}}$ (Card, 2012, p. 200).

Continuous moderator variables within subgroups utilized the Conventional T test: $T = r_{Xd}/s_{rXd}$ where $s_{rXd} = [(1 - r_{Xd}^2)/(k - 2)]^{1/2}$ (Sanchez-Meca & Marin-Martinez, 1998). According to Sanchez-Meca and Marin-Martinez (1998), “the correlational nature of the meta-analytic data

implies that the relationship between the X and d variables assumes a bivariate distribution” (p. 314). There are two other more current methods of determining effects of continuous data including the Rosenthal and Rubin procedure and the Hedges and Olkin procedure. Differences across the three methods were considered negligible regarding Type I error rates (Sanchez-Meca & Marin-Martinez, 1998).

For the purpose of the current study, results may be generalized to the population of the current studies and “to a population of studies with similar characteristics than those of represented in the meta-analysis” (Huedo-Medina et al., 2006, p. 4). Therefore, the meta-analysis utilized a conditional inference model with fixed-effects procedures. “The model is called conditional because it can be conceived as a model that holds fixed, or conditions on, the characteristics of studies that might be related to the effect-size parameter” (Hedges, 2009, p. 39). The statistical model and moderator variables were selected a priori based on a conceptual assumption that the effects within the current meta-analysis would be heterogeneous.

Pilot Study

A pilot study was implemented to estimate the feasibility, consistency, and transparency of the meta-analytic process. The four pilot study research questions included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children younger than 5 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task; and (3) what are the statistical effects of study characteristics within and across the following moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period; and (4) was the analytic process feasible? The pilot meta-

analysis was determined feasible and successful in establishing a procedure for the main study.

Additional pilot study information may be found in Appendix D.

CHAPTER IV

RESULTS

The purpose of the study was to address the varied and somewhat stratified study results within the area of singing ability and instruction by statistically summarizing the data of related studies through meta-analysis. Toward this purpose, three meta-analysis research questions were developed. These included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children ages 5 through 11 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task; and (3) what are the statistical effects of study characteristics within and across the following secondary moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period.

The population of the current meta-analysis was comprised of 34 studies with a combined sample size of $N = 5,497$ participants who ranged in age from 5- to 11-years ($M = 7.38$, $SD = 1.86$). Studies represented all regions of the United States, Canada, and England. Refer to Table 3 for information regarding each study grouped by teaching condition and instructional details listed for treatment (T) and control (C) groups. Refer to Table 4 for information regarding all other primary and secondary moderators by study.

Table 3

Teaching Conditions by Moderator Subgroup

Study	Year	<i>N</i>	<i>The effect of teacher skills or behaviors on singing ability</i>
			Vocal model
Mang	1997	83	(T) Female model (T) Male model
Michaud	2014	66	(T) Baritone model (T) Falsetto model

(table continues)

Teaching Conditions by Moderator Subgroup (continued)

Study	Year	N	<i>The effect of teacher skills or behaviors on singing ability</i>
			Vocal Model
Montgomery	1988	40	(T) Falsetto (T) Normal range
Persellin	2006	134	(T) Teacher sang for students (T) Teacher sang with students (C) Teacher sang for and with students
			Feedback
Buck	1991	42	(T) Pitch master (C) Traditional
Paladino	1991	171	(T) Behavioral training with non-technical, self-monitored aural feedback device (C) Behavioral training without feedback device
Paney & Kay	2014	2021	(T) Practice with computer program (C) No practice with computer program
Rutkowski & Snell Miller	2003	38	(T) Specific feedback after singing in a large group, small group, or individually followed by modeling (C) Generic feedback
Welch, Howard, & Rush	1989	32	(T) Experimental interactive regular singing sessions using SINGAD with adult feedback (T) Experimental non-interactive singing sessions using SINGAD (C) Traditional music program
			<i>The effect of instructional procedures on singing ability</i>
			Literature dissemination (process of teaching a song)
Klinger, Campbell, & Goolsby	1998	39	(T) Rote (T) Immersion
Marshall	2002	110	(T) Immersion with text (T) Phrase-by-phrase with text (T) Phrase-by-phrase with neutral syllables (T) Rote song with neutral syllables
			Presence and type of singing instruction
Aaron	1990	69	(T) Vocal coordination instruction (C) No vocal coordination instruction
Apfelstadt	1984	61	(T) Vocal instruction with melodic perception (T) Vocal instruction with imitation alone (C) Traditional music instruction
Jordan-DeCarbo	1982	89	(T) Same/different discrimination techniques (T) No same/different discrimination techniques
Kimble	1983	143	(T) Melody pre-training (T) General song practice

(table continues)

Teaching Conditions by Moderator Subgroup (continued)

Study	Year	N	<i>The effect of instructional procedures on singing ability</i>
			Presence and type of singing instruction
Kramer	1985	201	(T) Curriculum with accompaniment, teacher sang with students, various tonalities, all text (T) Gould Specialized Music Program
McGraw	1996	66	(T) Vocal training with songs E2–G1 (T) Repertoire in E2–G1 with no vocal training © Repertoire with songs in lower range of C1–A1
Miyamoto	2005	57	(T) Yuba method © No remedial treatment
Phillips & Aitchison	1997	269	(T) Psychomotor skills instruction © Traditional song approach
Porter	1977	80	(T) Multiple discrimination training © No multiple discrimination training
Richner	1976	77	(T) Traditional music instruction (T) Instruction emphasizing all areas of music education (T) Singing songs only (T) Remedial vocal training
Roberts & Davies	1975	90	(T) Remedial training group (T) Traditional training group © Control
			Pattern Instruction
Feierabend	1984	75	(T) Patterns easy to perceive aurally regardless of difficulty (T) Patterns easy to regardless of aural difficulty (T) Patterns easy to sing and easy to discriminate aurally © No tonal pattern instruction
Jarjisian	1981	96	(T) Diatonic pattern instruction (T) Pentatonic pattern instruction (T) Combination
Kane	1994	87	(T) Tonal preparatory audiation teaching © Traditional teaching
			Use of accompaniment
Atterbury & Silcox	1993	205	(T) No piano accompaniment © Piano accompaniment
Guilbault	2009	136	(T) A cappella (T) Root accompaniment
Hedden & Baker	2010	26	(T) A cappella (T) Accompaniment

(table continues)

Teaching Conditions by Moderator Subgroup (continued)

Study	Year	N	<i>The effect of instructional procedures on singing ability</i>	
			Use of accompaniment	
Stauffer	1985	310	(T) Neither melodic nor harmonic contexts (T) Melodic context (T) Harmonic context/accompaniment (T) Both melodic and harmonic contexts	
			<i>The effect of performance context on singing ability</i>	
			Individual versus group singing	
Rutkowski	1996	99	(T) Small group and individual participation in singing activities (C) No small group and individual	
			Response modes	
Levinowitz	1987	102	(T) Most songs with words (T) Most songs without words (T) All songs with words	
			Movement while singing	
Kim	2000	78	(T) Movement activities while singing (T) No movement activities while singing (C) No movement or music	
Newell	2013	143	(T) Continuous fluid movement instruction (T) Steady-beat movement (C) Listening map instruction	
			Song range	
Rutkowski	1986	162	(T) Restricted range song material C3–B3 (T) Traditional range song material (C) No music instruction	

Note. (T) = treatment; (C) = control.

Table 4

Moderator Variables by Study

Primary Moderators						
Author	Year	N	Age	Gender	Test	Task
Aaron	1990	69	9	x	VPA	Pitch accuracy
Apfelstadt	1984	61	10	-	Boardman Rote song test	Pitch accuracy, use of singing voice Song accuracy
			5			
Atterbury & Silcox	1993	205	5	-	SVDM	Pitch accuracy, use of singing voice (range)
Buck	1991	42	8	-	VPT	Pitch accuracy
			11			
Feierabend	1984	75	6	-	Unnamed	Pattern accuracy (table continues)

Moderator Variables by Study (continued)

Author	Year	N	Primary Moderators			Task
			Age	Gender	Test	
Guilbault	2009	136	5 6	-	TARS	Pattern accuracy, tonality accuracy, use of singing voice
Hedden & Baker	2010	26	7	-	Unnamed	Pitch accuracy
Jarjisian	1981	96	6	-		Interval accuracy, melodic contour, use of singing voice
Jordan-DeCarbo	1982	89	5	x	CST-1 CST-2	Pattern accuracy Song accuracy
Kane	1994	87	5	-	SAT	Song accuracy
Kim	2000	78	6	-	SVDM MST	Pitch accuracy, use of singing voice (range) Tonality accuracy
Kimble	1983	143	9 10	x	Smith	Pitch accuracy
Klinger et al.	1998	39	7	-	Error Frequency	Song accuracy
Kramer	1985	201	8 9	-	Gould	Pitch accuracy
Levinowitz	1987	102	5 6	-	Levinowitz	Pattern accuracy, use of singing voice
Mang	1997	83	6	-	Unnamed	Pattern accuracy, tonality accuracy, use of singing voice
Marshall	2002	110	8	-	SARS PARS	Pattern accuracy
McGraw	1996	66	7	-	Register- related pitch	Pitch accuracy
Michaud	2014	66	5	-	SVDM Unnamed	Pitch accuracy, use of singing voice (range) Pitch accuracy
Miyamoto	2005	57	9 10 11	-	Sona-Speech	Pitch accuracy
Montgomery	1988	40	8	-	Boardman	Pitch accuracy, use of singing voice (range)
Newell	2013	143	8.44	x	PAS SVDM	Pitch accuracy Pitch accuracy, use of singing voice (range)

(table continues)

Moderator Variables by Study (continued)

Primary Moderators						
Author	Year	N	Age	Gender	Test	Task
Paladino	1991	171	6	x	SAT-I	Pitch accuracy, melodic contour, pattern accuracy
			8		SAT-II	Interval accuracy, melodic contour, use of singing voice
Paney & Kay	2014	2021	8	x	Singing Coach	Pitch accuracy
Persellin (2006)	2006	134	5	-	VAll	Pitch accuracy
Phillips & Aitchison	1997	269	9	-	Unnamed	Pitch accuracy, range
			10			
			11			
Porter	1977	80	9	-	VPM	Pitch accuracy
			10			
Richner	1976	77	8	-	PPT	Pitch accuracy
			9			
			10			
Roberts & Davies	1975	90	6	-	Unnamed	Song accuracy
			7			
			8			
Rutkowski	1986	162	5	-	SVDM	Pitch accuracy, use of singing voice (range)
Rutkowski	1996	99	5	-	SVDM	Pitch accuracy, use of singing voice (range)
Rutkowski & Snell Miller	2003	38	6	-	SVDM	Pitch accuracy, use of singing voice (range)
Stauffer	1985	310	6	x	TSA	Pitch accuracy, melodic contour, tonality accuracy, use of singing voice
			7		MET	Pattern accuracy
Welch, Howard, & Rush	1989	32	7	-	SINGAD	Pitch accuracy
Secondary Moderators						
Author	Year	State	Source		Research Design	Treatment Period
Aaron	1990	IA	Dissertation		Pretest-posttest control group	20 weeks
Apfelstadt	1984	WI	JRME		Pretest-posttest control group	11 weeks
Atterbury & Silcox	1993	ME	JRME		Pretest-posttest control group	8 months
Buck	1991	WA	Dissertation		Pretest-posttest control group	6 weeks

(table continues)

Moderator Variables by Study (continued)

Secondary Moderators					
Author	Year	State	Source	Research Design	Treatment Period
Feierabend	1984	PA	Dissertation	Multiple treatment pretest-posttest control group	7 weeks
Guilbault	2009	MI	JRME	Posttest-only control group design	25 weeks
Hedden & Baker	2010	KS	CRME	Multiple treatment posttest-only	3 sessions
Jarjisian	1981	PA	Dissertation	Multiple treatment posttest-only	16 weeks
Jordan-DeCarbo	1982	OH	JRME	Pretest-posttest control group	11 weeks
Kane	1994	PA	Dissertation	Posttest-only control group	4 months
Kim	2000	MS	Dissertation	Multiple treatment pretest-posttest control group	10 weeks
Kimble	1983	GA	Dissertation	Multiple treatment posttest-only	3 weeks
Klinger et al.	1998	WA	JRME	Multiple treatment posttest-only	1 week
Kramer	1985	NJ	Dissertation	Pretest-posttest control group	10 weeks
Levinowitz	1987	NJ	Dissertation	Multiple treatment posttest-only	1 academic year
Mang	1997	Canada	CJRME	Multiple treatment posttest-only	5 weeks
Marshall	2002	PA	Dissertation	Multiple treatment posttest-only	10 weeks
McGraw	1996	GA	Dissertation	Multiple treatment pretest-posttest control group	8 weeks
Michaud	2014	MA	RIME	Multiple treatment pretest-posttest control group	20 weeks
Miyamoto	2005	HI	Dissertation	Pretest-posttest control group	12 weeks
Montgomery	1988	VA	Dissertation	Multiple treatment pretest-posttest	12 weeks

(table continues)

Moderator Variables by Study (continued)

Secondary Moderators					
Author	Year	State	Source	Research Design	Treatment Period
Newell	2013	DE	Dissertation	Multiple treatment pretest-posttest control group	4 months
Paladino	1991	OH	Update	Pretest-posttest control group	6 weeks
Paney & Kay	2014	IA, MN, ND, WI	CRME	Pretest-posttest 1 group	4-7 months
Persellin	2006	TX	JRME	Multiple treatment pretest-posttest control group	8 months
Phillips & Aitchison	1997	IA	Dissertation	Posttest-only control group	27 weeks
Porter	1977	NY	JRME	Multiple treatment pretest-posttest control group	19.06-94.13 minutes
Richner	1976	WA	Dissertation	Multiple treatment pretest-posttest	8 weeks
Roberts & Davies	1975	England	JRME	Multiple treatment pretest-posttest control group	8 weeks
Rutkowski	1986	PA	Dissertation	Pretest-posttest control group	15 weeks
Rutkowski	1996	PA	JRME	Pretest-posttest control group	9 months
Rutkowski & Snell Miller	2003	PA	CRME	Pretest-posttest control group	8 months
Stauffer	1985	MI	Dissertation	Multiple treatment pretest-posttest	12 weeks
Welch, Howard, & Rush	1989	England	POM	Multiple treatment pretest-posttest control group	7 weeks

With regard to the first research question, the current meta-analysis yielded an overall small weighted mean effect size for instruction ($\overline{ES} = 0.43$, 34 studies, 433 unique effects, 95% CI [0.42, 0.44]). Mean effect sizes across the studies ranged from large ($g = 0.71$, 95% CI [0.63, 0.79]) to trivial ($g = 0.03$, 95% CI [-0.10, 0.17]). The largest overall study effect sizes across categorical variables included Jordan-DeCarbo's (1982) investigation regarding the effects of

same and different discrimination techniques on mean score gains, favoring the treatment group ($g_{\text{change}} = 0.71$, 95% CI [0.63, 0.70]) and Porter's (1977) investigation regarding the effects of multiple discrimination training on mean score gains, favoring the treatment group ($g_{\text{change}} = 0.69$, 95% CI [0.46, 0.92]). The smallest overall study effect sizes across categorical variables included Rutkowski's (1996) investigation regarding the effects of small group and individual participation in singing activities on mean scores, favoring posttest treatment scores ($g = 0.03$, 95% CI [-0.10, 0.17]) and Klinger et al.'s (1998) investigation regarding the effects of rote versus immersion song teaching on mean scores, favoring between-group scores of students receiving immersion song instruction ($g = 0.04$, 95% CI [-0.59, 0.67]).

Effect sizes were examined based on calculations of Hedge's g between two independent group mean scores and within one group from pretest scores to posttest scores (g_{change}). Between-group effect sizes were indicative of the estimated magnitude of mean differences between the control and experimental groups. Within-group effect sizes were indicative of the estimated magnitude of gain or change within one group from pretest to posttest mean scores. Refer to Table 5 for additional descriptive information regarding overall mean effects, within-group effects, and between-group effects by study.

The overall effect size regarding pretest to posttest mean gain scores yielded an effect of 0.47 (95% CI [0.46, 0.48]). Mean gain effect sizes across the studies ranged from 0.82 (95% CI [0.51, 1.13]) to (-0.03, 95% CI [-0.17, 0.11]). The largest overall study effect sizes across categorical variables included Porter's (1977) investigation regarding the effects of same and different discrimination techniques on mean gain scores, favoring the treatment group ($g_{\text{change}} = 0.82$, 95% CI [0.51, 1.13]) and Jordan-DeCarbo's (1982) investigation regarding the effects of same and different discrimination techniques on mean gain scores, favoring the treatment group

($g_{\text{change}} = 0.80$, 95% CI [0.75, 0.85]). The smallest overall study effect sizes across categorical variables included Kane's (1994) investigation regarding the effects of tonal preparatory audiation instruction on mean gain scores, favoring the control group ($g_{\text{change}} = 0.04$, 95% CI [-0.59, 0.67]) and Rutkowski's (1996) investigation regarding the effects of small group and individual participation in singing activities on mean gain scores, favoring the control group ($g_{\text{change}} = -0.03$, 95% CI [-0.17, 0.11]).

The overall effect size between two independent group mean scores yielded an effect of 0.25 (95% CI [0.22, 0.27]). Between-group mean effect sizes across the studies ranged from 0.86 (95% CI [0.60, 1.12]) to -0.18 (95% CI [-0.35, -0.01]). The largest overall between-group study effect sizes across categorical variables included Michaud's (2014) investigation regarding the effects of a male falsetto model versus a male baritone model on mean scores, favoring between-group pretest scores (prior to male vocal model instruction) ($g = 0.86$, 95% CI [0.60, 1.12]) and Hedden and Baker's (2010) investigation regarding the effects of accompaniment versus a cappella singing on mean scores, favoring a cappella singing ($g = 0.67$, 95% CI [0.19, 1.15]). The smallest overall between-group effect sizes across categorical variables included Klinger et al.'s (1998) investigation regarding immersion versus rote song instruction, favoring students who received song instruction through immersion ($g = 0.04$, 95% CI [-0.59, 0.67]) and Kramer's (1985) investigation regarding the effects of the Gould Specialized Music Program versus a traditional curriculum, favoring the traditional music instruction group's scores on both the pretests and the posttests ($g = -0.18$, 95% CI [-0.35, -0.01]).

