Task-based Variability in Children's Singing Accuracy

Bryan E. Nichols

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Steven M. Demorest, Chair

Patricia Shehan Campbell

Min Li

Steven J. Morrison

Program Authorized to Offer Degree:

School of Music

University of Washington

Abstract

Task-based Variability in Children's Singing Accuracy

Bryan E. Nichols

Chair of the Supervisory Committee:
Professor Steven M. Demorest
School of Music

The purpose of this study was to explore task-based variability in children's singing accuracy performance. The research questions were:

1. Does children's singing accuracy vary based on the nature of the singing assessment employed?
2. Is there a hierarchy of difficulty and discrimination ability among singing assessment tasks?
3. What is the interrelationship among different tasks and how few tasks might be employed in a comprehensive measure of accurate singing?

A 2 X 4 factorial design was used to examine the performance of 4th grade children (n = 120) in both solo and doubled response conditions. Every child sang four task types: solo pitch, interval, pattern, and the song *Jingle Bells*. To account for the effect of tonal memory, test items in all tasks were presented in four total pitches by an adult female

vocal model. Each task type contained five items, and the fifth item replicated the first to provide a measure of stability. Scoring was done by the researcher and one other judge with high reliability. Pitch matching was scored dichotomously and song singing was scored using an eight-point scale. Data were transformed to a 0-1 scale to express difficulty and discrimination indices.

The results indicated that there was significant task-based variability in children's singing accuracy. Difficulty levels varied by task type, with patterns and songs indicating lower performance than single pitches and intervals. Performance was significantly higher for all tasks in the doubled condition than in the solo condition, and a significant interaction indicated task-based performance varied by response mode. Students who indicated a history of private lessons (n = 54) evidenced significantly higher performance than those without.

An exploratory factor analysis demonstrated that all tasks load onto one factor. Internal reliability was satisfactory, and the results suggest that a minimum of three items can be included in each task in future research for a reliability coefficient of .75, or a minimum of four items for a coefficient greater than .80. These singing tasks were significantly inter-correlated, and the easiest item was also lowest in range. Vocal scooping and its implications for singing accuracy assessment were discussed.

The three main findings were that doubled singing was more accurate than solo singing, summative assessment should include as many task types as is feasible, and within- and between-student performance should be compared using the same task type. Future research should explore variables affecting differing variability between lower- and higher-performing singers. Additionally, there is a relationship between

history of private lessons and singing accuracy, and motivation and general musical experience should be explored as mediating or moderating variables so that teachers can best encourage student development. Last, it remains possible that doubled singing primes students to perform better in solo singing. Future research should examine under which conditions this may be true so that teachers can better target remediation for the majority of students who sing better when doubled by another voice.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgments

This project is the result of the combination of support I received from my parents Henry and Jo Nell Nichols and my sister Christy, who encouraged my musical participation at an early age. Thanks to Joe and Brooksy Carrington for providing my first piano: a Kincaid spinet they passed down when I was in elementary school. Thanks to Granny Nichols for loving to sing and serving as my favorite "vocal model." Thanks to chorus teacher Shirley Jones and band director Bruce Soderstrom who helped guide my early teaching far more than they realize.

Thanks to the Seattle area students who agreed to participate in this study. Christopher Roberts decided my research questions were worthwhile enough to allow his own students to participate in this study, and led me to other teachers and schools in the area. This study could not have been done without his and their support. Thanks to Alison Farley for her help on this project and encouraging messages of support. Thanks also to Eliza, Luke, Erin, Karen and Ben, Aaron and Jasmine, Miah and Leslie, and Tyler and Jenn for bringing out the best in me during these Seattle years.

In her dissertation acknowledgments, Mary Goetze says, "the spouses of all dissertation authors deserve special mention, for inevitably these efforts become 'family affairs.'" Thanks to Jonny K. Miller for his constant support at home and for agreeing to move to Seattle when the time was just right. His well-timed questions (and cookies) helped me to be successful.

Thanks to Cecilia Wang and Judy Bowers, who convinced me I was capable of more graduate study, and for encouraging me to venture as far away from home as I wanted. Finally, thanks to Steven Demorest for his mentorship and for pushing me to express myself on paper more clearly. Along with Patricia Campbell and Steven Morrison, these three professors have forever changed my understanding of "what's important."

Dedication

To Gibbie Horsley and the students of Meade County, Kentucky.

Chapter 1: Introduction

Singing (in)accuracy is one of the biggest problems vocal music teachers face in the classroom (Goetze & Cooper, 1990). Teachers can attest to the idiosyncratic nature of this ability to match pitches accurately, as some children seem to have trouble coordinating their voice to match what they hear, while others do not (Rutkowski & Barnes, 2000). Students who experience singing problems may become timid singers or dislike singing altogether. Since self-perception and positive feedback have been linked to future musical participation (Clements, 2002), it is critical that teachers have ways to assess and encourage young children's singing development.

Researchers in psychology have examined singing accuracy in normal (Dalla Bella & Berkowska, 2009; Dalla Bella, Giguère, & Peretz, 2007; Pfordresher et al., 2010) and deficient (Berkowska & Dalla Bella, 2009; Bradshaw & McHenry, 2005; Cuddy, 2005; Martinez-Castilla & Sotillo, 2008; Sloboda, 2005) adult samples, and Pfordresher and Brown (2007) proposed four causes of "poor pitch" singing. First, they suggested that deficits in singing are caused by a motor control problem: some individuals have not coordinated their voices to correspond with what they hear. Second, the problem may exist more often as a perception issue, whereby individuals cannot translate sounds into pitches, perhaps indicating neuropsychological or cognitive dysfunction. A third possibility is an imitative or sensorimotor deficit, which is a combination of motor control problems and perception issues. Some other reports have also referred to this as a problem in the *vocal sensorimotor loop* (Berkowska & Dalla Bella, 2009; Tsang, Friendly, & Trainor, 2011). Fourth, it also remains possible that

individuals perform poorly due to memory deficits, whereby a lack of musical memory leads to the inability to recall pitches even shortly after they are given.