Table 5

Overall Mean Effects, Within-Group Effects, and Between-Group Effects by Study

Citation	Year	<i>n</i>	Effects	ES			<i>SE_{ES}</i>	95 % CI	
				<i>g_{change}</i>	<i>g_{b/t}</i>	<i>M</i>		<i>LL</i>	<i>UL</i>
Jordan-DeCarbo	1982	89	20	0.80	0.36	0.71	0.04	0.63	0.79
Porter	1977	80	7	0.82	0.52	0.69	0.12	0.46	0.92
Persellin	2006	134	3	0.68	-	0.68	0.07	0.54	0.81
Hedden & Baker	2010	26	3	-	0.67	0.67	0.25	0.19	1.15
Miyamoto	2005	57	7	-	0.65	0.65	0.14	0.37	0.93
Apfelstadt	1984	61	6	-	0.61	0.61	0.19	0.25	0.98
Paney & Kay	2014	2021	14	0.58	-	0.58	0.01	0.56	0.59
Michaud	2014	66	10	0.49	0.86	0.54	0.05	0.45	0.63
Montgomery	1988	40	12	0.60	0.49	0.54	0.05	0.44	0.64
Roberts & Davies	1975	90	12	0.54	0.53	0.51	0.04	0.47	0.62
Jarjisian	1981	96	9	-	0.51	0.51	0.09	0.34	0.67
Richner	1976	77	37	0.51	0.37	0.46	0.04	0.38	0.55
Levinowitz	1987	102	12	-	0.47	0.40	0.14	0.18	0.74
Welch, Howard, & Rush	1989	32	9	0.51	0.23	0.40	0.11	0.18	0.62
Aaron	1990	69	13	0.42	0.31	0.40	0.05	0.30	0.49
Stauffer	1985	310	82	0.43	0.24	0.36	0.01	0.33	0.39
Kimble	1983	143	11	-	0.36	0.36	0.05	0.25	0.46
Kim	2000	78	18	0.60	0.16	0.33	0.04	0.24	0.42
Feierabend	1984	75	16	0.26	0.29	0.28	0.06	0.17	0.39
Mang	1997	83	12	-	0.27	0.27	0.07	0.12	0.42
Paladino	1991	171	6	0.22	0.20	0.22	0.05	0.12	0.32
Rutkowski	1986	162	23	0.23	0.13	0.19	0.02	0.15	0.23
Phillips & Aitchison	1997	269	4	-	0.18	0.18	0.06	0.06	0.30
Atterbury & Silcox	1993	205	2	0.18	-	0.18	0.05	0.08	0.29
McGraw	1996	66	9	0.15	0.23	0.18	0.08	0.03	0.33
Newell	2013	143	30	0.18	0.09	0.17	0.02	0.13	0.21
Guilbault	2009	136	11	-	0.16	0.16	0.07	0.04	0.29
Kramer	1985	201	9	0.19	-0.18	0.14	0.03	0.08	0.20
Marshall	2002	110	6	-	0.13	0.13	0.11	-0.09	0.35
Rutkowski & Snell Miller	2003	38	6	0.12	-	0.12	0.07	-0.02	0.25
Buck	1991	42	6	0.11	-	0.11	0.09	-0.29	0.29

(table continues)

Overall Mean Effects, Within-Group Effects, and Between-Group Effects by Study (continued)

Citation	Year	<i>n</i>	Effects	ES		<i>M</i>	<i>SE_{ES}</i>	95% CI	
				<i>g_{change}</i>	<i>g_{b/t}</i>			<i>LL</i>	<i>UL</i>
Kane	1994	87	4	0.04	0.11	0.05	0.07	-0.08	0.18
Klinger et al.	1998	39	1	-	0.04	0.04	0.32	-0.59	0.67
Rutkowski	1996	99	3	-0.03	0.54	0.03	0.07	-0.10	0.17
Within-group effect <i>g_{change}</i>		-	186			0.47	0.01	0.46	0.48
Between-group overall effect <i>g</i>		-	247			0.25	0.01	0.22	0.27
Overall Mean Effect Size		5497	433			0.43	0.01	0.42	0.44

Note. *n* = frequency of unique effect sizes per study.

A *Q* test for heterogeneity of effect sizes (Card, 2012) and an I^2 index of the magnitude of heterogeneity indicated that the study effect sizes were largely heterogeneous ($Q = 1039.90$, $df = 33$, $I^2 = 96.83\%$, $p < 0.0001$). Moderator variables selected a priori were evaluated to observe effect sizes by study characteristics. Refer to Table 6 for a summary of the overall mean effects, within-group effects, between-group effects, and heterogeneity statistics by moderator variable.

With regard to the second research question, moderator variable analyses within each of the eight primary moderator variables yielded overall effect sizes within each subgroup. For the eight primary moderator variables, effect sizes were examined based on calculations of Hedges' *g* between two independent group mean scores and pretest to posttest gain scores (*g_{change}*). In addition to the overall effect sizes, between-group effect sizes and pretest to posttest gain effect sizes were calculated for the primary moderator variables: (a) condition, (b) age, (c), gender, (d) measurement instrument, and (e) measured task. Between-group effect sizes and pretest to posttest gain effect sizes were omitted from secondary moderator variable analyses because the secondary moderators described the study demographics, not the instructional characteristics of the primary moderators.

Eight categorical moderator variables were analyzed for heterogeneity and effects. The magnitude of heterogeneity within categorical groups was medium to large as indicated by an I^2 range of 70.73%–80.31%, but should be considered with caution. An observation of similar effects within several moderator variables indicated that the variables might not have been heterogeneous. The heterogeneity might have been a consequence of including all unique effects instead of the number of studies within the degrees of freedom. The decision to include all unique effects seemed to have been supported by Card (2012). Additionally, several of the variables required only some effects from studies as opposed to all. Therefore, all moderator variables utilized unique effects as degrees of freedom for the purpose of continuity between variables.

The largest overall effect sizes across categorical variables included research design: pretest-posttest 1 group design ($g = 0.58$, 95% CI [0.56, 0.59]) and publication source: other ($g = 0.55$, 95% CI [0.53, 0.57]). The smallest overall effect sizes across categorical variables included teaching condition: literature dissemination ($g = 0.12$, 95% CI [0.09, 0.21]) and age: 11-Years ($g = 0.02$, 95% CI [-0.33, 0.38]).

The condition moderator variables included seven subgroups in which the population of studies were categorized: (a) feedback, (b) vocal model, (c) presence and type of singing instruction, (d) use of accompaniment, (e) pattern instruction, (f) performance context, and (g) literature dissemination. Overall mean effects by subgroup across the studies ranged from 0.56 (95% CI [0.54, 0.57]) to 0.12 (95% CI [0.09, 0.21]). The largest overall effect sizes across condition subgroups included feedback (5 studies, 41 unique effects, $g = 0.54$, 95% CI [0.52, 0.56]) and vocal model (4 studies, 37 unique effects, $g = 0.53$, 95% CI [0.47, 0.58]). The smallest overall effect sizes across condition subgroups included performance context (5 studies, 86

unique effects, $g = 0.19$, 95% CI [0.17, 0.22]) and literature dissemination (2 studies, 7 unique effects, $g = 0.12$, 95% CI [0.09, 0.21]).

Between-group effect sizes were calculated for the seven condition subgroups. Between-group mean effect sizes across the studies ranged from 0.45 (95% CI [0.36, 0.54]) to 0.12 (95% CI [-0.09, 0.33]). The largest overall between-group study effect sizes across condition subgroups included vocal model ($n = 4$, $g = 0.45$, 95% CI [0.36, 0.54]) and pattern instruction ($n = 3$, $g = 0.35$, 95% CI [0.25, 0.45]). The smallest overall between-group effect sizes across condition subgroups included performance context ($n = 5$, $g = 0.14$, 95% CI [0.09, 0.19]) and literature dissemination ($n = 2$, $g = 0.12$, 95% CI [-0.09, 0.33]).

Pretest to posttest gain effect sizes were calculated for the seven condition subgroups. Mean gain effect sizes across the studies ranged from 0.59 (95% CI [0.52, 0.66]) to 0.21 (95% CI [0.18, 0.24]). The largest overall pretest to posttest gain effect sizes across condition subgroups included vocal model ($n = 4$, $g_{\text{change}} = 0.59$, 95% CI [0.52, 0.66]) and feedback ($n = 5$, $g_{\text{change}} = 0.54$, 95% CI [0.52, 0.56]). The smallest overall pretest to posttest gain effect sizes across condition subgroups included performance context ($n = 5$, $g_{\text{change}} = 0.21$, 95% CI [0.18, 0.24]) and pattern instruction ($n = 3$, $g_{\text{change}} = 0.14$, 95% CI [0.03, 0.25]).

The age moderator variable included seven subgroups in which the studies were categorized: (a) 5-year-olds, (b) 6-year-olds, (c) 7-year-olds, (d) 8-year-olds, (e) 9-year-olds, (f) 10-year-olds, and (g) 11-year-olds. Overall mean effect sizes by subgroup across the studies ranged from 0.50 (95% CI [0.49, 0.52]) to 0.02 (95% CI [-0.33, 1.49]). The largest overall effect sizes across age subgroups included 5-year-olds (10 studies, 80 unique effects, $g = 0.50$, 95% CI [0.49, 0.52]) and 9-year-olds (3 studies, 7 unique effects, $g = 0.44$, 95% CI [0.33, 0.56]). The smallest overall effect sizes across age subgroups included 5-year-olds (10 studies, 80 unique

effects, $g = 0.31$, 95% CI [0.28, 0.34]) and 11-year-olds (1 study, 2 unique effects, $g = 0.02$, 95% CI [-0.33, 1.49]).

Between-group effect sizes were calculated for the seven age subgroups. Between group mean effect sizes across the studies ranged from 0.44 (95% CI [0.9, 0.79]) to 0.20 (95% CI [0.14, 0.28]). The largest overall between-group study effect sizes across age subgroups included 10-year-olds ($n = 2$, $g = 0.44$, 95% CI [0.9, 0.79]) and 7-year-olds ($n = 4$, $g = 0.27$, 95% CI [0.09, 0.45]). The smallest overall between-group effect sizes across age subgroups included 8-year-olds ($n = 9$, $g = 0.21$, 95% CI [0.14, 0.28]) and 5-year-olds ($n = 10$, $g = 0.20$, 95% CI [0.15, 0.25, 0.33]).

Pretest to posttest gain effect sizes were calculated for the seven age subgroups. Mean gain effect sizes across the studies ranged from 0.45 (95% CI [0.39, 0.51]) to 0.02 (95% CI [-0.33, 0.38]). The largest overall pretest to posttest gain effect sizes across age subgroups included 9-year-olds ($n = 3$, $g_{\text{change}} = 0.45$, 95% CI [0.39, 0.51]) and 10-year-olds ($n = 2$, $g_{\text{change}} = 0.45$, 95% CI [0.39, 0.51]). The smallest overall pretest to posttest gain effect sizes across age subgroups included 5-year-olds ($n = 10$, $g_{\text{change}} = 0.36$, 95% CI [0.33, 0.39]) and 11-year-olds ($n = 1$, $g_{\text{change}} = 0.02$, 95% CI [-0.33, 0.38]).

The gender moderator variable included three subgroups in which the population of studies were categorized: (a) girls, (b) boys, and (c) between gender, favoring girls. The largest overall effect size across gender subgroups was girls (5 studies, 9 unique effects, $g = 0.55$, 95% CI [0.51, 0.59]), and the smallest overall effect size across gender subgroups was between genders, favoring girls (6 studies, 20 unique effects, $g = 0.45$, 95% CI [0.34, 0.55]). The largest overall between-group study effect size across gender subgroups was girls ($n = 5$, $g = 0.50$, 95% CI [0.21, 0.70]), and the smallest overall between-group effect size across gender subgroups was

boys ($n = 5$, $g = 0.35$, 95% CI [0.00, 0.70]). The largest overall pretest to posttest gain effect sizes across gender subgroups was girls ($n = 5$, $g_{\text{change}} = 0.55$, 95% CI [0.53, 0.47]) and the smallest overall pretest to posttest gain effect size across gender subgroups was boys ($n = 5$, $g_{\text{change}} = 0.52$, 95% CI [0.50, 0.54]).

The measurement instrument moderator variables included six subgroups in which the population of studies were categorized: (a) Boardman, (b) acoustic measures of pitch accuracy, (c) researcher as acoustic measurer of pitch accuracy, (d) researcher as acoustic measurer of song accuracy, (e) researcher as acoustic measurer of pattern and interval accuracy, and (f) Singing Voice Development Measure (SVDM). Overall mean effects by subgroup across the studies ranged from 0.54 (95% CI [0.45, 0.64]) to 0.19 (95% CI [0.17, 0.22]). The largest overall effect sizes across measurement instrument subgroups included Boardman (5 studies, 15 unique effects, $g = 0.54$, 95% CI [0.45, 0.64]) and acoustic measures of pitch accuracy (9 studies, 112 unique effects, $g = 0.52$, 95% CI [0.50, 0.53]). The smallest overall effect sizes across measurement instrument subgroups included researcher as acoustic measurer of pattern and interval accuracy (8 studies, 117 unique effects, $g = 0.34$, 95% CI [0.31, 0.37]) and Singing Voice Development Measure (7 studies, 63 unique effects, $g = 0.19$, 95% CI [0.17, 0.22]).

Between-group effects were calculated for the six measurement instrument subgroups. Between-group mean effect sizes across the studies ranged from 0.50 (95% CI [0.37, 0.62]) to 0.13 (95% CI [0.08, 0.18]). The largest overall between-group study effect sizes across measurement instrument subgroups included Boardman ($n = 2$, $g = 0.50$, 95% CI [0.37, 0.63]) and researcher as acoustic measurer of pitch accuracy ($n = 9$, $g = 0.29$, 95% CI [0.24, 0.34]). The smallest overall between-group effect sizes across measurement instrument subgroups included

acoustic measures of pitch accuracy ($n = 9$, $g = 0.22$, 95% CI [0.15, 0.29]) and Singing Voice Development Measure ($n = 7$, $g = 0.13$, 95% CI [0.08, 0.18]).

Pretest to posttest gain effects were calculated for the six measurement instrument subgroups. Mean gain effect sizes across the studies ranged from 0.60 (95% CI [0.45, 0.75]) to 0.22 (95% CI [0.19, 0.25]). The largest overall pretest to posttest gain effect sizes across measurement instrument subgroups included Boardman ($n = 2$, $g_{\text{change}} = 0.60$, 95% CI [0.45, 0.75]) and acoustic measures of pitch accuracy ($n = 9$, $g_{\text{change}} = 0.51$, 95% CI [0.50, 0.52]). The smallest overall pretest to posttest gain effect sizes across measurement instrument subgroups included researcher as acoustic measurer of pattern and interval accuracy ($n = 8$, $g_{\text{change}} = 0.40$, 95% CI [0.36, 0.44]) and Singing Voice Development Measure ($n = 7$, $g_{\text{change}} = 0.22$, 95% CI [0.19, 0.25]).

The measured task moderator variable included three subgroups in which the population of studies was categorized: (a) song accuracy, (b) pitch accuracy, and (c) pattern accuracy. The largest overall effect size across measured task subgroups was song accuracy (5 studies, 36 unique effects, $g = 0.49$, 95% CI [0.43, 0.54]), and the smallest overall effect size across measured task subgroups was pattern accuracy (10 studies, 126 unique effects, $g = 0.34$, 95% CI [0.31, 0.37]). The largest overall between-group study effect size across measured task subgroups was song accuracy ($n = 5$, $g = 0.28$, 95% CI [0.19, 0.37]), and the smallest overall between-group effect size across measured task subgroups was pitch accuracy ($n = 23$, $g = 0.22$, 95% CI [0.19, 0.24]). The largest overall pretest to posttest gain effect size across measured task subgroups was song accuracy ($n = 5$, $g_{\text{change}} = 0.61$, 95% CI [0.54, 0.68]), and the smallest overall pretest to posttest gain effect size across measured task subgroups was pattern accuracy ($n = 10$, $g_{\text{change}} = 0.40$, 95% CI [0.38, 0.42]).

With regard to the third research question, a secondary moderator variable analysis within each of the three categorical moderator variables yielded overall effect sizes within each categorical subgroup: (a) publication source, (b) population, and (c) research design. A secondary moderator variable analysis within each of the two continuous moderator variables yielded overall correlation coefficients and T Test statistics within each continuous subgroup: (a) publication year and (b) treatment period. Between-group effect sizes and within-group effect sizes were omitted from secondary moderator variable analyses as the secondary moderators described the study demographics as opposed to the instructional characteristics of the primary moderators.

The publication source moderator variable included four subgroups in which the population of studies were categorized: (a) Journal of Research in Music Education (JRME), (b) The Bulletin of the Council for Research in Music Education (CRME), (c) doctoral dissertations, and (d) other (Canadian Journal of Research in Music Education, Psychology of Music, Research Issues in Music Education, and Update: Applications of Research in Music Education). Overall mean effects by subgroup across the studies ranged from 0.55 (95% CI [0.53, 0.57]) to 0.28 (95% CI [0.26, 0.30]). The largest overall effect sizes across publication source subgroups included other (4 studies, 42 unique effects, $g = 0.55$, 95% CI [0.53, 0.57]) and JRME (9 studies, 66 unique effects, $g = 0.41$, 95% CI [0.37, 0.45]). The smallest overall effect sizes across publication source subgroups included CRME (3 studies, 12 unique effects, $g = 0.41$, 95% CI [0.32, 0.50]) and doctoral dissertations (18 studies, 312 unique effects, $g = 0.28$, 95% CI [0.26, 0.30]).

The population moderator variable included six subgroups in which the population of studies were categorized: (a) Northeast, (b) Southeast, (c) Southwest, (d) West, (e) Midwest, and

(f) other (Hawaii, England, and Canada). Overall mean effects by subgroup across the studies ranged from 0.52 (95% CI [0.50, 0.53]) to 0.20 (95% CI [0.01, 0.18]). The largest overall effect sizes across the population subgroups included the Midwest (9 studies, 159 unique effects, $g = 0.52$, 95% CI [0.50, 0.53]) and other (4 studies, 40 unique effects, $g = 0.49$, 95% CI [0.42, 0.55]). The smallest overall effect sizes across the population subgroups included the West (3 studies, 44 unique effects, $g = 0.39$, 95% CI [0.31, 0.47]) and the Northeast (13 studies, 137 unique effects, $g = 0.20$, 95% CI [0.18, 0.22]).

The research design moderator variable included six subgroups in which the population of studies were categorized: (a) pretest-posttest control group design, (b) multiple treatment pretest-posttest control group design, (c) multiple treatment pretest-posttest design, (d) posttest-only control group design, (e) multiple treatment posttest-only control group design, and (f) pretest-posttest 1 group design. Overall mean effects by subgroups across the studies ranged from 0.58 (95% CI [0.56, 0.59]) to 0.14 (95% CI [0.06, 0.21]). The largest overall effect sizes across the research design subgroups included pretest-posttest 1 group design (11 studies, 14 unique effects, $g = 0.58$, 95% CI [0.56, 0.59]) and multiple treatment pretest-posttest design (9 studies, 131 unique effects, $g = 0.30$, 95% CI [0.36, 0.41]). The smallest overall effect sizes across research design subgroups included pretest-posttest control group design (7 studies, 101 unique effects, $g = 0.25$, 95% CI [0.22, 0.28]) and posttest-only control group design (1 studies, 19 unique effects, $g = 0.14$, 95% CI [0.06, 0.21]).

Categorical moderator variables and continuous moderator variables are generally treated differently. The two continuous moderator variables, publication year and treatment period, were analyzed to determine the relationships with the overall study weighted mean effect sizes.

Publication years across the studies ranged from 1977 to 2014 ($N = 34$, $SD = 11.14$, $Mdn = 1992$,

Mode = 1991), and treatment periods across the studies ranged from one week to 36 weeks ($N = 34$, $M = 14.59$, $SD = 10.01$, $Mdn = 12$, Mode = 10). A Pearson r correlation coefficient yielded a weak, negative relationship between publication years and corresponding effect sizes ($r = -0.13$). A Pearson r correlation coefficient also yielded a weak, negative relationship between treatment periods and corresponding effect sizes ($r = -0.13$). T Tests results indicated that there were no significant relationships between effect sizes and variables ($T = -0.74$, $p = 0.23$).

Publication Bias

There may have been evidence from publication source subgroups indicating that the results were impacted by publication bias within the current meta-analysis. Unpublished studies ($g = 0.28$, 95% CI [0.26, 0.30]) had lower effect sizes than published studies ($g = 0.53$, 95% CI [0.52, 0.54]). Card (2012) defined publication bias as:

The possibility that studies finding null (absence of statistically significant effect) or negative (statistically significant effect in opposite direction) results are less likely to be published than studies finding positive effects (statistically significant effects in expected direction). (p. 257)

Borenstein (2009) stated that evidence exists demonstrating a trend toward publication bias in both meta-analyses and reviews of literature due to higher publication acceptance rates of studies that include larger effects.

Within the current study, effects across publication source yielded a larger effect for studies in the other category ($g = 0.55$, 95% CI [0.53, 0.57]) than the subsequent three subgroups. Studies within Journal of Research in Music Education ($g = 0.41$, 95% CI [0.37, 0.45]) and the Bulletin of the Council for Research in Music Education ($g = 0.41$, 95% CI [0.32,

0.50]) yielded identical effect sizes and were both larger than the overall doctoral dissertation effects ($g = 0.28$, 95% CI [0.27, 0.30]).

Table 6

Overall Mean Effects, Within-Group Effects, Between-Group Effects, and Heterogeneity Statistics by Moderator Variable

Group	<i>n</i>	<i>g</i>	95 % CI _{b/t}		<i>g</i> _{change}	95% CI _{change}		<i>df</i>	<i>M</i>		95% CI		Heterogeneity Statistics			
			<i>LL</i>	<i>UL</i>		<i>LL</i>	<i>UL</i>		ES	<i>SE</i> _{ES}	<i>LL</i>	<i>UL</i>	<i>Q</i> _{win}	<i>Q</i> _{b/t}	<i>Q</i> _{total}	<i>I</i> ²
<i>Teaching Condition</i>																
Feedback	5	0.21	0.01	0.41	0.54	0.52	0.56	40	0.54	0.01	0.52	0.56	370.25	1393.19		
Vocal model	4	0.45	0.36	0.54	0.59	0.52	0.66	36	0.53	0.03	0.47	0.58	93.97	1669.48		
Presence/type of singing instruction	11	0.31	0.26	0.36	0.44	0.40	0.48	134	0.40	0.02	0.36	0.43	558.41	1205.43		
Use of accompaniment	4	0.24	0.20	0.28	0.40	0.37	0.43	97	0.34	0.01	0.31	0.37	185.19	1578.25		
Pattern instruction	3	0.35	0.25	0.45	0.14	0.03	0.25	28	0.25	0.04	0.18	0.33	48.14	1715.31		
Performance context	5	0.14	0.09	0.19	0.21	0.18	0.24	85	0.19	0.01	0.17	0.22	341.21	1422.23		
Literature dissemination	2	0.12	-0.09	0.33	-	-	-	6	0.12	0.11	0.09	0.21	0.82	1762.63		
Overall	34							426					1597.98	165.46	1763.44	72.84%
<i>Age</i>																
5-years	10	0.20	0.15	0.25	0.36	0.33	0.39	79	0.31	0.01	0.28	0.34	404.34	1359.11		
6-years	9	0.26	0.20	0.33	0.40	0.35	0.45	82	0.34	0.02	0.30	0.38	229.75	1533.69		
7-years	4	0.27	0.09	0.45	0.38	0.34	0.42	31	0.36	0.04	0.29	0.43	139.10	1624.35		
8-years	9	0.21	0.14	0.28	0.51	0.05	0.52	81	0.50	0.01	0.49	0.52	326.29	1437.15		
9-years	3	-	-	-	0.45	0.39	0.51	6	0.44	0.06	0.33	0.56	8.54	1754.91		
10-years	2	0.44	0.9	0.79	0.45	0.39	0.51	4	0.38	0.12	0.15	0.61	9.96	1753.49		
11-years	1	-	-	-	0.02	-0.33	0.38	1	0.02	0.18	-0.33	1.49	1.49	1761.95		
Overall	34							284					1119.46	643.99	1763.44	75.17%
<i>Gender</i>																
Girls	5	0.50	0.21	0.79	0.55	0.53	0.57	8	0.55	0.02	0.51	0.59	63.09	1700.36		

(table continues)

Overall Mean Effects, Within-Group Effects, Between-Group Effects, and Heterogeneity Statistics by Moderator Variable (continued)

Group	<i>n</i>	<i>g</i>	95% CI _{b/t}		<i>g_{change}</i>	95% CI _{change}		<i>df</i>	<i>M</i>		95% CI		Heterogeneity Statistics			
			<i>LL</i>	<i>UL</i>		<i>LL</i>	<i>UL</i>		<i>ES</i>	<i>SE_{ES}</i>	<i>LL</i>	<i>UL</i>	<i>Q_{w/in}</i>	<i>Q_{b/t}</i>	<i>Q_{total}</i>	<i>I</i> ²
<i>Gender</i>																
Boys	5	0.35	0.00	0.70	0.52	0.50	0.54	8	0.51	0.02	0.47	0.56	112.06	1651.38		
Between genders, , favoring girls	6	0.45	0.35	0.55	-	-	-	19	0.45	0.05	0.34	0.55	22.89	1740.55		
Overall	8							35					198.04	1565.40	1763.44	80.31%
<i>Measurement Instrument</i>																
Boardman	2	0.50	0.37	0.63	0.60	0.45	0.75	14	0.54	0.05	0.45	0.64	16.11	1747.33		
Acoustic measures/pitch accuracy	9	0.22	0.15	0.29	0.51	0.50	0.52	111	0.52	0.01	0.50	0.53	692.36	1071.08		
Researcher as acoustic/pitch accuracy	9	0.29	0.24	0.34	0.49	0.45	0.53	92	0.41	0.02	0.38	0.44	326.65	1436.79		
Researcher as acoustic/song accuracy	5	0.25	0.19	0.31	0.47	0.37	0.57	26	0.38	0.04	0.31	0.46	102.92	1660.53		
Researcher as acoustic/pattern and interval accuracy	8	0.27	0.22	0.32	0.40	0.36	0.44	116	0.34	0.02	0.31	0.37	198.65	1564.79		
SVDM	7	0.13	0.08	0.18	0.22	0.19	0.25	62	0.19	0.01	0.17	0.22	290.35	1473.09		
Overall	34							421					1627.05	136.30	1763.44	73.33%
<i>Measured Task</i>																
Song accuracy	5	0.28	0.19	0.37	0.61	0.54	0.68	35	0.49	0.03	0.43	0.54	167.32	1596.12		
Pitch accuracy	23	0.22	0.19	0.25	0.47	0.46	0.48	270	0.44	0.01	0.43	0.45	1351.42	412.02		
Pattern accuracy	10	0.27	0.22	0.32	0.40	0.38	0.42	125	0.34	0.02	0.31	0.37	205.35	1558.10		
Overall	34							430					1724.09	39.3577	1763.44	74.83%

(table continues)

Overall Mean Effects, Within-Group Effects, Between-Group Effects, and Heterogeneity Statistics by Moderator Variable (continued)

Group	<i>N</i>	<i>df</i>	ES	<i>SE</i> _{ES}	95% CI		Heterogeneity Statistics			
					<i>LL</i>	<i>UL</i>	<i>Q</i> _{within}	<i>Q</i> _{between}	<i>Q</i> _{total}	<i>I</i> ²
<i>Publication Source</i>										
Other	4	41	0.55	0.01	0.53	0.57	131.34	1632.10		
Journal of Research in Music Education	9	65	0.41	0.02	0.37	0.45	309.99	1453.46		
The Bulletin of the Council for Research in Music Education	3	11	0.41	0.05	0.32	0.50	67.16	1696.29		
Doctoral dissertations	18	312	0.28	0.01	0.27	0.30	1000.34	763.10		
Overall	34	429					1508.83	254.61	1763.44	71.24%
<i>Population</i>										
Midwest	9	158	0.52	0.01	0.50	0.53	578.33	1185.11		
Other: Hawaii, England, Canada	4	39	0.49	0.03	0.42	0.55	83.33	1680.11		
Southwest	2	11	0.46	0.05	0.36	0.56	68.78	1694.66		
Southeast	3	40	0.40	0.03	0.34	0.46	151.09	1612.36		
West	3	43	0.39	0.04	0.31	0.47	95.61	1667.83		
Northeast	13	136	0.20	0.01	0.18	0.22	551.09	1212.37		
Overall	34	427					1528.22	235.23	1763.44	71.60%
<i>Research Design</i>										
One-group pretest-posttest design	1	13	0.58	0.01	0.56	0.59	87.90	1675.55		
Multiple treatment pretest-posttest design	3	130	0.38	0.01	0.36	0.41	238.32	1525.12		
Multiple treatment posttest-only control group design	7	53	0.35	0.03	0.28	0.42	81.19	1682.26		
Multiple treatment pretest-posttest control group design	9	113	0.31	0.01	0.28	0.34	589.10	1174.34		
Pretest-posttest control group design	11	100	0.25	0.01	0.22	0.28	457.30	1306.14		
Posttest-only control group design	3	18	0.14	0.04	0.06	0.21	31.13	1732.31		
Overall	34	427					1482.95	280.50	1763.44	70.73%

Note. n = frequency of analyzed studies

CHAPTER V

CONCLUSIONS

The fifth chapter includes the following: (a) a summary of the results, (b) a discussion of the results, their interpretations in light of previous research where available, and implications for teaching, (d) a review of publication bias, (e) a review of study quality, (f) limitations of the current meta-analysis, (f) implications for research pedagogy, and (g) suggestions for future research. Primary moderator variables, measurement instrument and instrument task, are discussed with regard to study quality. The secondary moderator variables, population and research design, are also discussed with regard to study quality. The secondary moderator variable, population source, is discussed with regard to publication bias.

The purpose of the study was to address the varied and somewhat stratified study results within the area of singing ability and instruction by statistically summarizing the data of related studies through meta-analysis. Toward this purpose, three meta-analysis research questions were developed. These included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children ages 5 through 11 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task; and (3) what are the statistical effects of study characteristics within and across the following secondary moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period. The magnitude of the results is discussed in terms of the following effect size ranges: trivial = < 0.20 (Ellis, 2010), small = 0.20–0.49, medium = 0.50–0.79, large = > 0.80 (Cohen, 1988).

Summary of Results

To answer the first research question, the current meta-analysis yielded a small overall mean effect size for instruction across 34 studies and 433 unique effect sizes ($\overline{ES} = 0.43$, $N = 5,497$). Mean effect sizes ranged from trivial ($g = 0.03$) to medium ($g = 0.71$). Study effect sizes were largely heterogeneous ($I^2 = 96.83\%$).

To answer the second research question, overall mean effects by primary moderator variable ranged from trivial ($g = 0.02$) to moderate ($g = 0.56$). Regarding teaching condition, effect sizes ranged from trivial ($g = 0.12$) to moderate ($g = 0.56$), with feedback yielding the largest effect. Regarding age, effect sizes ranged from trivial ($g = 0.02$) to moderate ($g = 0.50$), with 8-years yielding the largest effect. Regarding gender, effect sizes ranged from small ($g = 0.45$) to moderate ($g = 0.55$), with girls yielding the largest effect. Regarding measurement instrument, effect sizes ranged from trivial ($g = 0.19$) to moderate ($g = 0.54$), with Boardman yielding the largest effect. Regarding measured task, effect sizes were all within the small range ($g = 0.34, 0.49$), with song accuracy yielding the largest effect.

To answer the third research question, overall mean effects by secondary moderator variable ranged from trivial ($g = 0.14$) to medium ($g = 0.58$). Regarding publication source, effect sizes ranged from small ($g = 0.28$) to medium ($g = 0.55$), with other yielding the largest effect. Regarding population, effect sizes ranged from small ($g = 0.20$) to medium ($g = 0.52$), with Midwest yielding the largest effect. Regarding research design, effect sizes ranged from trivial ($g = 0.14$) to medium ($g = 0.58$), with pretest-posttest 1 group design yielding the largest effect.

Effect sizes within the current meta-analysis may suggest implications for (a) teaching, (b) research pedagogy, and (c) research within music education. Regarding teaching and singing

ability development, suggestions can be made for the specific population of studies included in the present meta-analysis. Regarding research pedagogy, suggestions can be made implying an ideal skill set conducive to the creation of quality studies. Regarding research within music education, more original research studies, replication studies, and meta-analyses are needed on the topic of singing ability and its many facets.

Discussion

Research Question One: Overall Effects

What is the overall statistical effect of instructional characteristics on singing ability in children from ages 5 through 11 years old? Overall, instructional interventions within the current meta-analysis were deemed as an effective means in the development of singing ability with young children. The overall differences for between- and within-group mean scores that were attributed to singing instruction were nearly one-half of a standard deviation greater than other treatment, control, or pretest scores. The within-group effect size was greater than the between-group effect size and the overall effect size. The between-group effect was a quarter of a standard deviation smaller than the within-group effect.

With regard to other music-education-specific meta-analyses, music instruction has been documented as an overall effective means of improving sight-singing scores, music reading scores, and serving as a “contingency for education and therapeutic objectives” (Standley, 1996, p. 105). Mishra (2013) found a small overall effect for instructional interventions on sight-reading mean scores and a “moderately strong” effect for pretest-posttest gain scores. Conclusions of the meta-analysis suggested that sight-reading instructional interventions contributed to increased mean scores. Standley (2008) found a small overall effect for music

interventions on reading mean scores and suggested that it was a moderately strong and significant effect due to the confidence intervals. Standley (2008) also mentioned that,

Because the overall 95% confidence interval did not include 0, this effect size is considered statistically significant and indicates that music interventions generally had a positive and significant effect on the teaching of reading skills. This is the most important finding in the study, but a very general statement. (p. 26)

The results of the current meta-analysis generally fit among results from the other studies. This is a likely expectation given that most of the studies within the other meta-analyses were also within public school settings with public school setting-specific research designs as well as similar sample sizes. The only exception can be found in Standley's (1996) meta-analyzed results that included case studies; effect sizes were expected to be much higher when instruction was directed to specific participant needs. Standley (1996) stated that,

One reason for different effect sizes by design type may be that the studies using subjects as their own control were often single-subject or small-sample studies with contingencies carefully tailored to individual preferences and reinforcement schedules. Such highly individualized techniques are a common aspect of behavioral practices and are designed to enhance effectiveness. (pp. 124–125)

Comparing the current study with other music-education-specific meta-analyses provided an external context that enabled the following conclusion within the current meta-analysis: a small effect size has been regarded as effective in past research. Therefore, instructional interventions may be effective to the development of singing ability with young children. Larger sample sizes and attention to a priori power analyses may benefit singing ability research and researchers, specifically regarding the detection of differences within and between groups.

It may also be important to explore the effects of instructional interventions across studies within the current meta-analysis. The overall small effect size representative of the current meta-analysis is an estimate of the magnitude between and within-group mean scores. Discussing the overall effect with regard to outside meta-analyses is relevant, because the other meta-analyses also pooled within- and between-group effects. Exploring within-group effects detached from between-group effects may, however, may yield a more meaningful discussion of effects within the current study than pooling within- and between-group effects.

Pretest to posttest group analyses yielded an overall small effect representative of unique effects that ranged from negative to large. There was an effect size difference of one-third of a standard deviation between control group pretest to posttest effects and treatment pretest to posttest effects. Between-group analyses also yielded an overall small effect representative of unique effects that ranged from negative to large. Although the ranges may be similar, it may be important to observe how the effects by studies differ across within-group effects and between-group effects. Referring back to Table 5 in Chapter IV where within-group and between-group effects were calculated for both groups, within-group effects were consistently greater than approximately three-quarters of the studies.

In conclusion, singing instruction may be seen as an overall effective means of improving singing ability scores in children between the ages of 5 and 11, specific to the studies and participants represented within the current meta-analysis. The importance of within-group and between-group effects in relation to one another has been discussed with regard to assessment in the music classroom. The difference between treatment group gain scores over control group gain scores may indicate that singing instruction is more effective than (a) what has been commonly referred to as a traditional curriculum, or one which includes song-singing without

vocal development guidance, and (b) no music instruction at all. The importance of within-group and between-group effect sizes is explored further through moderator variables.

Research Question Two: Primary Moderator Variables

What are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age, (c) gender, (d) measurement instrument, and (e) measured task? Moderator variables, measurement instrument and measured task, are discussed with regard to study quality. The current meta-analysis was largely heterogeneous. Heterogeneity implies that effects within moderator variables account for variation among overall effects of the meta-analysis. The moderator variables included within the current study were chosen from literature reviews regarding the topic of singing ability.

Teaching Condition

Overall, instructional conditions yielded positive, trivial to moderate effects and, for all but one subgroup, positive confidence intervals. The moderator variable, teaching condition, was comprised of studies sub-grouped by instructional similarities. Subgroups included instructional strategies regarding (a) feedback, (b) vocal model, (c) presence and type of singing instruction, (d) use of accompaniment, (e) pattern instruction, (f) performance context, and (g) literature dissemination. Observations and interpretations may be made regarding overall effects and effects within each subgroup.

Feedback yielded a medium effect, which was the largest effect size across all conditions. Vocal model also yielded a medium effect and was nearly as large as feedback by one-tenth of a standard deviation, but had the largest within-group and between-group effect sizes across all conditions. Performance context and literature dissemination were the only two conditions that yielded trivial effects. All other conditions yielded medium effects.

Feedback.

Feedback was an overall effective means of improving singing ability scores across studies. Within the context of the current meta-analysis, feedback yielded the largest effect size of all teaching conditions. Several studies utilized feedback as a condition through computer programs. There was a greater effect for practicing singing accuracy with computer programs than receiving traditional instruction in third grade. Practicing fewer times (between one and five) with computer programs yielded a greater effect on mean scores than practicing more than five times with the computer programs. Additionally, teacher feedback combined with computer feedback had a greater effect than computer feedback alone. Specific feedback was also found to be effective without the use of a computer, especially when compared to generic, unspecific feedback.