While there may be several causes of inaccuracy, music education researchers have been primarily concerned with the assessment and remediation of this skill. Indeed, researchers have studied the remediation of component skills for singing with some success in improvement (e.g., breath control in Phillips & Aitchison, 1997). Singing accuracy and melodic perception skills have been improved in selected samples of both “normal” and “poor pitch” students (Apfelstadt, 1984), and certain abilities like pitch matching have been shown to improve across the school year while others like song singing do not (Demorest & Nichols, 2012; Welch, Sergeant, & White, 1997).

A music teacher equipped with a useful measure can make sound pedagogical decisions and execute accurate academic reporting. Measures of singing accuracy have been put forth by researchers for use with various singing tasks (Salvador, 2010). Such measures indicate students’ ability by placing them along a scale, or by expressing individual ability relative to other students. The resulting data are often used as interval data, though these scales may not actually represent equal intervals. Another term, rubric, generally indicates lack of intervallic data, though like scales, rubrics are almost always associated with a scoring scale based on whole numbers.

Teachers can and do design their own assessments, or they use researcher-designed and validated measures or scales. Researcher-designed measures can be used in the classroom when they are made available by way of methods or techniques books and are matched to the teacher’s needs. These kinds of measures are useful

because they have been validated for certain grades; that is, they are shown to represent certain characteristics for certain ages.

The term assessment is used in this paper to refer to the general evaluation of student ability or to describe a specific instrument for testing skill. To create a valid assessment, an instrument must be shown to indicate the ability it is intended to test (validity), and it must be shown to provide reliable scores. This has proven challenging for singing accuracy partly because task-based variability has not been explored in detail; that is, the specific task used for assessing singing skill may itself have an effect on the outcome. A complicating factor is that certain intervals are more difficult than others, and overall, intervals may be more or less difficult to sing than single pitches or patterns of varying lengths. Similarly, some songs are more difficult than others, and yet another factor like singing alone or doubled may differentially affect performance of songs or other tasks.

Factors in Assessment

The remainder of this chapter is used to outline the decisions teachers must make when assessing singing in their classrooms. The test design depends on the purpose of assessment. If the purpose is to examine children's voice development, the appropriate tasks and scales for that construct should be used, such as the Singing Voice Development Measure (SVDM, Rutkowski, 1990). That measure is intended for the assessment of developmental characteristics like range and use of singing versus speaking voice, which Rutkowski argues must be established before measuring for accuracy. That measure has been shown to provide reliable score interpretations in grades 1-5 (Levinowitz et al., 1998). If, alternatively, the teacher's purpose is to

examine singing *accuracy* – the ability to sing in tune, the appropriate tasks and scales should be considered. It is this last construct of in-tune singing that is the focus of this study.

For teachers designing their own tests, there are many important decisions based on the purpose of the testing, including the type of task, key, range, text, singing alone or doubled, and scoring (Figure 1.1). First, the specific task(s) must be chosen. Single pitches, intervals, patterns and song singing have been shown to be

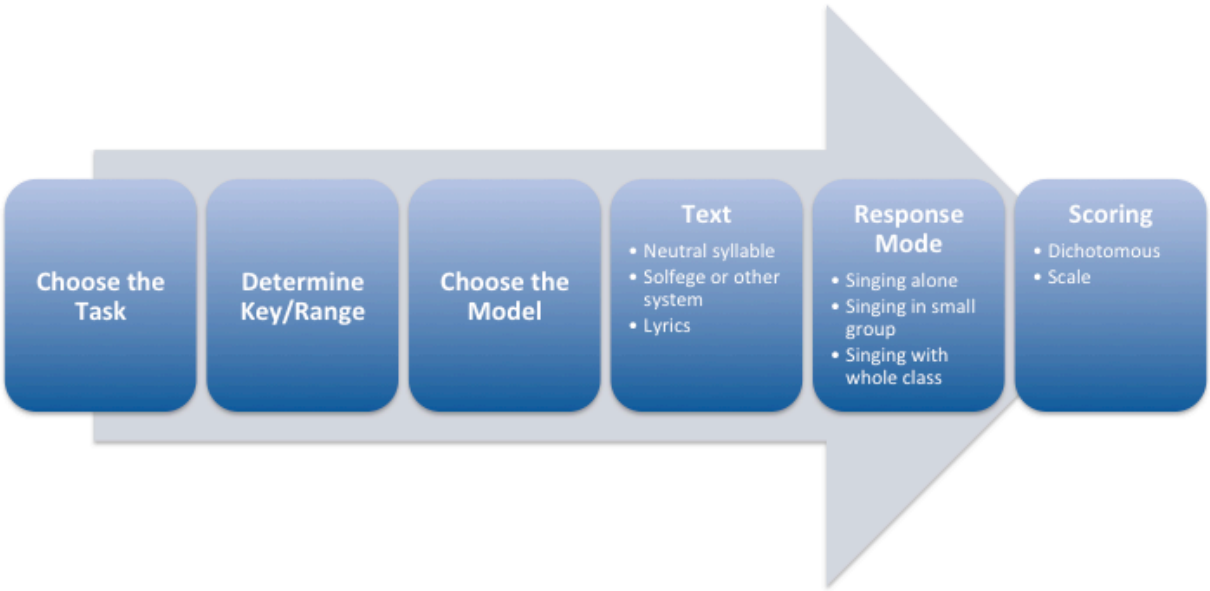


Figure 1.1. Flow of assessment choices. This figure demonstrates a teacher's *a priori* decision-making for evaluation of student singing accuracy.

discriminators of accuracy (Roberts & Davies, 1975). Usually the briefest task is most feasible, and the teacher can be expected to choose as few tasks (and items within those task types) as necessary.