Implications for teaching may be made with regard to feedback. Children are living in an age filled with technology and information. Children have access to information beyond that which was available when their teachers and parents were young. If children have access to information outside of the music classroom, perhaps there would be a benefit to having access to information within the music classroom. Information within the music classroom regarding the development of singing ability can be disseminated from teacher to student and from technology to student.

Public school districts often encourage teachers to embrace technology through interactive white-boards and iPads, to name a few. Observations made at various conferences and workshops have demonstrated that teachers are consistently in search of technology that will be efficient and effective within the music classroom. Ideally, technology would enhance a lesson rather than replace or take away from it. Investing in computer software that can provide

specific feedback to students, provide independent practice, and provide information to both the teacher and the student may be a useful instructional addition.

Vocal model.

The type of vocal model influenced the improvement of singing ability scores across studies. Differences between vocal models yielded a medium overall and within-group effect size and a small effect size for between-group scores. Within the context of the current meta-analysis, the effect of a vocal model was nearly as large as the effect of feedback. Research regarding vocal models was represented within the current meta-analysis by comparisons of male models, female models, and whether the model sang for or with the students. Within the current meta-analysis, a female model yielded a greater effect on mean scores than a male model. The effect within the current meta-analysis aligns with previous results (Goetze et al., 1990; Hermanson, 1971; McCachern, 1983; Simms, Moore, & Kuhn, 1982; Small & Green, 1987; Yarbrough, Green, Benson, & Bowers, 1991). In male models, singing with either a comfortable range or falsetto voice both yielded similar effects on mean scores, although a comfortable range yielded higher mean scores than singing with a falsetto voice.

Whether or not a teacher sang for students, with students, or for and with students may make a difference regarding the development of singing ability. Although groups yielded similar effect sizes within the current meta-analysis, a combination of singing for and with students yielded a greater effect on mean scores than singing only with students. Additionally, singing only with students had a greater effect than singing only for students. These findings were based on one study; replications of the study are needed to strengthen the evidence of the finding given that they are contradictory to common pedagogical recommendations.

Implications for teaching may be made with regard to vocal models. Overall, students may benefit more from a female vocal model than a male vocal model and more from a male model singing in his natural range than singing in falsetto. Singing abilities, however, should develop with proper instruction regardless of whether the vocal model is a male or female. Green (1990) suggested that neither adult model was as effective as a child vocal model. Regardless of model, consequences of quality instruction may surpass consequences of vocal modeling. Perhaps there is a need to encourage alternate views of instruction including gender-specific recommendations and gender-neutral recommendations.

Presence and type of singing instruction.

The presence and type of singing instruction yielded a small effect size for overall, within-, and between-group scores. Pretest to posttest scores yielded the largest effects for instruction that included discrimination training. Discrimination training included an element of differentiating between tones that were the same or different. Vocal coordination instruction yielded a medium effect for vocal development in elementary music and a small effect regarding range and higher pitches. This finding aligns with previous research (Phillips, 1983). Studies including remedial training as a treatment yielded small to medium effects. This finding also aligns with previous research (Gaiser, 1961).

Implications for teaching may be made with regard to the presence and type of singing instruction. Students may benefit from increased small-group work, if time and schedules allow. Elementary music specialists are often encouraged to lead choir programs at the elementary level in addition to teaching classroom music. Although rehearsals could be built into the school day, most are after or before school and exist as an extracurricular activity. Having a remedial choir in

addition to the full choir may benefit vocal development for children needing extra attention in the upper levels. Doing so may outweigh the cost of time required of the teacher.

Centers built into a lesson plan may provide teachers with opportunities for small-group work within the school day. Not all students, and not all weak singers, are interested in devoting extra time outside of the school day for choir. Perhaps utilizing centers that provided time for increased small-group work would benefit vocal development, especially if the specific goal of the centers was to build vocal development.

Use of accompaniment.

The use of accompaniment yielded a small effect size for overall, within-, and between-group scores. Whether or not piano accompaniment influences singing ability was investigated across four studies in the current meta-analysis. Gain scores were similar for groups who received instruction with and without accompaniment, although a slightly greater effect was found for groups without accompaniment overall. Previous research aligns with this finding (Guilbault, 2004). Recommendations for teachers include limiting the use of accompaniment until vocal accuracy is consistently high for most students.

Pattern instruction.

Pattern instruction yielded a small effect size overall and for between-group scores, but a trivial effect size for within-group scores. Pattern instruction has been used to provide children aural practice with the tonal vocabulary found in children's literature while focusing on two to three pitches at a time. Although pattern instruction yielded a positive effect, pattern instruction did not yield greater effect sizes for treatment conditions with pattern instruction than a music curriculum without pattern instruction. There may be, however, differences regarding which

patterns were most effective. A combination of pentatonic and diatonic patterns yielded a greater effect size than pattern instruction limited to diatonic patterns or limited to pentatonic patterns.

Implications for teaching may be made with regard to pattern instruction. Pattern instruction can be a valuable tool for teachers who enjoy pattern instruction and who develop the skills to implement pattern instruction effectively. A combination of pentatonic and diatonic patterns may be preferable to pentatonic or diatonic alone. Therefore, it may be beneficial for the literature used to also include pentatonic and diatonic patterns rather than be limited to all diatonic songs or all pentatonic songs.

Performance context.

Performance context yielded a trivial effect size overall, and for between-group scores, as well as a small effect size for within-group scores. Within the current meta-analysis, Performance context included studies where how (or what) children sang was an instructional condition. conditions included individual versus group singing, response modes, movement while singing, and song range. There was a greater effect size for instruction that included individual and small-group singing than instruction limited to classroom singing. This finding aligns with previous research (Goetze, 1985; Goetze & Horii, 1989), but also conflicts with other previous research (Cooper, 1995; Green, 1994). The use of songs with words in an elementary curriculum had the greatest effect for low-scoring kindergarten and first grade students when compared to singing most songs without words. The use of movement in a music curriculum had a greater effect on accuracy scores than without movement, and whether the movement was fluid or beat-centered yielded similar results. This finding aligns with previous research where 5- and 6-year-olds demonstrated improved vocal accuracy when associating movement with pitch levels (Liao, 2008). Regarding range, the song range in which literature was disseminated had a greater

effect, favoring literature of a restricted singing range (from C3–B3) over literature of a higher singing range representative of music textbook literature. This finding aligns with previous research (Moore et al., 1995/1996; Phillips, 2014).

Implications for teaching regarding performance context are challenging, because there were too few performance context studies represented to make conclusive recommendations. That in itself may imply more exploration is needed regarding the effect of performance context on vocal development, specifically in terms of pretest to posttest research designs.

Literature dissemination.

Literature dissemination yielded a trivial effect size overall, and for between-group scores. Findings should be interpreted with caution regarding the literature dissemination subgroup. Two studies constituted the subgroup, and neither study included a pretest. There was a trivial effect found between groups, favoring immersion. Within the current meta-analysis, there was little difference between accuracy scores based on how the songs were taught. This finding aligns with previous research (Gault, 2002). Another previous study, however, suggested that rote song instruction was significantly more effective than immersion (Apfelstadt, 1984). Differences across research designs and the specific songs used for the studies may attribute to conflicting results. Replications of the study (Klinger et al., 1998) as well as additional studies are needed to gain further context of the findings given that they are contradictory to common pedagogical recommendations that favor rote song instruction.

Age

Age yielded overall effect sizes that ranged from trivial to medium. The 8-year-old subgroup yielded a medium effect size, which was the largest effect across the group. The 11-year-old group yielded a trivial effect, which was the smallest effect across the group. All other

effect sizes were small. Precautions must be made regarding the overall effects, as fewer studies were available for the older age groups than the younger age groups.

Regarding the current meta-analysis, singing instruction had the greatest overall effect on group scores for 8-year-old (third grade) children. After 8-year-old children, the second largest and subsequent effect sizes were with 9-year-olds, 10-year-olds, 7-year-olds, and 6-year-olds. Singing instruction had the lowest effect for 5-year-old (Kindergarten) children and 11-year-old children. This finding aligns with past research (Bentley, 1969; Gould, 1969). Children enter the public school system with varying background experiences (Welch, 2015) and generally spend much of Kindergarten exploring vocal timbres and foundational vocal development. The current meta-analysis also demonstrated a trend that aligns with previous research: accuracy tends to improve across the ages of 5, 6, 7, and 8. Accuracy then decreases between ages 8 and 9, 10, and 11. Mizener (1993) found a similar decline in accuracy scores with upper singing ranges. In contrast, though, some studies have found that children tended to improve with age (Geringer, 1983; Rutkowski, 1990).

Implications for teaching may be made with regard to age. Children in Kindergarten through third grade may benefit from an elementary music program that prioritizes vocal development. If there is a social shift and decrease in accuracy scores after third grade, great care must be taken before third grade to build a respectful, safe classroom environment to negate or cushion the shift. Teachers may be doing so regardless, but doing so with the intention of vocal development in mind may offer a new perspective on building trust and skills in an elementary music program.

Gender

Within the current meta-analysis, singing instruction was approximately as effective for boys as it was for girls. Instruction had the greatest effect on girls, yielding a medium effect size. The medium effect size for boys, however, was only four-hundredths of a standard deviation lower than the effect size for girls. Furthermore, within-group effects were two-hundredths of a standard deviation apart, favoring girls. Previous research has yielded conflicting results regarding gender. Some studies have found that girls tended to demonstrate higher singing accuracy than boys (Franklin & Franklin, 1988; Goetze & Horii, 1989; Leighton & Lamont, 2006; Mang, 2006; Moore et al., 2004; Moore et al., 1995/1996; Welch, Sergeant, & White, 1997). Other studies have found that there were no statistical differences between genders and singing accuracy (Apfelstadt, 1984, Hedden & Baker, 2010; Moore, 1994; Paney & Kay, 2014; Petzold, 1963; Yarbrough et al., 1992).

Implications for teaching may be made with regard to gender. It may be beneficial to encourage an alternate view regarding the existing trend of comparing male singers to female singers as an independent variable. Both groups may benefit from singing instruction. It seems that conflicting results may be attributed to other variables, such as student attitude towards music or student attitude towards the teacher. Research has also demonstrated that social pressures can increase as a child reaches middle school. The interaction between a researcher's interest in gender differences and a child's vocal insecurities might be problematic for all involved. Rather than demonstrating concern regarding differences between genders, boys and girls alike may benefit from being compared to themselves. Having knowledge of one's own singing development may encourage a sense of empowerment over one's unique ability.

Publication Bias

Music-education-specific meta-analyses have managed publication bias through reporting bias within the body of the study and by eliminating studies until the study population is homogeneous. Standley (2008) reported publication bias, referred to as editorial bias, by providing the effect sizes for both published and unpublished literature. In another study, Standley, (1996) discussed publication sources but did not address publication bias within the body of the paper. Publication bias may have been eliminated in doing so. Mishra (2013) eliminated studies from a sight-reading meta-analysis until the remaining studies were homogeneous. Effect sizes across publication sources may or may not have been homogeneous prior to study elimination in Mishra's (2013) meta-analysis.

Publication bias may be less of a concern within the current meta-analysis. Mishra (2013) grouped all published studies together to investigate if effect sizes varied between unpublished and published studies. When applied to the current meta-analysis, published studies yielded an overall effect of 0.53 whereas unpublished studies yielded an overall effect of 0.28. The difference between effects was reduced to approximately a quarter of a standard deviation, which may have lessened the concern of publication bias. It should be noted, however, that one inclusion criteria statement may have contributed to publication bias within the current meta-analysis: (g) *Studies must have been published within a peer-reviewed journal or a doctoral dissertation.*

Study Quality

Internal Validity: Research Design

Threats to internal validity may have contributed to differences in effects across studies within the meta-analysis. Attempts to prevent threats to internal validity were written into the

inclusion and exclusion criteria statements, specifically regarding research design. The current meta-analysis calculated effects for studies limited to quasi-experimental and experimental designs, specifically those with a pretest to posttest or posttest-only design. Studies within the research design moderator variable group were considered moderately heterogeneous, but closer observations are warranted.

An examination of subgroups within the Research Design moderator variable indicated that threats to internal validity might be accounted for in part by the medium effect size of the pretest-posttest 1 group design and/or the trivial effect size of the posttest-only control group design. The two groups measured nearly half of a standard deviation apart.

In terms of specific designs, threats to internal validity exist for all research designs within the current meta-analysis, including the 1 group pretest-posttest design and posttest-only control group design (Campbell & Standley, 1963). Internal threats exist for all research designs in the current meta-analysis; however, the pretest-posttest 1 group design and the posttest-only control group design included the most extreme effect sizes. The pretest-posttest 1 group design was two-tenths of a standard deviation higher than the subsequent subgroup, the posttest-only control group design was one-tenth and one-hundredth lower than the previous subgroup, and the total difference between the pretest-posttest 1 group design and posttest-only control group design was nearly half of a standard deviation.

Implications for classroom teachers may be made with regard to research design. The pretest-posttest control group design [referred by Campbell and Standley (1963) as the nonequivalent control group design] controls for more threats to internal validity than when researching without a pretest or without a comparison group. When implementing new strategies for vocal development, teachers may see the most easily interpreted benefits by making a limited

number of small changes in instruction with one or two intact classes across a predetermined time period. The other classes could serve as control groups and should receive similar instruction without the added strategies or small changes. A pretest and posttest should be chosen that are efficient, accessible, and will provide information specific to the informal research questions the teacher may be asking. The Singing Voice Development Measure has been most commonly used for investigating stages of vocal development, and computer programs are highly recommended if vocal accuracy is the primary goal.

External Validity

Population source.

The spread of population representation within the current meta-analyses may indicate a threat to external validity. Midwest populations yielded the greatest effects whereas Northeast populations yielded the smallest effects. Although studies within the current meta-analysis represented all regions of the United States as well as Canada and England, the Midwest and Northeast accounted for approximately 65% of the total population. More study populations originated from the Northeast than from anywhere else, with over half from Pennsylvania. The current meta-analysis utilized a fixed-effects model, which limits generalization of effects to the population of included studies as well as similarly-constructed outside studies (Card, 2012). Care must be taken to limit over generalizing.

Sampling.

Qualifications need to be made based on sampling, as well. Of the 34 studies, less than 10% reported random sampling (participants randomly selected from a population), and over half reported random assignment (classes or groups were randomly assigned to a treatment). Random sampling and random assignment were neither accounted for in terms of inclusion and exclusion

criteria nor through moderator variables. Threats existed across the population of studies as many, being quasi-experimental, relied on convenience sampling to do research within classrooms.

Implications for classroom teachers may be made with regard to the generalizability of research study results to other classroom settings, specifically in terms of population and random sampling. The population represented within a study has implications for generalizability. It would be unwise to interpret research results from one population as generalizable to a different population. Demographic information is important for an audience to infer the possible generalizability of the results. Randomization regarding research conducted within a classroom can include random assignment. Overall, the greater the randomization, the more generalizable the results may be to another group of students from the same, or similar, population (also taking into account sample size and sample demographics). Random selection is commonly problematic in a classroom setting. Therefore, teachers would benefit from being cautious when reading results yielded from intact classroom samples. While study results may be significant and meaningful to one sample, with limited random assignment and random selection, the same results may not be found in another sample.

In conclusion, teachers would benefit from taking into account sample or population demographics and randomization when reading results. It would be unadvisable to assume results of one study will work for all. Finally, teachers may want to question the results and research what other similar studies found regarding a topic.

Construct Validity

Measurement instruments.

An analysis of the moderator variable, measurement instruments, indicated a threat to construct validity by the frequency of tests and moderate heterogeneity of the subgroup. All of the effects, however, were generally similar with the exception of the Singing Voice Development Measure. The similar effect sizes may indicate that differences between effects can be attributed to variables other than what measurement instrument was used.

Across the 34 studies that measured vocal accuracy of some kind, nearly all were unique measurement instruments. Many more exist within research outside of the scope of the current meta-analysis. Measurement instruments were grouped by the designer of the instrument, who/what assessed the stimulus, and specific characteristics that were measured. Most were rating scales, although the acoustic measures evaluated pitches based on intonation. Of the 27 instruments, over a third included the use of singing voice as a variable, and about one-fifth included melodic contour. All measures assessed a specific type of accuracy including pitch, pattern, interval, or song.

The breadth of instruments created to measure singing ability has been addressed across literature reviews (Goetze, Cooper, & Brown, 1989; Salvador, 2010). The frequency of instruments may be indicative of the lack of consensus in defining singing ability or the lack of prior knowledge regarding established instruments at the onset of a study. The latter may be more relevant given the similarities across measures within the current meta-analysis.

Measured task.

An analysis of the moderator variable, measured task, indicated a threat to construct validity as indicated by the frequency of tasks and moderate heterogeneity of the subgroup.

measured tasks were divided into three subgroups: (a) song accuracy, (b) pitch accuracy, and (c) pattern accuracy. Similar to measurement instruments, the differences between subgroup overall effects were minimal, which may indicate that differences in effect sizes can be attributed to other variables. As mentioned regarding measurement instruments, the frequency of instruments may be indicative of the lack of consensus in defining singing ability. Additionally, future research regarding singing ability development would benefit from the refinement of current tests rather than continued accumulation of teacher-developed tests.

Limitations of the Current Meta-analysis

The current meta-analysis utilized a fixed-effects model instead of a random-effects model. Although cited as most common (Card, 2012), a fixed-effect model produces results that should be generalized only to the population of meta-analyzed studies as well as other similar studies. A random-effects model, on the other hand, would have produced results that could have been generalized to similar and less similar studies outside of the current population. A replication of the current meta-analysis utilizing both a fixed-effect and random-effect model would provide additional context to the results of the current meta-analysis.

Results of the current meta-analysis were primarily calculated with the use of a graphing calculator. Calculating results without the use of a meta-analysis-specific computer program was both advantageous and problematic. Advantages included the in-depth understanding that was required in order to complete the project. The researcher made every decision to understand the process holistically and atomistically. The purpose was to inform future research endeavors as well as to inform research pedagogy. As Cohen (1990) stated regarding performing his dissertation by hand,

The sheer laboriousness of the task assured that throughout the entire process I was in intimate contact with the data and their analysis. There was no chance that there were funny things about my data or intermediate results that I didn't know about, things that could vitiate my conclusions. (p. 1306)

Although a valuable learning experience, miscalculations may have been possible given the thousands of calculations that were completed. Every effort was made to minimize inaccuracies. However, there was room for human error. The current meta-analysis would benefit from a replication that used an analysis program.

While forgoing meta-analysis software and higher-level statistical programs meant that slight inaccuracies were unavoidable, every attempt was made to avoid inaccuracies. Computer programs generally perform calculations with multiple numbers behind the decimals whereas numbers were rounded to four decimal places in the current meta-analysis. Every effort was made to control for effect inflation and deflation. The current meta-analysis was most sensitive to the inflation or deflation due to the ranking of effects across numerous studies. A follow-up meta-analysis would be beneficial to support and correct the possible rounding issues of the current meta-analysis.