If the purpose of student evaluation is formative assessment, and the teacher wishes to know which students can accurately sing a specific interval from concert repertoire so that (s)he can design the next day's lesson plan, the teacher may choose short phrases from the current repertoire. Likewise, a teacher may wish to know which patterns a child has mastered and the teacher would present several recently-practiced patterns for assessment. For song singing, any song could be used as a criterion for tuneful singing, but performance on any one song may not be representative of a student's overall ability. A teacher who is asked to present a summative assessment for grading may wish to use several tasks. For instance, a teacher could use two songs, or a pattern task and a song, since some songs are more difficult than others based on factors like range, specific intervals, tonality and other features, and thus represent singing ability differently (Wolf, 2005).

The range used in pitch sequences and melodies is a critical consideration for teachers who must determine in what key the assessment task should be presented. If a song from class repertoire is used, the key selected is likely to be the key selected for use in class. As grade levels increase and changing voices are present, or in the case of ensembles based on multi-part music distributed among voice types such as soprano, alto, tenor and bass, several keys must be presented if the same pitch sequence is used. If the key is chosen at the time of testing and tailored to each individual student, the testing time increases unless the teacher knows each unique

voice already. The teacher will know whether the selected range is too high or low if the student can sing some of the pitches but not others, or if the student sings in a higher or lower octave. The astute teacher can differentiate between students who choose to do this and those who switch because they must (Hedden, 2012).

Moore (1991) reported the singing range of 8-11 year olds as approximately two octaves from G3 to high G5-sharp. Singing such a large range requires changes of register, from chest voice to head voice, which can be difficult, and not everyone can use all parts of the possible voice range. Rutkowski (1990) described the varying degrees of range and singing voice use as *singing voice development*. This construct was first described in five levels as 1) Pre-singer (chants), 2) Speaking range singer, 3) Uncertain singer, 4) Initial range singer, and 5) Singer. It is the fifth level that represents the student who can use all parts of the singing voice range, below and above the register "lift."

Once the range and register have been chosen, it must be decided how test items will be presented to students. Teachers sometimes ask students to echo pitch sequences by call-and-response vocal modeling. For elementary-aged students, male teachers must decide whether to sing in the falsetto register or to sing in the male timbre, which may affect the results (Price, Yarbrough, Jones, & Moore, 1994; Sims, Moore, & Kuhn, 1982). Teachers can also have students respond to pitch sequences from a piano or other instrument. Although children have been shown to respond best to another child's voice, and after that an adult female's voice (Green, 1990), it is difficult to administer classroom assessments for echo tasks using a child's voice. Using song

tasks from memory eliminates this concern, though the choice still must be made to establish a key signature (above), or to allow the student to self-select a starting pitch.

Next, the teacher must decide whether to ask students to sing using text or a neutral syllable, a decision often based on the role of text in the task used. If memorization is important to the purpose for testing, text may be included. For testing single pitches or intervals, the use of text may be unnecessary. If an echo task is conceptualized as a pitch matching task, neutral syllables may be used. If an echo task is conceptualized as a phrase singing task, song text may be used. If the purpose for assessment does not prescribe the use of text, the teacher could rely on the research in this area to suggest the mode that is chosen. Unfortunately, the effect of text on accuracy is mixed; some suggest the use of text sometimes may elicit better performance (Gault, 2002), but others suggest no difference (Levinowitz, 1987; Sims et al., 1982; Smale, 1987), or that they are more accurate on neutral syllables (Goetze, 1985).

Another area of conflicting research is whether testing students individually, rather than in groups when the singer's voice is doubled by others, would be the best way to evaluate student performance. Since many teachers see students only once a week or teach very large ensemble classes, one-on-one assessment may be impractical. For these teachers, evaluating students while they perform as a part of a group is the logical solution to these time constraints. In this scenario, students sing with an added stimulus: the voices of their peers. If students are being tested based on their performance in an ensemble, doubled singing may be more ecologically valid since it represents the type of singing the test is intended to replicate. If determining

individual student progress is the purpose, singing alone may be a more valid response mode. Unfortunately, there are mixed results on whether students sing more accurately individually or with others, but the chosen response mode is dependent on the purpose for assessment.

Finally, before testing students, the teacher must decide whether the sung responses are to be scored dichotomously or using a rubric or scale. Dichotomous scoring simply requires a cut-off for the consideration of "in-tune" and is usually scored by the teacher's perception of the pitch at the time of testing. Researchers sometimes record sung responses for later scoring by one or more judges; alternatively, researchers can measure the actual cent deviation of the sung pitch from the given pitch as a way to specify the degree and direction of the pitch response. Teachers, however, operate under great time constraints and are likely to make scoring decisions during test administration. They must choose dichotomous scoring or choose a scale, many of which are viable for specific kinds of tasks or grade levels, though teachers also commonly design their own rubrics that are specific to the purpose.

Conclusion

In this study, singing accuracy refers to the degree to which a sung pitch matches a given (model) pitch, or pitches from memory. Accuracy in this definition does not include anything about students' perceptual skills, the tone quality of their voice, vocal register development, or other aspects of vocal technique. While several of these may be related to accurate singing, the goal of this study was simply to explore the best means of measuring singing accuracy. A review of singing accuracy research must examine studies that address all those factors which affect accuracy measures: stimulus

presentation characteristics, such as the effect of the vocal model used, and the content of tests, including the specific tasks that have been used for accuracy measurement. Types of tasks include pitch matching tasks of differing length and complexity and song singing tasks. Comparisons of these will be discussed, where possible. Additionally, the conditions under which these tasks are presented are important considerations. Some reports include analyses of the difficulty of specific items on tests, or on the ability of test item performance to discriminate between singers' abilities. All these findings represent important factors researchers must consider when constructing tests of singing accuracy.