Research Pedagogy and Music Education

Research pedagogy in the field of music education may benefit from including effect size instruction early on in master's or doctoral coursework. Undertaking the current meta-analysis was a task that challenged and utilized every moment in every research course through my doctoral graduate student career. Though I had taken numerous classes regarding research methodology, I found myself regularly visiting course content. The meta-analytic process requires the researcher to be knowledgeable about most research methods—at least enough so

that a researcher can quickly scan and make informed decisions regarding whether or not the study meets the inclusion criteria.

In a survey research study regarding senior researchers' perceptions of doctoral student preparation, Rohwer and Svec (2014) concluded that a cohesive pedagogy might have been lacking in music education doctoral programs across the nation. While doctoral candidates graduated with the same PhD in Music Education, research experiences varied across programs. Factors that may have attributed to varying experiences included (a) limited resources, (b) limited research knowledge of the professors, (c) varying opinions regarding what is important for a doctoral student to know, and (d) bias for or against branches of research. As Heller and O'Connor (2006) stated,

The researcher has a very clear idea of what the question is that will be investigated. The techniques and methodology used to search for an answer to a question should be appropriate to the question posted. (p. 38)

In order to improve research in music education, students may benefit from quality experiences and instruction in and outside of the field. It would be useful for opportunities to be sought by students or provided by faculty to allow time and space to practice research methods. One or two research courses may not be enough experience for doctoral students to find success as tenured, published professors and life-long researchers.

Implications for Future Research

As Cohen (1990) wrote, "less is more" (p. 1304). The included studies of the current meta-analysis yielded over 400 unique effect sizes. Over a third of the studies included more than two groups. About half of the studies included more than one age group, while other studies included gender as an independent variable or more than one dependent variable. Given all the

variables that could be feasibly present within one study, it seems likely that the number of unique effect sizes would be high.

A high frequency of unique effects across few studies may have been problematic in the current meta-analysis. One consequence may have included conflicting results found between the current study findings and findings of the original studies that were limited to statistical significance for interpretation. There were also issues with calculating heterogeneity due to the large numbers of unique effects.

Studies that included multiple independent and/or dependent variables were more difficult from which to glean needed effect size data. Challenges increased beyond the 2x2 ANOVA table. Additionally, confounding variables meant more data to report, but journals have limited space. Upon requesting more data, several emails were returned indicating that enough room was not provided by a journal for publication. As a result, information reported was limited to the necessities.

Therefore, future research may benefit from adopting a “less is more” philosophy. Research may benefit from spending more time and energy gathering a larger sample rather than overusing a small sample that is available. A larger sample may more closely align statistical significance with effect sizes and decrease sensitivity to Type I errors. Perhaps with fewer variables to take up precious space in a journal article, more room may be devoted to data that were often omitted from the current studies within the meta-analysis, including correlation information for repeated measures.

Future research may also benefit from the publication of raw data sets within dissertations. Writing a dissertation provides young researchers with an opportunity to report every piece of information possible. There are no publication limitations with dissertations.

Additionally, meta-analyses have grown in popularity outside of music education, and there are now at least six meta-analyses, including the current one, within the field. In order to continue meta-analyzing music education research, accurately, original data are needed.

Future research may benefit from consistently publishing descriptive statistics in music education research. Publishing overall descriptive statistics has been a common practice. When analyzing multiple independent variables, the inclusion of group sample sizes, group means, and group standard deviations with published results would benefit researchers involved in replication research and meta-analytic research.

Future research may also benefit from consistent reporting of effect sizes and confidence intervals. Out of all of the studies included in the current meta-analysis, one reported d effect sizes, but only for groups with significant differences. To encourage growth of meta-analytic procedures within music education research, it would be beneficial for researchers to include effects for both significant and non-significant results. Without effect size and confidence intervals, information may be lost or go unreported within studies. Future research may also benefit from utilizing repeated-measures more often, since between-group comparisons yield higher threats to validity and smaller effect sizes.

Future research may benefit from more research studies and statistical syntheses regarding singing ability. Several meta-analyses could be beneficial, including (a) a meta-analysis of longitudinal studies, (b) a meta-analysis of between-group comparisons, (c) an expansion of the current meta-analysis to include a broader age range, and (d) a meta-analysis regarding music aptitude in childhood. Future meta-analyses will continue to be possible if primary researchers either report, or store, original means and standard deviations for the purpose of a replication or synthesis study.

Future research regarding singing ability may benefit from time spent developing trusted measures that are accessible to teachers, accessible to students for independent use, and accessible to researchers to gather data. At this time, one of the most widely used measurement instruments has been the Singing Voice Development Measure. Additional measures that may need further development and accessibility include the vocal assessment computer programs. While computer programs may be useful to judge pitch accuracy, perhaps programs can also be developed that can be sensitive to timbre.

In conclusion, singing instruction may be seen as an overall effective means of improving singing ability scores in children between the ages of 5 and 11, specific to the studies and participants represented within the current meta-analysis. Interventions are important to music education, overall and with regard to singing ability. Singing instruction is more effective than, (a) what has been commonly referred to as a traditional curriculum, or one which includes song-singing without vocal development guidance, and (b) no music instruction at all.

I, as a researcher and educator, believe that singing is important and singing instruction is meaningful. As educators, we are consistently asking how instruction can be better and of what we can do more. We are always weighing the question: is what we are doing good enough for our students and for the future of music education? As Zoltan Kodaly said,

If one were to attempt to express the essence of this education in one word, it could only be – singing... Our age of mechanization leads along a road ending with man himself as a machine; only the spirit of singing can save us from this fate. (Bónis, 1974, p. 206)

APPENDIX A
CONTENT VALIDITY PANEL LETTER

Dear Dr.____ ,

Thank you for reviewing this document. I hope you have decided to review the inclusion and exclusion criteria statements as well as the initial list of studies to be meta-analyzed as part of my dissertation: A Meta-analysis of Singing Ability in Children Ages 5-11. I provided my purpose statement, a short explanation of the process, the lists of inclusions and exclusion criteria statements, and the lists of questions to guide your feedback. You should have also received an attachment that includes the list of gathered studies based on the inclusion and exclusion criteria statements. After making revisions based on your recommendations, the studies on the final research list will serve as the population for the current meta-analysis. I look forward to receiving your feedback, questions, and concerns regarding the statements and research list soon!

Purpose Statement

The purpose of the current meta-analysis was two-fold:

1. to investigate overall statistical effects of instructional characteristics on singing ability in children ages 5-11; and
2. to investigate statistical effects within and across the following moderator variables: (a) instruction (b) age of children, (c) gender, (d) population, (e) research design, (f) measurement instrument, and (g) singing task measured (i.e., pitch accuracy, use of singing voice).

Process Explanation

After establishing research questions and goals of a meta-analysis, inclusion and exclusion criteria stating desired and undesired study characteristics are commonly specified and listed in the literature for the purpose of providing a transparent literature search. If another

researcher were to replicate the current meta-analysis, he/she would compile an identical list of studies (with the addition of studies published after the original meta-analysis). For the current meta-analysis, the process includes defining initial criteria statements, compiling a list of literature based on the statements, checking the statements and list of studies with a content validity panel of three field-specific research experts, and revising the list of statements and studies as needed.

Inclusion Criteria Statements

- Studies must have investigated singing ability, also referred to in research as singing achievement or singing development;
- studies must have included effect size information or sufficient data used to compute an effect size;
- singing ability must have been represented by measuring a characteristic or characteristics (i.e., pitch accuracy, use of singing voice) of singing performance;
- singing performances must have required students to sing independently for at least part of the data collection;
- studies must have been quantitative with experimental or quasi-experimental research designs;
- study sample population must have been limited to children between the ages of five and eleven/kindergarten through fifth grade (as labeled in the American school system) as well as sixth grade under the circumstance that sixth grade was part of the elementary school;
- studies must have been written in English; and
- studies must have been published within a peer-reviewed journal or doctoral dissertation.

Exclusion Criteria Statements

- Studies must not have been longitudinal;
- studies must not have investigated relationships between singing ability and other constructs (absent of instruction);
- studies must not have included an investigation regarding the singing ability of children who speak a tonal language (i.e., Cantonese, Taiwanese);
- studies must not have included preschool, middle school, and junior high school participants;
- studies must not have measured the singing ability of groups of children (as in a choral setting) where the group was measured as $n = 1$;
- studies must not have included results that are only generalizable to specific populations (e.g., special needs children, children's choirs, private voice students); and
- studies must not have included investigations of singing ability with language acquisition.

Please refer to the following questions when providing feedback:

1. How could the inclusion/exclusion criteria statements be modified to better align with the purpose of the meta-analysis?
2. What additional characteristics of singing ability should be described in the inclusion/exclusion criteria statement lists?
3. How could the list of studies more accurately represent the inclusion statements?
 - a. What statements should be added?
 - b. What studies should be added?
 - c. What statements should be omitted?

- d. What studies should be omitted?
- 4. How does the list of studies conflict with the exclusion statements?
 - a. What statements should be added?
 - b. What statements should be omitted?

APPENDIX B

GEOGRAPHIC REGIONS OF THE UNITED STATES

Geographic Regions of the United States

Northeast	Southeast	Southwest	West	Midwest
Connecticut	Alabama	Arizona	California	Illinois
Delaware	Arkansas	New Mexico	Colorado	Indiana
Maine	Florida	Oklahoma	Idaho	Iowa
Massachusetts	Georgia	Texas	Montana	Kansas
New Hampshire	Kentucky		Nevada	Michigan
New Jersey	Louisiana		Oregon	Minnesota
New York	Maryland		Utah	Missouri
Pennsylvania	Mississippi		Washington	Nebraska
Rhode Island	North Carolina		Wyoming	North Dakota
Vermont	South Carolina			Ohio
	Tennessee			South Dakota
	Virginia			Wisconsin
	Washington DC			
	West Virginia			

Note. National Geographic Society (2009). *United States Regions*. Retrieved from http://education.nationalgeographic.com/education/maps/united-states-regions/?ar_a=1

APPENDIX C
EXCLUDED ARTICLES BY RATIONALE

Table C.1

Rationale for Excluding Articles: Age

< 5 Years Old	> 11 Years Old	< 5 Years and > 11 Years
Boardman (1964)	Hedden & Johnson (2008)	Colwell (1963)
Dansereau (2005)	Lamant, Daubney, & Spruce (2012)	Fuchs et al. (2009)
Feierabend, Saunders, Holahan, & Getnick (1998)	Mang (2007)	Price, Yarbrough, Jones, & Moore (1994)
Flowers & Dunne-Sousa (1990)	Bella & Berkowska (2009a)	
Geringer (1983)	Tremblay-Champoux, Bella, Phillips-Silver, Lebrun, & Peretz (2010)	
Kreutzer (2001)	Wang & Sogin (1997)	
Moog (1976)		
Ramsey (1983)		
Runfola, Etopio, Hamlen, & Rozendal (2012)		
Trollinger (2003)		
Young (1974)		

Table C.2

Rationale for Excluding Articles: No Instructional Treatment

No Instructional Treatment			
Barlow & Howard (2005)	Goetze & Horii (1989)	Killian (1999)	Sergeant & Welch (2009)
Bennett (1991)	Green (1990)	Levinowitz et al. (1998)	Smith (1973)
Brand (2000)	Green (1994)	Mizener (1993)	Sterling (1984)
Cooper (1992)	Guerrini (2002)	Moore (1994)	Steenwyk (2004)
Cooper (1995)	Guerrini (2006)	Moore, Brotons, Fyk, & Castillo (1997)	Tatem (1990)
Davies & Roberts (1975)	Hedden (1982)	Moore, Fyk, Frega, & Brotons, (1995/1996)	Tsang & Conrad (2011)
Elmer (2000)	Hermanson (1971)		Wassum (1979)
Froehlich (1977)	Hickey (1995)	Nichols (2013)	Welch & White (1993/1994)
Froehlich (1979)	Hornbach & Taggart (2005)	Patrick (1978)	
Fyk (1985)	Howard, Williams, & Herbst (2014)	Seltzer (1983)	Wolf (2003)
Goetze (1985)			Wolf (2005)
			Yarbrough, Bowers, & Benson (1992)
			Yarbrough, Green, Benson, & Bowers (1991)

Table C.3

Rationale for Excluding Articles: Research Design/Genre

Literature Review	Historical Research
Atterbury (1984)	Britton (1989)
Berkowska & Bella (2009)	Coffman (1987)
Brand (2000)	Fisher (1973)
Davies & Roberts (1975)	John (1954)
Elmer (2011)	
Ferguson (2005)	Qualitative Research
Hedden (2012)	Liao & Davidson (2007)
Hedden (2011)	
Liao & Davidson (2007)	Longitudinal Design Research
Mackworth-Young (1978)	Leighton & Lamont (2006)
	Miller (1972)
Welch (1979a)	Petzold (1963)
Welch (1979b)	Rutkowski & Snell Miller (2002)
Welch (1985)	Wassum (1980)
Welch (1994)	Welch, Sergeant, & White (1997)
Welch (1998)	Welch, Sergeant, & White (1998)

Table C.4

Rationale for Excluding Articles: Topic, Tonal Language, Population, and Insufficient Data

Topic	Tonal Language
<i>Sight-singing</i>	Chen-Hefteck (1999)
Baker (1980)	Chen-Hefteck, Niekerk, Labaka, & Masuetete (1999)
Blyler (1960)	Keller & Tyler (2011)
Bowles (1998)	Liao (2008)
<i>Preference</i>	Mang (2006)
Reifinger (2009)	Moore, Chen, & Brotons (2004)
Reifinger (2012)	Trollinger (2003)
<i>Attitude</i>	Trollinger (2004)
	Population
Jones (1993)	Eyck (1985)
Siebenaler (2008)	Hamdan et al. (2009)
<i>Conservation</i>	Odom (1991)
Foley (1975)	Data
<i>Listening</i>	Buckton (1977)
McLean (1999)	Doneski (2005)
<i>Singing games</i>	Gaiser (1961)
Moore (2002)	Gault (2000)
<i>Improvisation</i>	Gault (2002)
Guilbault (2009)	Holmes (2009)
<i>Auditory encoding</i>	Martin (1991)
Phillips-Silver & Trainor (2007)	Mathias (1997)
<i>Aptitude</i>	Petzold (1969)
Lang (1999)	Phillips (1985)
	Roberts & Davies (1976)
	Small & McCachern (1983)
	Welch (1985)
	Wyatt (1993)

APPENDIX D
PILOT STUDY

A pilot study was implemented to estimate the feasibility, consistency, and transparency of the meta-analytic process. The four pilot study research questions included: (1) what is the overall statistical effect of instructional characteristics on singing ability in children younger than 5 years old; (2) what are the statistical effects within and across the following primary moderator variables: (a) instruction, (b) age of children, (c) gender, (d) measurement instrument, and (e) singing task; and (3) what are the statistical effects of study characteristics within and across the following moderator variables: (a) publication source, (b) publication year, (c) population, (d) research design, and (e) treatment period; and (4) was the analytic process feasible?

Pilot study data were extracted from the pool of unused/excluded articles (see Tables C.1 for exclusion rationale) from the full study in order to maintain consistency of topic and moderator variables. Of the 130 unused and excluded articles, 10 articles investigated singing instruction with children under five years old. Of those 10 articles, seven were excluded for the following two reasons: (a) use of research designs other than pretest/posttest control group or posttest-only control group designs (Boardman, 1964; Flowers & Dunne-Sousa, 1990; Geringer, 1983; Moog, 1976; Trollinger, 2003) and (b) use of dependent variables other than singing ability (Feierabend, Suanders, Holahan, & Getnick, 1998; Runfolo, Etopio, Hamlen, & Rozendal, 2012). The remaining three studies were coded for the following variables: (a) citation, (b) sample size, (c) age, (d) gender, (e) teaching condition, (f) test, (g) task, (h) population, (i) source, (j) research design, and (k) treatment period.

Pilot Discussion and Limitations

While the pilot meta-analysis was successful in establishing a procedure for the main study, there were limitations. Statistical issues with study effect sizes came to light through the course of the pilot study that is adjusted in the main study. Standard deviations and means were

provided within each of the studies. For the pretest-posttest designs, effect sizes were calculated for the experimental group mean scores pretest to posttest. Experimental group effect sizes were then used throughout the remainder of the pilot. Experimental group effect sizes, however, may be inflated as the control group effects were omitted from overall effect sizes. Factoring in the experimental group effect sizes accounted for the natural gain that may have happened regardless of instruction. Additionally, between-group effect sizes were also used to determine the overall effect size for pretest-posttest control group designs to account for group differences. Therefore, in the main meta-analysis, the effect size of a pretest-posttest control group design study is a reflection of the pretest-posttest experimental group means and standard deviations, the pretest-posttest control group means and standard deviations, and the between-group posttest means and standard deviations.

An inability to offer holistic conclusions from the results of the pilot meta-analysis was an additional limitation given the small number of studies that were meta-analyzed. However, the pilot does indicate that the meta-analytic process was found to be feasible for the primary study. Moderator variables persisted for the primary study, and the larger frequency of studies yielded thorough results; all moderators were more adequately represented in the full study.

Pilot Procedures

Several options were available to analyze current meta-analytic processes. Card (2012) recommended Excel, SPSS, or SAS for young researchers over meta-analysis computer programs. Card (2012) elaborated,

The danger of relying on them [meta-analysis programs] exclusively – especially when you are first learning to conduct meta-analyses – is that they may encourage erroneous use when you are not adequately familiar with the techniques. (p. 14)

Although SPSS macros were located to aid in the use of SPSS for meta-analytic analyses, the syntax commands were out of date and did not align with SPSS 22. Therefore, Excel was used to calculate equations in the current meta-analysis where warranted, specifically regarding calculating p-values and sums. All other calculations were done on a TI-83 graphing calculator.

Intrarater Reliability

In a meta-analysis, intrarater reliability has been cited as the extent to which study characteristics were coded consistently (Card, 2012). The Agreement Rate (AR) formula has been generally accepted as the most commonly used formula where $AR = (\text{number of observations agreed upon}) / (\text{total number of observations})$ (Card, 2012; Orwin & Vevea, 2009, p. 187). AR does not control for coding agreement happening by chance (as does Cohen's Kappa, for example). Controlling for chance, however, was not necessary in the current study because, cells were either in agreement or disagreement.

To check for intrarater reliability, the researcher recoded all of the pilot studies two weeks after initial coding. Twelve categories across four studies yielded 48 total observations. Of the 48 total observations, there were 38 agreements and 10 disagreements, yielding an AR of 79.20%.

The AR of 70% or higher has been commonly cited as acceptable across studies (Card, 2012), however a higher percentage rate was warranted given the increased volume of coded information to be included in the full study. Four observations were made regarding reasons for disagreement between coding and recoding. First, the singing task was not clearly labeled within two of the four studies. Therefore, *singing task* as a category was left blank in the first coding, but in the recoding, the researcher included all information for *singing task*. Second, the category *scoring singing* was problematic because it was not clearly defined before coding. The researcher

added the category of *scoring singing* after observing several differences regarding scoring procedures. While the data pulled from *scoring singing* as a moderator variable might have provided useful information, more investigation may be warranted after completion of the current meta-analysis. Other differences included additional information provided by primary researchers during the recoding period, and information being omitted from cells in the initial coding stage.