Singing accuracy assessment is a challenging area of music assessment, given the many factors involved in testing this skill. While various singing tasks have been shown to be useful in singing assessment, it remains unclear which tasks are necessary for general assessment, which are most difficult, and how many items within those task types may be required. Additionally, singing accuracy performance may be different when the subject sings alone versus along with another voice. There is conflicting evidence on doubled singing (Goetze & Cooper, 1990) or the role of aural feedback in singing performance (Rutkowski & Miller, 2003), referred to here as the response mode, which may have an important interaction with the type of task used. A better understanding of assessment in these two response modes combined with an understanding of task-based variability can help teachers design useful classroom assessments that address content standards as they exist in local schools and states. As test design becomes better understood, music teachers can more confidently

implement formal assessments of singing than may currently occur in classrooms (Salvador, 2010).

Inaccurate singing among students, which may have several possible causes, presents a problem that music educators have a keen interest in identifying and improving. In order to do so, teachers need to be able to properly evaluate students, whether using their own or others' tests and measures for students singing alone or doubled by others. These evaluations should incorporate a task or a set of tasks that best represent students' abilities based on the purpose for assessment. The task-based variability of singing accuracy across a range of task types is still unknown, and an understanding of this variability would benefit both researchers and teachers of singing. It would be helpful to understand the different conditions under which children might sing more or less accurately so that teachers can design appropriate instructional responses.

Chapter 2: Review of Literature

Researchers in music education have created unique measures of singing (Salvador, 2010) and have tested many different samples across the elementary grades in order to describe properties like singing range, accuracy, and other features (Hedden, 2012). Children have demonstrated widely varying competencies, from 9-38% of children deemed "inaccurate" (Goetze & Cooper, 1990) to 75-90% in one sample defined as "non-singers" (Levinowitz et al., 1998). However, these figures reflect samples of students that may be dissimilar and tasks that vary in difficulty. Additionally, different models of evaluation are applied to these assessments, which in turn further obscure the true percentage of the student population with undeveloped or under-developed singing skills. In other words, the reported prevalence of accurate singing depends wholly on the nature of the assessment (item + measure) and the definition of accurate singing.

As indicated in the previous chapter, there are two main areas of study that are relevant to singing accuracy assessment. The first is the effect of how singing tasks are presented to students and under which conditions they are asked to respond. The second area is the effect of the specific type of task and items used. This review will focus on studies from the music education literature, which sample school-age children for participation. At times, references will be made to studies outside music education with adult samples where those findings can inform this review.

Task Presentation Variables

Students are asked to sing in music classroom and rehearsal situations in a variety of ways. Specific pitch sequences (referred to as specific "items" when used in

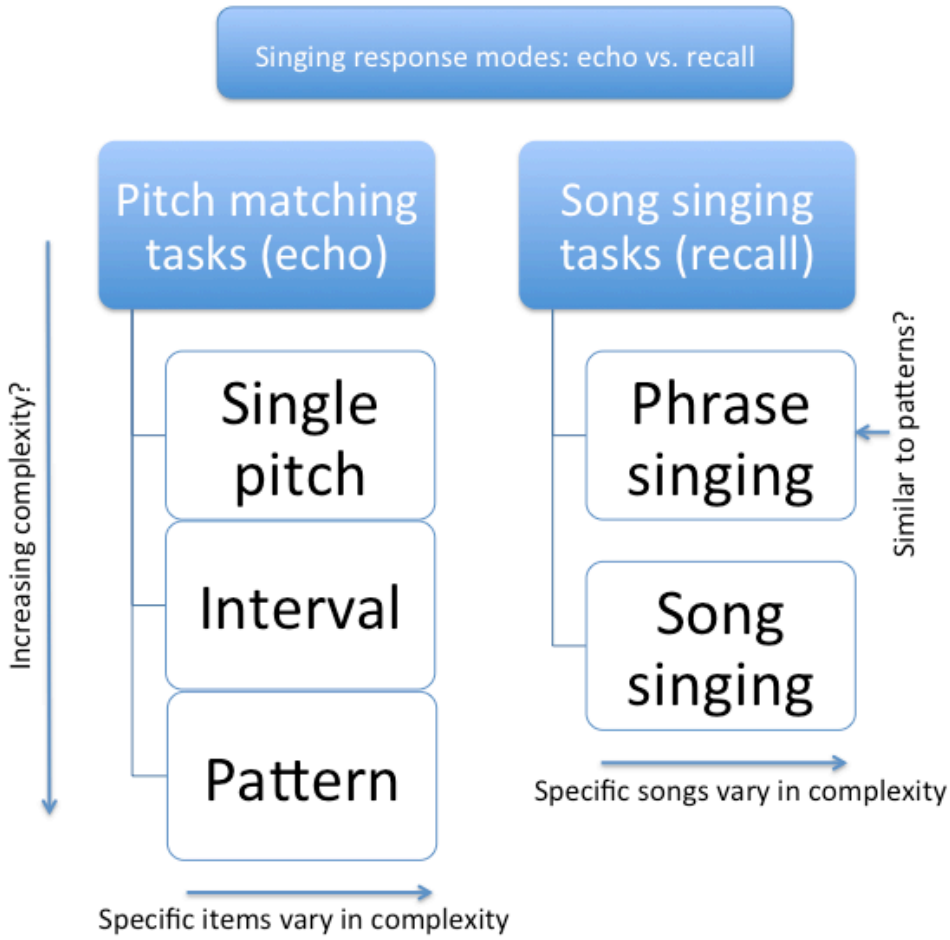


Figure 2.1. Singing response modes. This figure illustrates the types and relationships of singing tasks.

an assessment) can be grouped into types of tasks in two response modes, one in which subjects "echo" a given pitch(es) and another in which students recall melodic material (Figure 2.1). For assessment, these response modes are represented by pitch matching tasks and song singing tasks, respectively. Within pitch matching, students can be given a single pitch, an interval, or patterns of varying lengths. For song singing, students can be asked to recall songs or fragments of songs, like phrases. Before discussing and comparing these tasks in the second part of this review, however, it is

first important to consider the way in which pitch sequences are presented to subjects, including considerations such as model characteristics and contextualized stimuli, as well as varied response modes like solo and doubled singing. Since no significant effect for the presence of accompaniment has been reported, it is not included as a consideration here (Atterbury & Silcox, 1993; Guilbault, 2004).

A number of researchers have explored the role of model characteristics in sung performance, and these studies can inform the process of designing a study of task-based variability. Children have responded more accurately to a female model than a male model (Yarbrough, Green, Benson, & Bowers, 1991). Children also respond to another child's voice best, then to an adult female's voice, then to an adult male's voice (Green, 1990). Overall, children sing more accurately when less vibrato is used (Yarbrough, Bowers, & Benson, 1992). Green (1990) presented the same stimuli in all three conditions (child's voice, adult female, adult male) in order to compare the conditions. Yarbrough et al et al. (1992) presented the same stimuli in three conditions also (child's voice, female with vibrato, female with minimal vibrato). Both studies used a 2 X 3 condition to isolate variables for direct comparison, which could be effective for future research comparing similar conditions.

For classroom applications, rather than research use, the possibilities for types of vocal model vary as much as do teachers and instruments. For example, male timbre, falsetto, and sine-wave models have also been compared to demonstrate that male stimuli may be more effective than sine waves for both boys and girls (Price et al., 1994). Overall, girls responded better to higher stimuli, whereas boys responded better to lower stimuli. Additionally, the octave of the stimuli affected the octave of the

response. Boys may need help in understanding whether they are expected to sing at or below (octave displacement) a high pitch given from a female model.

Two studies reported that students were more adept echoing a male falsetto voice than a male timbre (Price et al., 1994; Yarbrough, Morrison, Karrick, and Dunn, 1995), demonstrating that male teachers must choose in what octave they sing to best help their students. Otherwise, teachers have few choices: they can use their own voice, use their students as models for one another, or use a common instrument like the piano, knowing that these will affect singing measures. Generally, children respond best when the stimuli is presented in their register (Kramer, 1985/1986; Sims et al., 1982). For singing assessment in research or in classrooms, all these options exist as possibilities for test stimuli. In sum, studies with children now generally present stimuli with an adult female's voice instructed to employ minimal vibrato.

The characteristics of the stimuli presented to subjects are important for pitch matching performance, but so are the conditions under which subjects are asked to respond. Green (1994) reported that students performed better when singing in groups of eight than when singing alone, which suggests group performance elicits higher accuracy than individual performance. However, classrooms and choral rehearsals normally present a larger group size than eight people, and it is difficult to control for the social effects between individual performance, small-group performance, and large-group performance. A notable contrast to Green (1994) is a report by Joyner (1969), who examined students described as "monotones" based on their performance in a group setting. Joyner found each student to sing in-tune once they sang alone. Further, Cooper (1995) and Smith (1973) found no difference in accuracy performance when

students sang alone or along with the researcher. Others have found individual performance to be higher than “unison” singing (Goetze, 1985; Smale, 1987). Thus, the results are mixed as to the effect of doubling, and more research is needed on doubling as a test condition.

The differences between the above-cited studies may be due in part to the specific nature of the doubling used. Some studies had children singing in a small group of three, five, or eight students and sometimes also the researcher. One study used a recorded child model for doubling and another used the researcher only. Possibly, social effects account for differences in singing, as some children may perform differently in front of a few or more peers or in front of the researcher only. Additionally, some of their peers may not have sung as accurately as would a pre-recorded model, which could affect tuneful singing.

In addition to variation in doubling standards, these studies also used a variety of songs, or instead a short phrase was used. One study used a four-beat pattern on “loo” (Cooper, 1995). Some were scored by judges, and one study weighted equally two scores of accurate pitches and accurate intervals (regardless of starting pitch). Most studies used a design wherein the researcher taught a the pattern, phrase, or song to students but they varied in the number of repetitions and classroom visits, and thus also familiarity with the researcher which could affect student performance at the time of testing. The only variable these studies had in common was the range used for stimuli. With one exception, a range of no more than a fifth was used. Green (1994) used a range of a sixth and all the studies used only notes between C4 and B4.

Wise and Sloboda (2008) studied the effect of an “accompaniment” condition in adults. For pitch matching, a male or female vocal model was used depending on the participant and the vocal stimuli were re-played for participants to sing along with. For song singing, a piano was used for doubling rather than a vocal model; the researcher played three octaves of the melody along with the participant, who sang *Happy Birthday* preceded by a song of their choice. Pfordresher and Brown (2007) described this variable in terms of auditory feedback and used three conditions in a sample of adults: normal, augmented, and masked. For normal singing, participants sang tasks that were heard on headphones. For the augmented feedback condition, participants heard the stimuli played again on headphones while they sang. For the masked feedback condition, pink noise was played through the headphones.