Upon removing the category *scoring singing*, AR was checked again and found tenable at 84.10%. Eleven categories across four studies yielded 44 total observations. Of the 44 total observations, there were 37 agreements and seven disagreements, yielding an increased AR of 84.10%. *Scoring singing* was then omitted from the full study. The final coding has been listed in Table 3.

Pilot Study Results

The population of the current pilot study was comprised of three studies with a combined sample size of $N = 258$ participants. Pilot study participants ranged from three- to four-years old ($M = 3.67$, $SD = 0.58$). See Table D.1 for specific information regarding each study grouped by primary moderators and secondary moderators; see Table D.2 for all within and between study effect sizes. The current meta-analysis yielded a moderate overall weighted mean effect size for instruction ($\overline{ES} = 0.66$, 95% CI [0.41, 0.91]). A non-significant Q test indicated that studies were homogeneous ($Q = 1.26$; $df = 2$; $p = 0.53$). Moderator variables selected a priori were evaluated to observe effect sizes further. See Table D.3 for a summary of results.

Table D.1

Singing ability pilot studies and moderator variables included in analysis

Primary Moderators						
Citation	N	Age	Gender	Teaching Condition	Test	Task
Dansereau (2005)	N = 39 $n_c = 17$ $n_e = 22$	3	girls = 18 boys = 28	Music instruction vs. no music instruction	SVDM	Singing voice use/contour
Ramsey (1983)	N = 91 $n_e = 43$ $n_c = 48$	4	girls = 28 boys = 46	Instrumental vs. non-instrumental	PSALT	Pitch Acc. Contour Acc. Tonal Center Interval Acc.
Young (1974)	N = 128 $n_e = 64$ $n_c = 64$	4	-	Music instruction vs. no music Instruction	PMAT	Pitch Acc. Interval Acc. Pattern Acc.
Secondary Moderators						
Citation	Population	Source	Research Design		Treatment Period	
Dansereau (2005)	South	Unpublished dissertation	Pretest-posttest control group		19 weeks	
Ramsey (1983)	Midwest	JRME	Posttest-only two experimental groups		6 weeks	
Young (1974)	South	JRME	Pretest-posttest control group		9 weeks	

Note. “ - “ indicates unavailable information.

Study mean effect sizes ranged from 0.40 to 0.79. Variable effect sizes ranged from 0.93 (the effect of music instruction on unfamiliar phrase repetition) to -1.34 (the negative effect of non-instrumental instruction on maintaining a tonal center). Between group effect sizes were indicative of the estimated magnitude of mean differences between the control and experimental groups. Within group effect sizes were indicative of the estimated magnitude of mean differences within one group from pretest to posttest. The directionality of effect sizes indicated whether an instructional condition influenced a measure positively or negatively. Within effect sizes were positive except in Young’s (1974) study regarding mean control group gains from pretest to posttest, indicating that the absence of music instruction had an estimated minimal, but negative

effect on music achievement mean scores (ES = -0.02, -0.06, -0.07, -0.07, -0.21). Within group effect sizes were positive with the exception of three of Ramsey's (1983) measures of singing ability, indicating that the absence of instrumental accompaniment had an estimated medium to large negative effect on music achievement mean scores (ES = -0.57, -0.92, -1.34).

Table D.2

Effect Sizes Within Condition Groups, Between Condition Groups, and Overall

Citation	Test	Teaching Condition	N	ES		<i>M</i>	<i>CI</i>	
				w/in	b/t		<i>UL</i>	<i>LL</i>
Dansereau (2005)	SVDM		39			0.40	-0.20	1.00
		Music instruction	22	*0.40	-			
		No music	17	0.07				
Ramsey (1983)	PSALT		91			0.62	0.21	1.03
	Absolute pitch.	Instrumental inst.	43					
		Non-instrumental	48	-	0.32			
	Melodic rhythm	Instrumental inst.	43					
		Non-instrumental	48	-	0.15			
	Melodic contour	Instrumental inst.	43					
		Non-instrumental	48	-	-0.92			
	Tonal Center	Instrumental inst.	43					
		Non-instrumental	48	-	-1.34			
	Interval match.	Instrumental inst.	43					
		Non-instrumental	49	-	-0.57			
Young (1974)	Hill PAMT		128			0.79	0.43	1.15
	Interval match.	Music instruction	64	0.77	0.54			
		No music	64	- 0.07				
	Unfamiliar phrase	Music instruction	64	0.93	0.38			
		No music	64	- 0.06				
	Tone matching	Music instruction	64	0.63	0.45			
		No music	64	- 0.02				
	Familiar song	Music instruction	64	0.70	0.34			
		No music	64	- 0.21				
	Tonal melodic achievement	Music instruction	64	0.89	0.43			
		No music	64	- 0.07				

Note. - indicates unavailable information. * indicates adjusted effect size.

All moderators were measured categorically. To establish homogeneity between subgroups within moderator variables, each subgroup had to include at least two studies. Several moderator variable subgroups (measurement instrument, gender, and treatment period) had fewer than two studies, and were therefore omitted from further analysis. As expected from a

homogeneous group, all effect sizes were moderate to large, with little variation within moderator variable groupings. The effect of music instruction ($ES = 0.69$, 95% CI [0.38, 1.00]) and instrumental accompaniment ($\overline{ES} = 0.62$, 95% CI [-0.60, 1.84]) had a moderate, positive effect on singing ability. Age groupings yielded a moderate effect for three-year old children ($\overline{ES} = 0.41$, 95% CI [-0.19, 1.01]) and a large effect for four-year old children ($\overline{ES} = 0.78$, 95% CI [0.43, 1.13]). Measured task yielded a larger spread of effect sizes, though insignificant. Effect sizes ranged from a moderate effect on use of singing voice ($\overline{ES} = 0.41$, 95% CI [-0.19, 1.01]) to a large effect on tonal center accuracy ($\overline{ES} = 1.34$, 95% CI [0.89, 1.79]).

Several moderator variables were chosen a priori to reveal bias issues, assuming their existence within the current meta-analysis. The pilot meta-analysis population was grouped by the geographical location of the study participants (North, South, Midwest, West). Population may not have contributed to bias; population effect sizes were similar and homogeneous between the subgroups South ($\overline{ES} = 0.69$, 95% CI [0.38, 1.00]) and Midwest ($\overline{ES} = 0.62$, 95% CI [-0.60, 1.84]). Research design may not have contributed to bias; population and research design effect sizes were similar and homogeneous between the subgroups including pretest-posttest control group design ($\overline{ES} = 0.69$, 95% CI [0.38, 1.00]) and posttest-only ($\overline{ES} = 0.62$, 95% CI [-0.60, 1.84]). For publication type, unpublished dissertations yielded a moderate effect size ($\overline{ES} = 0.41$, 95% CI [-0.19, 1.01]) while studies published in the Journal of Research in Music Education yielded a large effect ($\overline{ES} = 0.78$, 95% CI [0.43, 1.13]).

Table D.3

Mean Effect Sizes for Moderator Variables

Group	N	n	ES	SE _{ES}	Teaching Condition		Heterogeneity Statistics		
					95% CI		Q _{w/i}	Q _{b/t}	Q _{total}
					LL	UL			
Music instruction vs. no music instruction	2	167	0.69	0.16	0.38	1.00	1.14		
Instrumental acc. vs. no instrumental acc.	1	91	0.62	0.62	-0.60	1.84	-		
								0.12	1.26
Population									
South	2	167	0.69	0.16	0.38	1.00	1.13		
Midwest	1	91	0.62	0.62	-0.60	1.84	-		
								0.13	1.26
Research Design									
Pretest-posttest control group	2	167	0.69	0.16	0.38	1.00	1.13		
Posttest-only	1	91	0.62	0.62	-0.60	-1.84	-		
								0.13	1.26
Sample Age									
Three years old	1	39	0.41	0.31	-0.19	1.01	-		
Four years old	2	219	0.78	0.18	0.43	1.13	0.08		
								1.18	1.26
Publication Type									
Unpublished Dissertation	1	39	0.41	0.31	-0.19	1.01	-		
JRME	2	219	0.78	0.18	0.43	1.13	0.08		
								1.18	1.26
Measured Task									
Use of singing voice	1	39	0.41	0.31	-0.19	1.01	-		
Contour accuracy	1	91	0.92	0.22	0.49	1.35	-		
Tonal center accuracy	1	91	1.34	0.23	0.89	1.79	-		
Pattern accuracy	1	128	0.93	0.18	0.58	1.28	-		
Pitch accuracy	2	219	0.56	0.14	0.29	0.83	1.23		
Interval accuracy	2	219	0.69	0.14	0.42	0.96	0.42		
								0.51	1.26

Note. Population of studies = *N*. Study sample size = *n*.

Pilot Discussion and Limitations

While the pilot meta-analysis was successful in establishing a procedure for the main study, there were limitations. In the pilot meta-analysis, intrarater reliability was checked using an agreement rate formula that divided the total number of agreements by the total number of studies. Although this method has been cited as common and intuitive across meta-analyses (Card, 2012), it has also been criticized for not factoring in the element of chance – that two cells

could have been identically coded across coding and recoding (Orwin & Vevea, 2009). To aid in controlling for the element of change, Cohen's Kappa agreement rate was used in the primary meta-analysis as it has been cited as controlling for the element of chance (Orwin & Vevea, 2009). Cohen's Kappa agreement rate was not utilized in the pilot due to the researcher's initial unfamiliarity with systematic coding procedures.

Statistical issues with study effect sizes came to light through the course of the pilot study that are adjusted in the main study. Standard deviations and means were provided within each of the studies. For the pretest-posttest designs, effect sizes were calculated for the experimental group mean scores pretest to posttest. Experimental group effect sizes were then used throughout the remainder of the pilot. Experimental group effect sizes, however, may be inflated as the control group effects were omitted from overall effect sizes. Factoring in the experimental group effect sizes accounted for the natural gain that may have happened regardless of instruction. Additionally, between group effect sizes were also used to determine the overall effect size for pretest-posttest control group designs to account for group differences. Therefore, in the main meta-analysis, the effect size of a pretest-posttest control group design study is a reflection of the pretest-posttest experimental group means and standard deviations, the pretest-posttest control group means and standard deviations, and the between group posttest means and standard deviations.

An inability to offer holistic conclusions from the results of the pilot meta-analysis was an additional limitation given the small number of studies that were meta-analyzed. However, the pilot does indicate that the meta-analytic process was found to be feasible for the primary study. Moderator variables persisted for the primary study, and the larger frequency of studies yielded thorough results; all moderators were more adequately represented in the full study.

REFERENCES

- *Aaron, J. C. (1990). *The effects of vocal coordination instruction on the pitch accuracy, range, pitch discrimination, and tonal memory of inaccurate singers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9103179)
- Acklin, A. I. (2009). *The effect of conducting on ensemble performance: A “best-evidence” synthesis* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3385224)
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- *Apfelstadt, 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. doi:10.2307/3345277
- Atterbury, B. W. (1984). Children’s singing voices: A review of selected research. *Bulletin of the Council for Research in Music Education*, 80, 51–63.
- Atterbury, B. W. (1992). Research on the teaching of elementary general music. In R. Colwell (Ed.), *Handbook of research on music teaching and learning* (pp. 594–501). New York, NY: Schirmer Books.
- *Atterbury, B. W., & Silcox, L. (1993). The effect of piano accompaniment on kindergartners’ developmental singing ability. *Journal of Research in Music Education*, 41, 40–47. doi:10/2307/3345378

- Baker, D. S. (1980). The effect of appropriate and inappropriate in-class song performance models on performance preference of third- and fourth-grade students. *Journal of Research in Music Education*, 28, 3–17.
- Barlow, C., & Howard, D. M. (2005). Electrolaryngographically derived voice source changes of child and adolescent singers. *Logopedics Phoniatrics Vocology*, 30, 147–157.
doi:10.1080/14015430500294031
- Bechtold, M. L., Puli, S. R., Othman, M. O., Bartalos, C. R., Marshall, J. B., & Roy, P. K. (2009). Effect of music on patients undergoing colonoscopy: A meta-analysis of randomized controlled trials. *Digestive Diseases and Sciences*, 54, 19–24.
doi:10.1007/s10620-008-0312-0
- Bella, S. D., & Berkowska, M. (2009a). Singing and its neuronal substrates: Evidence from the general population. *Contemporary Music Review*, 3, 279–291.
doi:10.1080/07494460903404386
- Bella, S. D., & Berkowska, M. (2009b). Singing proficiency in the majority: Normality and “phenotypes” of poor singing. *Annals of the New York Academy of Sciences*, 1169, 99–107. doi:10.1111/j.1749-6632.2009.04558.x
- Bella, S. D., Giguere, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America*, 121, 1182–1189. doi:10.1121/1.2427111
- Bennett, P. D. (1991). Children’s pattern perception, accuracy, and preference in three response modes. *Journal of Research in Music Education*, 39, 74–85. doi:10.2307/3344610
- Bentley, A. (1969). *Monotones. Music education research papers*. London, England: Novello.
- Berkowska, M., & Bella, S. D. (2009). Acquired and congenital disorders of sung performance: A review. *Advances in Cognitive Psychology*, 5, 69–83. doi:10.2478/v10053-008-0068-2

- Blyler, D. (1980). The song choices of children in the elementary grades. *Journal of Research in Music Education*, 8, 9–15.
- Boardman, E. L. (1964). *An investigation of the effect of preschool training on the development of vocal accuracy in young children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 6408354)
- Bónis, F. (Ed.). (1974). *The selected writings of Zoltán Kodály*. London, England: Boosey & Hawkes.
- Borenstein, M. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.
- Bowles, C. (1998). Music activity preferences of elementary students. *Journal of Research in Music Education*, 46, 193–207.
- Brand, E. (2000). Children's mental organizations as highlighted by their singing errors. *Psychology of Music*, 28, 62–80. doi:10.1177/0305735600281005
- Brittin, R. V., & Standley, J. M. (1997). Researchers in music education/therapy: Analysis of publications, citations, and retrievability of work. *Journal of Research in Music Education*, 45, 145–160.
- Britton, A. (1989). The how and why of teaching singing schools in eighteenth century America. *Bulletin of the Council for Research in Music Education*, 98, 23–41.
- *Buck, (1991). *An experimental study using the PITCH MASTER and TAP MASTER systems to improve music literacy and singing skills* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 303986562)
- Buckton, R. (1977). A comparison of the effects of vocal and instrumental instruction on the development of the melodic and vocal abilities in young children. *Psychology of Music*, 5, 36–47. doi:10.1177/030573567751006

- Butzlaff, R. (2000). Can music be used to teach reading? *Journal of Aesthetic Education*, 34, 167–178. Retrieved from <http://www.jstor.org/stable/3333642>
- Campbell, D. T., & Standley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Belmont, CA: Wadsworth.
- Campbell, P. S. (1998). *Songs in their heads: Music and its meaning in children's lives*. New York, NY: Oxford University Press.
- Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.
- Chabris, C. F. (1999). Prelude or requiem for the “Mozart effect”? *Nature*, 400, 826–827. Retrieved from <http://dx.doi.org/10.1038/23608>
- Chen-Hafteck, L. (1999). Discussing text-melody relationship in children's song-learning and singing: A Cantonese-speaking perspective. *Psychology of Music*, 27, 55–70. doi:10.1177/0305735699271006
- Chen-Hafteck, L., van Niekerk, C., Lebaka, E., & Masuetete, P. (1999). Effects of language characteristics on children's singing pitch: Some observations on Sotho- and English-speaking children's singing. *Bulletin of the Council for Research in Music Education*, 141, 26–31.
- Clark, I. N., Taylor, N. F., & Baker, F. A. (2012). Music interventions and physical activity in older adults: A systematic literature review and meta-analysis. *Journal of Rehabilitation Medicine*, 44, 710–719. doi:10.2340/16501977-1025
- Coffman, D. D. (1987). Vocal music and the classroom teacher, 1885 to 1905. *Journal of Research in Music Education*, 35, 92–102.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Taylor & Francis Group.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
Retrieved from <http://dx.doi.org/10.1037/0003-066X.45.12.1304>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Colwell, R. (1963). An investigation of musical achievement among vocal students, vocal-instrumental students, and instrumental studies. *Journal of Research in Music Education*, 11, 123–130.
- Colwell, R. (Ed.) (1992). *The handbook of research on music teaching and learning: A project of the Music Educators National Conference*. New York, NY: Schirmer Books.
- Colwell, R. (Ed.) (2006). *MENC handbook of musical cognition and development*. New York, NY: Oxford University Press.
- Colwell, R. (Ed.) (2006). *MENC handbook of research methodologies*. New York, NY: Oxford University Press.
- Colwell, C. M., & Heller, G. N. (2003). Lowell Mason's "The Song Garden" (1864-66): Its background, content, and comparison to a twentieth-century series. *Journal of Research in Music Education*, 51, 231–244. doi:10.2307/3345376
- Colwell, R., & Richardson, C. (Eds.) (2002). *The new handbook of research on music teaching and learning: A project of the Music Educators National Conference*. New York, NY: Oxford University Press.
- Colwell, R., & Webster, P. (Eds.) (2011). *MENC handbook of research on music learning volume 1: Strategies*. New York, NY: Oxford University Press.

- Colwell, R., & Webster, P. R. (Eds.) (2011). *MENC handbook of research on music learning volume 2: Applications*. New York, NY: Oxford University Press.
- Consortium of National Arts Education Associations. (1994). *National standards for arts education: What every young American should know and be able to do in the arts*. Reston, VA: Music Educators National Conference.
- Cooksey, J. M. (1993). Do adolescent voices “break” or do they “transform”? *Voice*, 2, 15–39.
- Cooksey, J. M. (2000). Male adolescent transforming voices: voice classification, voice skill development, and music literature selection. In L. Thurman, & G. Welch (Eds.), *Bodymind and voice* (pp. 821–841). Collegeville, MN: VoiceCare Network.
- Cooksey, J. M., Beckett, R. L., & Wiseman, R. (1984). A longitudinal investigation of selected vocal physiological, and acoustical factors associated with voice maturation in the junior high school male adolescent. In E. M. Runfola (Ed.), *Proceedings of the Research Symposium on the Male Adolescent Voice* (pp. 4–60). Buffalo: State University of New York at Buffalo Press.
- Cooper, H., Ledges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.
- Cooper, N. A. (1992). *Selected factors related to children’s singing accuracy* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9307489)
- 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

- Dastgheib, S. S., Layegh, P., Sadeghi, R., Foroughipour, M., Shoeibi, A., & Gorji, A. (2013). The effects of Mozart's music on interictal activity in epileptic patients: Systematic review and meta-analysis of the literature. *Current Neurology and Neuroscience Reports*, 14, 420. doi:10.1007/s11910-013-0420-x
- Dansereau, D. R. (2005). *The musicality of 3-year-old children within the context of research-based musical engagement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3180934)
- Davies, A., & Roberts, E. (1975). Poor pitch singing: A survey of its incidence in school children. *Psychology of Music*, 3, 24–36.
- Doneski, S. M. (2005). *The effects of wait time on the tonal pattern performance accuracy of second- and fourth-grade students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3188691)
- Doscher, B. M. (1994). *The Functional unity of the singing voice*. Lanham, MD: Scarecrow Press.
- Dreu, M. J., Wilk, A. S., Poppe, E., Kwakkel, G., & Wegen, E. E. (2012). Rehabilitation, exercise therapy and music in patients with Parkinson's disease: A meta-analysis of the effects of music-based movement therapy on walking ability, balance and quality of life. *Parkinsonism and Related Disorders*, 18S1, S114–S119. doi:10.1016/S1353-8020(11)70036-0
- Ebie, B. D. (2002). Characteristics of 50 years of research samples found in the "Journal of Research in Music Education," 1953–2002. *Journal of Research in Music Education*, 50, 280–291.