The studies cited here define simultaneous stimuli in different ways. Some refer to unison singing, which can be conceptualized as singers singing the same notes. Others use the term doubling, which can be conceptualized as singing along with an external stimulus. Wise and Sloboda (2008) use the term accompaniment, can be thought of as being *followed* or *supported* by another musical voice or instrument. Those authors also used the term synchronized condition to refer to what music education researchers refer to as “doubled” or “unison” singing. Pfordresher and Brown (2007) were exploring the role of external stimuli as auditory feedback to determine how singing was affected by simultaneous sounds versus pink noise.

Contextual Presentation

In addition to model characteristics and response mode, the tonal context can also influence singing performance. Researchers have designed contextually

presented pitch matching stimuli for singing research. Geringer (1983) established context for test items by asking children to sing back the final pitch in a simple three-measure song. Demorest and Clements (2007) adopted a similar procedure, and for comparison they employed a single pitch matching condition that replicated the typical classroom exercise in which the teacher sings a single note and the student attempts to sing it back. This latter pitch condition is also most similar to task demands in classrooms and choral rehearsals, and thus represents the most common kind of pitch matching task. In this study, singers classified as "certain" and "uncertain" singers differed from each other but did not differ significantly in their responses to the two conditions. Those singers classified as "inconsistent" (neither certain or uncertain), however, did differ. These inconsistent singers were between the certain and uncertain groups in terms of pitch matching ability. The authors report that if only the contextual condition were presented, the singers in this group would have been classified as good singers.

Both studies discussed in the previous paragraph were designed to investigate the relationship between pitch perception and pitch production in addition to stimuli presentation, so it is appropriate to address the relationship between these two variables. Just as the contextual presentation of pitches may affect the outcome, so can one's ability to *hear* and *discriminate* pitches. For example, Demorest (2001) found that unlike inconsistent singers, certain and uncertain singers performed differently on a measure of pitch discrimination, which could suggest that pitch discrimination ability plays a role in accurate pitch production. Geringer (1983) found no significant differences in pitch discrimination based on children's pitch matching ability, though the

tasks were relatively simple. Demorest and Clements (2007) reported that the results in these studies may vary due to the relative difficulty of items used, but that most studies have not directly compared tasks. They recommended further examination of task-based *variability* in singing accuracy.

Song singing can be described as contextualized because song presentation comes with additional information that is absent in single-pitch matching exercises. For example, songs provide cues like key, meter, and contour, which may support the developing student in the act of singing. This difference in context is important since pitch matching and song singing tasks are difficult to compare for these reasons.

Range

Children tend to select lower, rather than higher keys when singing, but when given a higher register, children can evidence wide ranges (Moore, 1991). Still, children often modulate keys in order to use pitches that may be easy or comfortable for them, and children show greater ranges when echoing pitch matching tasks than when singing songs from memory (Flowers, 1990). Rutkowski (1990) described range ability in terms of development, from a student with no use of the singing voice to a student with full use of an expanded range: 1) Does not sing but chants text, 2) Sustains tones with sensitivity to pitch, usually A2 to C3, 3) Wavers between speaking/singing, usually up to F3, 4) Has initial singing range, D3 to A3, and 5) Has full singing range, beyond register lift near B-flat-3 and beyond. Rutkowski (1996) then added levels between these five intervals to describe more specifically the developing singing voice. Wolf (2005) reported that test items that are low in range are easier than test items that are high in range, which corroborates Rutkowski's suggestion that range develops upwards.

Song lyrics could be theorized to add complexity to the task demands in singing and may be a component which takes priority over accurate singing. For example, the first level of the Singing Voice Development Measure (Rutkowski, 1996) states that a student "...does not sing but chants song text." Another scale for singing accuracy describes the first level the same way, a focus on text rather than melody (Welch, Sergeant, & White, 1995). Goetze (1985) found the neutral syllable "loo" to be easier to sing than text, and more so for the kindergarteners and first graders than the second and third graders in her sample. Smale (1987) also compared the use of text to the syllable "loo" but found no difference in performance using a sample of preschoolers.

Singing Tasks Used in Assessment

A recent review of literature suggests there is evidence supporting the use of "tonal patterns" and familiar songs for singing accuracy assessment tasks (Phillips & Doneski, 2011), but it remains unclear how these tasks differ when they are used for assessment. Pitch matching and song singing have been used extensively and sometimes combined in assessments, but when this occurs, little comparison is made. The single pitch task is briefest and has been used for singing accuracy assessment (Demorest, 2001; Demorest & Clements, 2007; Flowers & Dunn-Sousa, 1990; Geringer, 1983; Jones, 1971; Murry & Zwirner, 1991; Porter, 1977; Roberts & Davies, 1975). The choice of intervals as test items in singing assessments has been popular both as the only task (Green, 1990; Moore, Fyk, Frega, & Brotons, 1995; Price et al., 1994; Yarbrough et al., 1991; Yarbrough et al., 1992) and as a part of a larger assessment (Flowers & Dunn-Sousa, 1990; Joyner, 1969; Klemish, 1974; Roberts & Davies, 1975). Melodic patterns using three or more notes have also been used exclusively to inform

singing accuracy performance (Cooper, 1995; Phillips & Aitchison, 1997; 1999; Phillips, Aitchison, & Nompula, 2002; Rutkowski, 1990; Rutkowski, 1996) or as a part of a larger assessment (Flowers & Dunn-Sousa, 1990; Guerrini, 2006; Welch et al., 1997). To measure singing voice development, a separate construct from singing accuracy that primarily includes vocal features like voice range, researchers have used one-measure song fragments (Levinowitz et al., 1998; Rutkowski, 1983; Rutkowski, 1990; Rutkowski, 1996; Rutkowski & Miller, 2003).

Researchers have also chosen specific songs based on familiarity to their subjects and/or song simplicity. Play songs and folk songs have been the test stimuli for many children's studies (Apfelstadt, 1984; Brophy, 1997; Flowers & Dunn-Sousa, 1990; Gault, 2002; Goetze, 1985; Green, 1994; Guerrini, 2006; Guilbault, 2004; Joyner, 1969; Muse, 1993; Roberts & Davies, 1975; Smith, 1973; Welch et al., 1997; Western, 2002; Wurgler, 1990; Young, 1971). Studies of singing in adults have used similar songs, such as *Twinkle, Twinkle Little Star* (Pfordresher, Brown, Meier, Belyk, & Liotti, 2010), *Row-Row-Row Your Boat* (McCoy, 1997; Pfordresher et al., 2010), *Jingle Bells* (Berkowska & Dalla Bella, 2009; Pfordresher et al., 2010), *Brother John* (Berkowska & Dalla Bella, 2009), *Gens du Pays*, Quebecois birthday song (Dalla Bella, Giguère, & Peretz, 2007), *Happy Birthday* (Pfordresher et al., 2010), a Polish happy birthday song (Dalla Bella & Berkowska, 2009), *America* (Phillips & Vispoel, 1990; Watts, Moore, & McCaghren, 2005), and *O Music* (Phillips & Vispoel, 1990).

Comparisons of pitch matching and song singing tasks are few, but they do exist. Five-year-olds have been shown to perform better on single pitch matching, glides, and patterns than song singing (Welch et al., 1995). These young singers were evaluated

on a seven-point scale that was applied in the same way to each of the tasks. The authors did not publish the measure, which would have allowed the reader to understand how such a scale could be applied to the single pitch matching condition. Students in that sample performed single pitches better than patterns. In a similar study including older students, sung patterns were again demonstrated to be easier than songs in 5, 6, and 7-year-olds (Welch et al., 1997), and the authors suggested that accuracy is task-specific.

Demorest and Clements (2007) and Welch et al., (1997) have suggested that a) performance in singing accuracy is task-based and b) future research is warranted because there are important classroom and research implications for this possibility. Singing accuracy assessment depends on a thorough understanding of how performance differs based on the task given and under which conditions students respond most accurately. Once this is understood, remediation for poor singers can be refined by the specific transfer of skills from easy tasks to more difficult tasks, especially since these varying tasks may be inter-correlated.

Performance on pitch matching and song singing has been found to be moderately correlated in a sample of kindergarteners (Demorest & Nichols, 2012) and highly correlated in a sample of adults (Pfordresher et al., 2010). Pfordresher and Brown (2007) suggested that single pitch, interval, and pattern tasks vary in terms of complexity because of the number of unique pitches they contain and varied in memory demand because of the total number of notes that they contain. For their study, they presented the three tasks using four notes for every task to equalize memory demands. The authors presented the single pitch task in four (identical) notes, the interval pitches

in four notes (two for the first pitch and two for the second pitch), and the pattern in four unique pitches. Performance on the three pitch matching tasks varied for participants in this adult sample. For participants deemed "good" singers, performance improved across these three tasks. For "poor" singers, however, performance decreased across these tasks, suggesting that good singers may benefit from the added contextual information, whereas poor singers become more and more overwhelmed with additional complexity. Contrastingly, Wise and Sloboda (2008) found performance of "tone deaf" and "not tone deaf" groups to differ significantly from each other, but the performance of both groups across the three tasks (single pitches, intervals, patterns) was similiar in that all adults' performance decreased across tasks of additional complexity. Their pitch matching items varied by number of total notes (they did not control for memory effect by presenting the same number of notes in each pitch matching task type).

Unlike task complexity, familiarity with the pitch sequence may not have an effect on singing accuracy performance. Guerrini (2006) had 4th and 5th grade students sing patterns followed by the songs *America* and *Path to the Moon*. *America* was chosen as a familiar song for the students, and the second song was chosen as an unfamiliar song. Student responses were significantly higher to the pattern test than to the song singing tasks, which were not significantly different (students performed both the familiar and unfamiliar songs similarly). Guerrini interpreted these results to mean that students can sing patterns accurately before they can sing whole songs accurately and makes three conclusions related to singing development:

1. Many students learn to sing accurately by expanding their ranges first;

2. After the range has been expanded, many students become more vocally accurate by singing short patterns, but may not be able to apply this skill to entire songs;
3. Once students are able to sing one song accurately, using notes above the lift, they appear to be able to sing other songs accurately (p. 167).

Van Zee (1984) also had students sing patterns and a song, *Hot Cross Buns*, in an effort to explore remediation techniques for first-graders. After an analysis and description of individual students, she concluded that the ability to "match tonal patterns does not always carry over to singing the songs from which they were taken. A tonal memory needs to be developed for longer phrases" (p. 70). Van Zee suggested tonal memory plays a role, but more importantly that pitch matching exercises like patterns may not always correlate to functional singing test items like song singing.

Trying a different approach, Apfelstadt (1984) set out to test the effect of melodic perception instruction in kindergarteners. Her assessment included the song *Jingle Bells* in addition to patterns. Pitch matching performance was significantly improved after a period of instruction, but song singing was not. The author did not conjecture how the nature of pitch matching and song singing differ; however, these results combined with others described in this review lead to an understanding that pitch matching tasks differ from each other and from song singing and may vary by age. As is suggested in the conclusion to this review, a combination of these tasks may be necessary to test the skill of singing accuracy.

Item difficulty.

This section will review those studies that make an explicit attempt to define the difficulty of items within types of tasks in assessments, including discrimination power where available. Many of the applicable studies had interval or pattern comparisons as their main objective so that findings could be applied to the sequencing of instruction in classrooms. However, the data are also useful for researchers who wish to apply findings on the actual difficulty of specific pitch relationships (like intervals) to the assessment of student ability. First, I will relate the findings for intervals followed by patterns.