- Ellis, P. D. (2010). *The essential guide to effect sizes*. Cambridge, NY: Cambridge University Press.
- Elmer, S. S. (2000). A new method for analyzing and representing singing. *Psychology of Music*, 28, 23–42. doi:10.1177/0305735600281003
- Elmer, S. S. (2011). Human singing: Towards a developmental theory. *Psychomusicology: Music, Mind, and Brain*, 21, 13–30.
- Eyck, S. (1985). The effect of programmed materials on the vocal development of selected children's choruses. *Journal of Research in Music Education*, 33, 231–246. doi:10.2307/3345250
- *Feierabend, J. M. (1984). *The effects of specific tonal pattern training on singing and aural discrimination abilities of first grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8410243)
- Feierabend, J. M., Saunders, T. C., Holahan, J. M., & Getnick, P. E. (1998). Song recognition among preschool-age children: An investigation of words and music. *Journal of Research in Music Education*, 46, 351–359. doi:10.2307/3345547
- Ferguson, L. (2005). The role of movement in elementary music education: A literature review. *Update: Applications of Research in Music Education*, 23(2), 23–34. doi:10.1177/87551233050230020104
- Fisher, J. L. (1973). The roots of music education in Baltimore. *Journal of Research in Music Education*, 21, 214–224.
- Flowers, P. F., & Dunne-Sousa, D. (1990). Pitch-pattern accuracy, tonality, and vocal range in preschool children's singing. *Journal of Research in Music Education*, 38, 102–114. doi:10.2307/3344930

- Foley, E. A. (1975). Effects of training in conservation of tonal and rhythmic patterns on second-grade children. *Journal of Research in Music Education*, 23, 240–248.
- Franklin, E., & Franklin, A. D. (1988). The uncertain singer. *Update: Applications of Research in Music Education*, 7(1), 7–23.
- Froehlich, H. (1977). The relationship of selected observational variables to the teaching of singing. *Journal of Research in Music Education*, 25, 115–130.
- Froehlich, H. (1979). Replication of a study on teaching singing in the elementary music classroom. *Journal of Research in Music Education*, 27, 35–45.
- Froehling, H. (2010). *Meta-analysis: A critical introduction of its historical origins*. Cypress, CA: CreateSpace Independent Publishing Platform.
- Fuchs, M., Meuret, S., Thiel, S., Taschner, R., Dietz, A., & Gelbrich, G. (2009). Influence of singing activity, age, and sex on voice performance parameters, on subjects' perception and use of their voice in childhood and adolescence. *Journal of Voice*, 23, 182–189.
doi:10.1016/j.jvoice.2007.09.007
- Fyk, J. (1985). Vocal pitch-matching ability in children as a function of sound duration. *Bulletin of the Council for Research in Music Education*, 85, 76–89.
- Gaiser, P. E. (1961). *A study of tone-matching techniques as remedial instruction for non-singers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 302076475)
- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

- Garlin, F. V., & Owen, K. (2006). Setting the tone with the tune: A meta-analytic review of the effects of background music in retail settings. *Journal of Business Research*, 59, 755–764. doi:10.1016/j.jbusres.2006.01.013
- Gault, B. (2000). *Effects of pedagogical approach, presence/absence of text, and developmental music aptitude on the song performance accuracy of kindergarten and first-grade students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9964435)
- 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.
- 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. doi:10.2307/3345213
- Geringer, J. M., Nelson, J. K., & Kostka, M. C. (1981). Differential assessments of children and adult vocal ranges. *Contributions to Music Education*, 8, 39–45.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3–8.
- Green, G. A. (1987). The effect of vocal modeling on pitch-matching accuracy of children in grades one through sex. (Doctoral dissertation). Dissertation Abstracts International, 48, 1410A.
- Green, G. A. (1990). The effect of vocal modeling on pitch–matching accuracy of elementary schoolchildren. *Journal of Research in Music Education*, 38, 225–231. doi:10.2307/3345186

- Green, G. A. (1994). Unison versus individual singing and elementary students' vocal pitch accuracy. *Journal of Research in Music Education*, 42, 105–114. doi:10.2307/3345495
- Goetze, M. (1985). *Factors affecting accuracy in children's singing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8528488)
- Goetze M., Cooper, N., & Brown, C. J. (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 of Research in Music Education*, 52, 23–30.
- Gold, C., Solli, H. P., Kruger, V., & Lie, S. A. (2009). Dose-response relationship in music therapy for people with serious mental disorders: Systematic review and meta-analysis. *Clinical Psychology Review*, 29, 193–207. doi:10.1016/j.cpr.2009.01.001
- Gold, C., Voracek, M., & Wigram, T. (2004). Effects of music therapy for children and adolescents with psychopathology: A meta-analysis. *Journal of Child Psychology and Psychiatry*, 45, 1054–1063.
- Gould, A. O. (1969). Developing specialized programs for singing in the elementary school. *Bulletin of the Council for Research in Music Education*, 17, 9–22.
- Gordon, E. (2003). *A music learning theory for newborn and young children*. Chicago, IL: GIA.
- Guerrini, S. C. (2002). *The acquisition and assessment of the developing singing voice among elementary students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3040318)

- Guerrini, S. C. (2006). The developing singer: Comparing the singing accuracy of elementary students on three selected vocal tasks. *Bulletin of the Council of Research in Music Education*, 167, 21–31.
- Guilbault, D. M. (2004). The effect of harmonic accompaniment on the tonal achievement and tonal improvisations of children in kindergarten and first grade. *Journal of Research in Music Education*, 52, 64–76. doi:10.2307/3345525
- *Guilbault, D. (2009). The effects of harmonic accompaniment on the tonal improvisations of students in first through sixth grade. *Journal of Research in Music Education*, 57, 81–91. doi:10.1177/0022429409337201
- Hale, C. L. (2006). *Primary students' attitudes towards their singing voice and the possible relationship to gender, singing skill and participation in singing activities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3223373)
- Hamdan, A., Deeb, R., Sibai, A., Rameh, C., Rifai, H., & Fayyad, J. (2009). Vocal characteristics in children with attention deficit hyperactivity disorder. *Journal of Voice*, 23, 190–194. doi:10.1016/j.jvoice.2007.09.004
- Hattwick, M. S. (1933). The role of pitch level and pitch range in the singing of pre-school, first grade, and second grade children. *Child Development*, 4, 281–291.
- 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
- *Hedden, D. G., & Baker, V. A. (2010). Perceptual and acoustical analyses of second graders' pitch-matching ability in singing a cappella or with piano accompaniment. *Bulletin of the*

- Council for Research in Music Education*, 184, 35–48. Retrieved from <http://www.jstor.org/stable/27861481>
- Hedden, D. G., & Johnson, C. (2008). The effect of teaching experience on time and accuracy of assessing young singers' pitch accuracy. *Bulletin of the Council for Research in Music Education*, 2008, 63–72.
- Hedden, S. K. (1982). Prediction of music achievement in the elementary school. *Journal of Research in Music Education*, 30, 61–68.
- Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 37–47). New York, NY: Russell Sage Foundation.
- Heller, J. J. and O'Connor, E. J. (2006). Maintaining quality in research and reporting. In R. Colwell (Ed.), *MENC Handbook of research methodologies* (pp. 38–72). New York, NY: Oxford University Press.
- Hencke, D. E. (2010). *A meta-analysis of brain research on music learning and teaching: Applications for early classroom instruction* (Master's thesis). Available from ProQuest Dissertations and Theses Global database. (UMI No. 744089881)
- Hendley, J. A., & Persellin, D. C. (1996). The comparative effects of the lower adult male voice and the male falsetto voice on children's vocal accuracy. *Update: Applications of Research in Music Education*, 14(2), 9–14.
- Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, 34, 601–629.

- Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educational Researcher*, 39, 229–240.
- Hermanson, L. W. (1971). *An investigation of the effects of timbre on simultaneous vocal pitch acuity of young children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 7204169)
- Hetland, L. (2000a). Learning to make music enhances spatial reasoning. *Journal of Aesthetic Education*, 34, 179–238.
- Hetland, L. (2000b). Listening to music enhances spatial-temporal reasoning: Evidence for the “Mozart Effect.” *Journal of Aesthetic Education*, 34, 105–148. doi:10.2307/3333639
- Hetland, L. (2001). The relationship between music and spatial processes: A meta-analysis (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9992322)
- Hickey, K. M. (1995). *Intonation accuracy in the singing of elementary school children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9616968)
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557–560. doi:10.1136/bmj.327.7414.557
- Hoffer, C. R., Aebischer, D., Hood, M. V., Jipson, W. R., McManus, J. C., Smith, P. A., . . . Wyatt, A. D. (1986). *The school music program: Description and standards*. Reston, VA: Music Educators National Conference.

- Holmes, A. V. (2009). *Effect of fixed-do and movable-do solfege instruction on the development of sight-singing skills in 7- and 8-year-old children* (Doctoral dissertation). Retrieved from <http://purl.fcla.edu/fcla/etd/UFE0024398>
- Howard, D. M., Williams, J., & Herbst, C. T. (2014). "Ring" in the solo child singing voice. *Journal of Voice*, 28, 161–169. doi:10.1016/j.jvoice.2013.09.001
- Hornbach, C., & Taggart, C. (2005). The relationship between developmental tonal aptitude and singing achievement among kindergarten, first-, second-, and third-grade students. *Journal of Research in Music Education*, 53, 322–331. doi:10.1177/002242940505300404
- Huck, S. W. (2012). *Reading statistics and research*. Boston, MA: Pearson Education.
- Huedo-Medina, T., Sanchez-Meca, J., Marin-Martinez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, 11, 193–206. doi:10.1037/1082-989X.11.2.193
- *Jarjisian, C. S. (1981). *The effects of pentatonic and/or diatonic pitch pattern instruction on the rote-singing achievement of young children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8125481)
- Jarjisian, C. S. (1983). Pitch pattern instruction and the singing achievement of young children. *Psychology of Music*, 11, 19–25. doi:10.1177/0305735683111002
- Jersild, A. T., & Bienstock, S. F. (1934). A study of the development of children's ability to sing. *Journal of Educational Psychology*, 25, 281–503.
- John, R. W. (1954). Nineteenth century graded vocal series. *Journal of Research in Music Education*, 2, 103–118.

- Jones, M. (1971). A pilot study in the use of a vertically-arranged keyboard instrument with the uncertain singer. *Journal of Research in Music Education*, 19, 183–194.
doi:10.2307/3343822
- Jones, M. (1993). An assessment of audiation skills of accurate and inaccurate singers in grades 1, 2, and 3. *Update: Applications of Research in Music Education*, 11(2), 14–17.
doi:10.1177/875512339301100204
- *Jordan-DeCarbo, J. (1982). Same/different discrimination techniques, readiness training, pattern treatment, and sex on aural discrimination and singing of tonal patterns by kindergartners. *Journal of Research in Music Education*, 30, 237–246.
doi:10.2307/3345297
- Kampfe, J., Sedlmeier, P., & Renkewitz, F. (2011). The impact of background music on adult listeners: A meta-analysis. *Psychology of Music*, 39, 424–448.
doi:10.1177/0305735610376261
- *Kane, M. J. (1994). *The effects of teacher training upon the developmental music aptitude and music achievement of kindergarten students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9422653)
- Kantorski, V. J. (1995). A content analysis of doctoral research in string education, 1936–1992. *Journal of Research in Music Education*, 43, 288–297.
- *Kim, S. (2000). *The effects of sequential movement activities of first-grade students' solo singing abilities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9976141)

- *Kimble, E. P. (1983). *The effect of various factors on the ability of children to sing an added part* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8326407)
- *Klinger, R., Campbell, P. S., & Goolsby, T. (1998). Approaches to children's song acquisition: Immersion and phrase-by-phrase. *Journal of Research in Music Education*, 46, 24–34.
doi:10.2307/3345757
- Killian, J. (1999). A description of vocal maturation among fifth- and sixth-grade boys. *Journal of Research in Music Education*, 47, 357–369. doi:10.2307/3345490
- Killian, J., Liu, J., & Reid, J. F. (2013). The *Journal of Music Teacher Education*: A content analysis of articles 1991–2011. *Journal of Music Teacher Education*, 22, 85–99.
doi:10.1177/1057083712467637
- *Kramer, S. J. (1985). *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 Global database. (UMI No. 8524224)
- Kratus, J. (1992). Subjects in music education research, 1961–1990. *The Quarterly Journal of Music Teaching and Learning*, 3, 50–54.
- Kreutzer, N. J. (2001). Song acquisition among rural Shona-speaking Zimbabwean children from birth to 7 years. *Journal of Research in Music Education*, 49, 198–211.
- Lamont, A., Daubney, A., & Spruce, G. (2012). Singing in primary schools: Case studies of good practice in whole class vocal tuition. *British Journal of Music Education*, 29, 251–268.
doi:10.1017/S0265051712000083

- Lang, D. M. (1999). *The effect of the use of text in music instruction on the tonal aptitude, tonal accuracy, and tonal understanding of kindergarten students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9948130)
- Lebrun, M. A., Moreau, P., McNally-Gagnon, A., Mignault-Goulet, G., Peretz, I. (2012). Congenital amusia in childhood: A case study. *Cortex*, 48, 683–688.
doi:10.1016/j.cortex.2011.02.018
- Lehman, P. R., Adress, B., Getz, R. P., Graham, R. M., McManus, J. C., Meske, E. B., & Petzold, R. G. (1974). *The school music program: Description & standards*. Reston, VA: Music Educators National Conference.
- Leighton, G., & Lamont, A. (2006). Exploring children's singing development: Do experiences in early schooling help or hinder? *Music Education Research*, 8, 311–330.
doi:10.1080/14613800600957461
- *Levinowitz, L. M. (1987). *An experimental study of the comparative effects of singing songs with words and without words on children in kindergarten and first grade* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8716497)
- Levinowitz, L. M., Barnes, P., Guerrini, S., Clement, M., D'April, P., & Morey, M. J. (1998). Measuring singing voice development in the elementary general music classroom. *Journal of Research in Music Education*, 46, 35–47. doi:10.2307/3345758
- Liao, M. (2008). The effects of gesture use on young children's pitch accuracy for singing tonal patterns. *International Journal of Music Education*, 26, 197–211.
- Liao, M. Y., & Davidson, J. W. (2007). The use of gesture techniques in children's singing. *International Journal of Music Education*, 25, 82–96.

- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Loomba, R. S., & Arora, R. (2012). Effects of music on systolic blood pressure, diastolic blood pressure, and heart rate: A meta-analysis. *Indian Heart Journal*, 64, 309–313.
doi:10.1016/S0019-4832(12)60094-7
- Mackworth-Young, R. (1978). A new device for the measurement of pitch. *Psychology of Music*, 6, 65–67. doi:10.1177/030573567861007
- Mang, E. (1995). The effects of training with male and female vocal modeling on melodic singing achievement of grade one children (Master's thesis). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9967391)
- *Mang, E. (1997). The effects of training with male and female vocal modeling on the singing achievement of grade one children. *Canadian Journal of Research in Music Education*, 38(4), 25–33.
- Mang, E. (2006). The effects of age, gender and language on children's singing competency. *British Journal of Music Education*, 23, 161–174.
- Mang, E. (2007). Effects of musical experience on singing achievement. *Bulletin of the Council of Research in Music Education*, 174, 75–92.
- Mark, M. L., & Gary, C. L. (2007). *A history of American music education*. Lanham, MD: Rowman & Littlefield Education.
- *Marshall, H. D. (2002). *Effects of song presentation method on pitch accuracy of third-grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3079131)
- Martin, B. A. (1991). Effects of hand signs, syllables, and letters on first graders' acquisition of tonal skills. *Journal of Research in Music Education*, 39, 161–170. doi:10.2307/3344696

- Mathias, S. L. (1997). *A teaching technique to aid the development of vocal accuracy in elementary school students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9731675)
- *McGraw, A. G. (1996). *An assessment of the effectiveness of vocalizes in training elementary school children to sing using head voice* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9705397)
- McLean, C. (1999). Elementary directed listening to music: A singing approach versus a nonsinging approach. *Journal of Research in Music Education*, 47, 239–250.
- *Michaud, B. G. (2014). *Male music teachers and singing fathers: Effects on and correlations with kindergarten children's singing abilities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3581070).
- Miller, P. H. (1972). *An experimental analysis of the development of tonal capabilities of first grade children* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.
- Mishra, J. (2013). Factors related to sight-reading accuracy: A meta-analysis. *Journal of Research in Music Education*, 61, 452–465. doi:10.1177/0022429413508585
- Mishra, J. (2014). Improving sightreading accuracy: A meta-analysis. *Psychology of Music*, 42, 131–156. doi:10.1177/0305735612463770
- *Miyamoto, K. (2005). Effects of a remedial singing method on the vocal pitch accuracy of inaccurate elementary singers. *Research & Issues in Music Education*, 3(1), 1–16.
Retrieved from <http://www.stthomas.edu/rimeonline/vol3/miyamoto1.htm>
- Mizener, C. P. (1993). Attitudes of children toward singing and choir participation and assessed singing skill. *Journal of Research in Music Education*, 41, 233–245.
doi:10.2307/3345327

- *Montgomery (1988). *A study of the associations between two means of vocal modeling by a male music teacher and third-grade students' vocal accuracy in singing pitch patterns* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8822404)
- Moog, H. (1976). The development of musical experience in children of pre-school age. *Psychology of Music*, 4, 38–45. doi:10.1177/030573567642004
- Moore, R. (1991). Comparison of children's and adult's vocal ranges and preferred tessituras in singing familiar songs. *Bulletin of the Council for Research in Music Education*, 107, 13–22.
- Moore, R. (1994). Effects of age, sex, and melodic/harmonic patterns on vocal pitch-matching skills of talented 8–11-year-olds. *Journal of Research in Music Education*, 42, 5–13. doi:10.2307/3345332
- Moore, R. S., Brotons, M., Fyk, J., & Castillo, A. (1997). Effects of culture, age, gender, and repeated trials on rote song learning skills of children 6-9 years old from England, Panama, Poland, Spain, and the United States. *Bulletin of the Council for Research in Music Education*, 133, 83–88.
- Moore, R. S., Chen, H. S., & Brotons, M. (2004). Pitch and interval accuracy in echo singing and xylophone playing by 8 and 10 year-old children from England, Spain, Taiwan, and USA. *Bulletin of the Council for Research in Music Education*, 161/162, 173–180.
- Moore, R. S., Fyk, J., Frega, A. L., & Brotons, M. (1995/1996). Influences of culture, age, gender and two-tone melodies on interval matching skills of children from Argentina, Poland, Spain and the USA. *Bulletin of the Council of Research in Music Education*, 127, 127–135.

- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364–386. doi:10.1177/1094428106291059
- National Core Arts Standards (2014). Retrieved from www.nationalartsstandards.org
- *Newell, M. K. (2013). *Effects of movement instruction on children's singing achievement scores* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3564842).
- Nichols, B. E. (2013). *Task-based variability in children's singing accuracy* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3588972)
- Niet, G., Tiemens, B., Lendemeijer, B., & Hutschemaekers, G. (2009). Music-assisted relaxation to improve sleep quality: Meta-analysis. *Journal of Advanced Nursing, 65*, 1356–1364. doi:10.1111/j.1365-2648.2009.04982.x
- Odom, D. R. (1991). *The effect of male vocal modeling on the singing skill development of older children's choirs utilizing Southern Baptist curriculum* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9212259)
- Orwin, R., & Vevea, J. (2009). Evaluating coding decisions. In H. Cooper, L. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 177–203). New York, NY: Russell Sage Foundation.
- *Paladino, M. M. (1991). *The effect of aural feedback on the singing accuracy and melodic perception of children in grades one and three* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9214672)

- *Paney, A. S., & Kay, A. C. (2014). Developing singing in third-grade music classrooms: The effect of a concurrent-feedback computer game on pitch-matching skills. *Update: Applications of Research in Music Education*. doi:10.1177/8755123314548047
- Patrick, L. R. (1978). *An investigation of the pitch-matching abilities of first grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 7917764)
- Pedersen, M. (1997). *Biological development and the normal voice in puberty* (Unpublished doctoral dissertation). University of Oulu, Finland. Retrieved from http://www.mpedersen.org/images/uploads/Thesis_in_English.pdf
- Pelletier, C. L. (2004). The effect of music on decreasing arousal due to stress: A meta-analysis. *Journal of Music Therapy*, 41, 192–214. doi:10.1093/jmt/41.3.192
- *Persellin, D. C. (2006). The effects of vocal modeling, musical aptitude, and home environment on pitch accuracy of young children. *Bulletin of the Council for Research in Music Education*, 169, 39–50.
- Petzold, R. G. (1963). The development of auditory perception of musical sounds by children in the first six grades. *Journal of Research in Music Education*, 17, 21–43.
- Petzold, R. G. (1969). Auditory perception by children. *Journal of Research in Music Education*, 17, 82–87.
- Pfordrescher, P. Q. (2007). Poor-pitch singing in the absence of “tone deafness.” *Music Perception*, 25, 95–115. Retrieved from <https://libproxy.library.unt.edu/login?url=http://search.proquest.com/docview/222220162?accountid=7113>

- Pfordrescher, P. Q., & Halpern, A. R. (2013). Auditory imagery and the poor-pitch singer. *Psychonomic Bulletin & Review*, 20, 747–753. doi:10.3758/s13423-013-0401-8
- Phillips, K. H. (1983). The effects of group breath control training on selected vocal measures related to the singing ability of elementary students in grades two, three, and four. (Doctoral dissertation). *Dissertation Abstracts International*, 44, 1017A.
- Phillips, K. H. (1985). The effects of group breath-control training on the singing ability of elementary students. *Journal of Research in Music Education*, 33, 179–191. doi:10.2307/3344805
- Phillips, K. H. (2014). *Teaching kids to sing*. Boston, MA: Schirmer.
- *Phillips, K. & Aitchison, R. (1997). Effects of psychomotor instruction on elementary general music students' singing performance. *Journal of Research in Music Education*, 45, 185–196. doi:10.2307/3345580
- Phillips, K. H. & Doneski, S. M. (2011). Research on elementary and secondary school singing. In R. Colwell & P. R. Webster (Eds.), *MENC handbook of research on music learning volume 2: Applications* (pp. 176–232). New York, NY: Oxford University Press.
- Phillips, K. H., Williams, J. W., & Edwin, R. (2012). The young singer. In G. McPherson & G. Welch (Eds.), *Oxford handbook of music education, volume 1*, (pp. 594–607). New York, NY: Oxford.
- Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition*, 105, 533–546. doi:10.1016/j.cognition.2006.11.006
- Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect-Shmozart effect: A meta-analysis. *Intelligence*, 38, 314–323. doi:10.1016/j.intell.2010.03.001

- Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual representation enhances the appreciation of music performance. *Music Perception*, 30, 71–83. Retrieved from <https://libproxy.library.unt.edu/login?url=http://search.proquest.com/docview/1041061271?accountid=7113>
- *Porter, S. Y. (1977). The effect of multiple discrimination training on pitch-matching behaviors of uncertain singers. *Journal of Research in Music Education*, 25, 68–82.
doi:10.2307/3344846
- Price, H., Yarbrough, C., Jones, M., & Moore, R. (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
- Radocy, R. E., & Boyle, J. D. (2003). *Psychological foundations of musical behavior*. Springfield, IL: Charles C Thomas.
- Ramsey, J. H. (1983). The effects of age, singing ability, and instrumental experiences on preschool children's melodic perception. *Journal of Research in Music Education*, 31, 133–145.
- Reifinger, F. L. (2009). An analysis of tonal patterns used for sight-singing instruction in second-grade general music class. *Journal of Research in Music Education*, 57, 203–216.
doi:10.1177/0022429409343099
- Reifinger, F. L. (2012). The acquisition of sight-singing skills in second-grade general music: Effects of using solfège and of relating tonal patterns to songs. *Journal of Research in Music Education*, 60, 26–42. doi:10.1177/0022429411435683

- *Richner, S. S. (1976). *The effect of classroom and remedial methods of music instruction on the ability of inaccurate singers, in the third, fourth, and fifth grades, to reproduce pitches* (Unpublished doctoral dissertation). University of Idaho, Moscow, ID.
- *Roberts, E., & Davies, A. (1975). Poor pitch singing: Response of monotone singers to a program of remedial training. *Journal of Research in Music Education*, 23, 227–239.
doi:10.2307/3344852
- Roberts, E., & Davies, A. (1976). A method of extending the vocal range of “monotone” schoolchildren. *Psychology of Music*, 4, 29–43. doi:10.1177/030573567641004
- Rohwer, D., & Svec, C. (2014). Perceived value of research preparation opportunities for future music education professors. *Update: Application of Research in Music Education*, 33(1), 57–64.
- Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, 11, 446–453. doi:10.1111/1467-9280.00287
- Runfola, M., Etopio, E., Hamlen, K., & Rozendal, M. (2012). Effect of music instruction on preschoolers’ music achievement and emergent literacy achievement. *Bulletin of the Council for Research in Music Education*, 192, 7–27.
doi:10.5406/bulcouresmusedu.192.0007
- *Rutkowski, J. (1986). *The effect of restricted song range on kindergarten children’s use of singing voice and developmental music aptitude* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8619357)
- Rutkowski, J. (1990). The measurement and evaluation of children’s singing voice development. *The Quarterly: Center for Research in Music Learning and Teaching* 1(1-2), 81–95.

- *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. doi:10.2307/3345447
- Rutkowski, J., & Snell Miller, M. (2002). A longitudinal study of elementary children's acquisition of their singing voices. *Update: Applications of Research in Music Education*, 22(1), 5–14. doi:10.1177/87551233020220010901
- *Rutkowski, J., & Snell Miller, M. (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.
- Sanchez-Meca, J., & Marin-Martinez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology*, 51, 311–326.
- Schmidt, C. P., & Zdzinski, S. F. (1993). Cited quantitative research articles in music education research journals, 1975–1990: A content analysis of selected studies. *Journal of Research in Music Education*, 41, 5–18.
- Seashore, C. E. (1941). *Why we love music*. Philadelphia, PA: Theodore Presser.
- Seltzer, G. (1983). *Music making*. Jefferson, NC: McFarland.
- Sergeant, D. C., & Welch, G. F. (2009). Gender differences in long-term average spectra of children's singing voices. *Journal of Voice*, 23, 319–336. Retrieved from <http://search.proquest.com/docview/742782186?accountid=7113>

- Siebenaler, D. (2008). Children's attitudes toward singing and song recordings related to gender, ethnicity, and age. *Update: Applications of Research in Music Education*, 27(1), 49–56.
doi:10.1177/8755123308322275
- Silveira, J. M., & Diaz, F. M. (2014). Student teaching in music: A content analysis of research journals in music education. *Update: Applications of Research in Music Education*, 23(2), 92–104. doi:10.1177/1057083713487215
- Silverman, M. J. (2003). The influence of music on the symptoms of psychosis: A meta-analysis. *Journal of Music Therapy*, 40, 27–40. doi:10.1093/jmt/40.1.27
- Simms, W.L., Moore, R. S., & Kuhn, R. 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.
- Small, A. R., & McCachern, F. L. (1983). The effect of male and female vocal modeling on pitch-matching accuracy of first-grade children. *Journal of Research in Music Education*, 31, 227–233. doi:10.2307/3345175
- Smith, R. S. (1973). *Factors related to children's in-tune singing abilities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 7411404)
- Standley, J. M. (1996a). A brief introduction to meta-analysis. *Journal of Research in Music Education*, 44, 101–104. doi:10.2307/3345664
- Standley, J. M. (1996b). A meta-analysis on the effects of music as reinforcement for education/therapy objectives. *Journal of Research in Music Education*, 44, 105–133.
doi:10.2307/3345665

Standley, J. M. (2002). A meta-analysis of the efficacy of music therapy for premature infants.

Journal of Pediatric Nursing, 17, 107–113. doi:10.1053/jpdn.2002.124128

Standley, J. M. (2008). Does music instruction help children learn to read? Evidence of a meta-analysis. *Update: Applications of Research in Music Education*, 27(1), 17-32.

doi:10.1177/8755123308322270

State Education Agency Directors of Arts Education (2013). *National coalition for core arts*

standards music model cornerstone assessment: General music grade 2. Retrieved from

http://www.nationalartsstandards.org/sites/default/files/Music_MCA_Grade_2_GenMus_Performing.pdf

State Education Agency Directors of Arts Education (2014a). *National core arts standards:*

General music/creating. Retrieved from http://www.nationalartsstandards.org/customize-handbook?bundle_name%5B1%5D=1&bundle_name%5B2%5D=2&bundle_name%5B3%5D=3&bundle_name%5B4%5D=4&bundle_name%5B5%5D=5&bundle_name%5B6%5D=6&bundle_name%5B7%5D=7&bundle_name%5B8%5D=8&bundle_name%5B9%5D=9&bundle_name%5B10%5D=10&sm_vid_Process%5B1%5D=1&sm_vid_Process%5B2%5D=2&sm_vid_Process%5B3%5D=3&sm_vid_Process%5B4%5D=4&ss_grade_level%5B1%5D=1&ss_grade_level%5B5%5D=5&ss_grade_level%5B6%5D=6&ss_grade_level%5B7%5D=7&ss_grade_level%5B8%5D=8&include%5Beu%2F%5D=eu%2F%5D=&include%5Bpv%5D=pv&include%5Bresources%5D=resources

State Education Agency Directors of Arts Education (2014b). *National core arts standards:*

National core arts anchor standards. Retrieved from

<http://www.nationalartsstandards.org/content/national-core-arts-standards-anchor-standards#creating>

- *Stauffer, S. L. (1985). *An investigation of the effects of melodic and harmonic context on the development of singing ability in primary grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 8520992)
- Steenwyk, L. A. (2004). *An investigation of the use of singing games in measuring singing voice development in young children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 1424742)
- Sterling, P. A. (1984). *A developmental study of the effects of accompanying harmonic context on the children's vocal pitch accuracy of familiar melodies* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI no. 8425896)
- Stevens, C. J., Keller, P. E., & Tyler, M. D. (2011). Tonal language background and detecting pitch contour in spoken and musical items. *Psychology of Music*, *41*, 59–74.
doi:10.1177/0305735611415749
- Swanwick, K. (1994). *Musical knowledge: Intuition, analysis, and music education*. New York, NY: Routledge.
- Tam, W., Wong, E., & Twinn, S. (2008). Effect of music on procedure time and sedation during colonoscopy: A meta-analysis. *World Journal of Gastroenterology*, *14*, 5336–5343.
doi:10.3748/wjg.14.5336
- Tatem, F. L. (1990). *Effects of selected timbres, tasks, grade level, and gender on vocal pitch-matching accuracy of kindergarten through third-grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI no. 9105941)
- Timmerman, L. M., Allen, M., Jorgensen, J., Herrett-Skjellum, J., Kramer, M. R., & Ryan, D. J. (2008). A review and meta-analysis examining the relationship of music content with sex,

- race, priming, and attitudes. *Communication Quarterly*, 53, 303–324.
doi:10.1080/1463370802240932
- Tremblay-Champoux, A., Bella, S. D., Phillips-Silver, J., Lebrun, M., & Peretz, I. (2010). Singing proficiency in congenital amusia: Imitation helps. *Cognitive Neuropsychology*, 27, 463–476. doi:10.1080/02643294.2011.567258
- Trollinger, V. L. (2003). Relationships between pitch-matching accuracy, speech fundamental frequency, speech range, age, and gender in American English-speaking preschool children. *Journal of Research in Music Education*, 51, 78–94. doi:10.2307/3345650
- Trollinger, V. L. (2004). Preschool children's pitch-matching accuracy in relation to participation in Cantonese-immersion preschools. *Journal of Research in Music Education*, 52, 218–233. doi:10.2307/3345856
- Tsang, C. D., & Conrad, N. J. (2011). Music training and reading readiness. *Music Perception*, 29, 157–163.
- Tsang, C. D., Friendly, R. H., & Trainor, L. J. (2012). Singing development as a sensorimotor interaction problem. *Psychomusicology: Music, Mind & Brain*, 21(1/2), 31–44.
doi:10.1037/h0094002
- Ueda, T., Suzukamo, Y., Sato, M., & Izumi, S. (2013). Effects of music therapy on behavioral and psychological symptoms of dementia: A systematic review and meta-analysis. *Ageing Research Reviews*, 12, 628–641. doi:10.1016/j.arr.2013.02.003
- Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 129–146). New York, NY: Russell Sage Foundation.

- Vasionyte, I., & Madison, G. (2013). Musical intervention for patients with dementia: A meta-analysis. *Journal of Clinical Nursing*, 22, 1203–1216. doi:10.1111/jocn.12166
- Wang, C. C., & Sogin, D. W. (1997). Self-reported versus observed classroom activities in elementary general music. *Journal of Research in Music Education*, 45, 444–456.
- Wang, C., Sun, Y., & Zang, H. (2014). Music therapy improves sleep quality in acute and chronic sleep disorders: A meta-analysis of 10 randomized studies. *International Journal of Nursing Studies*, 51, 51–62. doi:10.1016/j.jnurstu.2013.03.008
- Wassum, S. (1979). Elementary school children's vocal range. *Journal of Research in Music Education*, 27, 214–226. doi:10.2307/3344708
- Wassum, S. (1980). Elementary school children's concept of tonality. *Journal of Research in Music Education*, 28, 18–33. doi:10.2307/3344709
- Welch, G. (1979a). Poor pitch singing: A review of the literature. *Psychology of Music*, 7, 50–58. doi:10.1177/030573567971006
- Welch, G. (1979b). Vocal range and poor pitch singing. *Psychology of Music*, 7, 13–31.
- Welch, G. F. (1985). Variability of practice and knowledge of results as factors in learning to sing in tune. *Bulletin of the Council for Research in Music Education*, 85, 238–247.
- Welch, G. F. (1994). The assessment of singing. *Psychology of Music*, 22, 3–19.
- Welch, G. F. (1998). Early childhood musical development. *Research studies in music education*, 11, 27–41. doi:10.1177/1321103X9801100104
- Welch, G. (2000). The developing voice. In L. Thurman, & G. Welch (Eds.), *Bodymind and Voice* (pp. 704–717). Collegeville, MN: VoiceCare Network.

- Welch, G. F. (2015). Singing and vocal development. In G. McPherson (Ed.), *The child as musician: A handbook of musical development* (n.p.). New York, NY: Oxford University Press.
- *Welch, G. F., Howard, D. M., & Rush, C. (1989). Real-time visual feedback in the development of vocal pitch accuracy in singing. *Psychology of Music*, 17, 146–157.
doi:10.1177/0305735689172005
- Welch, G. F., Sergeant, D. C., & White, P. J. (1996). The singing competencies of five-year-old developing singers. *Bulletin of the Council for Research in Music Education*, 127, 155–162.
- Welch, G. F., Sergeant, D. C., & White, P. (1997). “Age, sex and vocal task as factors in singing “in-tune” during the first years of schooling.” *Proceedings, Sixteenth International Society for Music Education Research Seminar*, 15–22 July, Frascati, Italy. *Bulletin of the Council for Research in Music Education*, 133, 153–160.
- Welch, G. F., Sergeant, D. C., & White, P. J. (1998). The role of linguistic dominance in the acquisition of song. *Research Studies in Music Education*, 10, 67–74.
doi:10.1177/1321103X9801000106
- Welch, G. F., & White, P. (1993/1994). The developing voice: Education and vocal efficiency – A physical perspective. *Bulletin of the Council for Research in Music Education*, 119, 146–156.
- Wolf, D. L. (2003). *Tonal pattern and rhythm pattern hierarchies and the long-term importance of kindergarten music instruction* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 3112322)

- Wolf, D. L. (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.
- Wyatt, S. C. (1993). *The effect of group instruction in pitch accuracy and in head voice on the intonation and range of second-grade children* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Global database. (UMI No. 9322010)
- Young, W. (1974). Musical development in preschool disadvantaged children. *Journal of Research in Music Education*, 22, 155–169.
- Yarbrough, C. (1984). A content analysis of the “Journal of Research in Music Education”, 1953–1983. *Journal of Research in Music Education*, 32, 213–222.
- Yarbrough, C. (2002). The first 50 years of the Journal of Research in Music Education: A content analysis. *Journal of Research in Music Education*, 50, 276–279.
- 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:10.2307/3345773
- 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.
- Zhang, J., Wang, P., Yao, J., Zhao, L., Davis, M. P., Walsh, D., & Yue, G. H. (2012). Music interventions for psychological and physical outcomes in cancer: A systematic review and meta-analysis. *Support Care Cancer*, 20, 3043–3053. doi:10.1007/s00520-012-1606-5
- Note.* * denotes reference that was included in the meta-analysis