For specific intervals, there is a perception that thirds, especially descending thirds are easiest to sing. However, in one study, four-note patterns and half steps were shown to be no more difficult than whole steps or thirds (Sinor, 1984), although descending third intervals were found in many of patterns indexed as "easy." Young (1971) presented intervals in minor and major designations to kindergarten and first-graders. His study reports a range from easy to difficult for intervals in this order: descending minor 3rd, ascending and descending 5th, ascending major 3rd, and descending minor 6th. Jones (1971) listed the intervals and also the patterns selected for his test in order of difficulty when used in a small pilot study using eight second and third graders: descending minor 3rd, ascending Perfect 4th, mi-re-do, ascending 2nd, and do-re-mi.

Van Zee (1984) studied students in the first ten grades and found the ascending fourth "sol-do" to be the easiest interval sung by participants, which corroborates data from Petzold (1963) and Boardman (1964). Most importantly, and related to task

difficulty, Van Zee questioned how much pitch matching tasks correlate with song singing tasks. She found little-to-no correlation and like the kindergarteners studied by Demorest and Nichols (2012), suggested these tasks may not measure the same construct (or at least not exactly the same combination of component skills). Perhaps song singing is a more complex task, which is why song singing was not shown to improve over six months' time in the kindergarten sample (and also Apfelstadt, 1984), but pitch matching did improve. Perhaps the isolated, even sterilized, task of pitch matching is a skill that can be easily taught, facilitated, and reinforced. The task may be a minimal challenge for short-term memory, and since intervals and patterns are brief, they may be easier and less intimidating to sing than whole songs.

The Singing Voice Development Measure was written with a minor melody (SVDM, Rutkowski, 1990), though tonality has been compared in song singing tasks, and children in grades 1-6 performed better on the major key song *Row, Row, Row Your Boat* than they performed on the minor key song *In the Sea* (Levinowitz et al., 1998). The authors of that study, however, suggested the two songs were not well-matched, so song singing in a major key cannot be said to easier. Indeed, singing a minor song has twice been demonstrated to be easier than singing a major song (DeYarman, 1972, and Dittemore, 1969, as cited in Levinowitz et al., 1998). Therefore, there are conflicting results as to whether children sing minor or major songs more accurately.

In a study exploring the effect of tonality on singing accuracy, Wolf (2005) administered the Tonal Performance Pattern Test (TPPT) to K-2nd graders in order to determine the difficulty of 40 two- and three-note patterns split between parallel forms of

a major key and a minor key section of the test. The scores were normally distributed and scores were categorized by difficulty in this way:

1. Difficult: lower than one standard deviation below the mean;
2. Moderate: within one standard deviation above or below the mean;
3. Easy: higher than one standard deviation above the mean.

Seven patterns were easy; seven were difficult; and the remaining 26 patterns were deemed moderate. Three factors did NOT seem to affect difficulty:

1. Modality (major or minor);
2. Contour (ascending or descending);
3. Harmonic function (the patterns were classified as Tonic or Dominant in function).

Three factors DID seem to determine difficulty in performance:

1. Interval (all easy patterns were thirds);
2. Length (all easy patterns were two notes);
3. Range (all easy patterns were low in range).

The second finding reinforces that as tasks increase in number of pitches, they increase in complexity, and also difficulty.

A sample of 480 6-9 year-olds from four countries was tested on 16 intervals (Moore et al., 1995). The intervals consisted of unison, 2nd, 3rd, 4th, 5th and octave, but not 6th or 7th in order to limit the test to ten minutes. Overall, 60% of intervals were sung correctly by the participants. Results indicated that the unison, descending minor 3rd, and perfect 4th were easier than the minor 2nd and octave. Specifically, the order of difficulty (pass rate) was: unison (80%), descending minor 3rd (73%), descending

perfect 4th (68%), ascending major 2nd (66%), descending major 3rd (65%), descending major 2nd (62%), ascending minor 3rd (59%), ascending major 3rd (58%), descending and ascending perfect 5th (both 54%), ascending perfect 4th (51%), descending half-step (46%), ascending octave (45%), ascending half step (38%), and descending octave (36%). There was no significant difference between ascending and descending intervals.

It is possible that task difficulty varies by student. Whereas studies usually describe data in terms of group performance and mean scores, Bergee (2012) found that the examination of individual student performance in a measure of rhythm reading ability was useful for understanding sight-reading performance. In a profile analysis of individual students, he described the ability of specific students in relation to others. This kind of analysis can be helpful for understanding how specific items function in test construction and make clear when items perform differently among students.

Task discrimination.

Roberts and Davies (1975) tested three pitch matching tasks plus song singing in a sample of 90 students identified by the classroom teacher as “monotone singers” to study the effects of remediation instruction. They asked students aged 6-8 years old to do a variety of tasks including single pitch, interval, melody, and song singing. The sample was randomly assigned to a control group, a traditional group, or a remedial group, with thirty participants each. A additional group of 30 “normal” singers was added for comparison. For testing on the song task, children were allowed to select any song or nursery rhyme to sing from memory. After twice weekly singing improvement sessions over eight weeks, both the normal singer group and the monotone singer

groups improved on all measures of singing production. After eight weeks the remedial group indicated better improvement of single pitch and interval production and their vocal range improved overall, but song singing did not. More recent samples of typical kindergarteners confirmed these findings (Apfelstadt, 1984; Demorest et al., 2012; Welch et al., 1997).

Each of the four vocal production tasks used by Roberts and Davies are presented in Table 2.1, which were shown to have significant discrimination power and were said to be interpreted reliably (since there was only one item for free song, split-half reliability could not be calculated). The authors used the *Cj* statistic, which is a expression of the ratio between the difference of means and the combined standard deviation for an item (the larger the *Cj* number, the better the discrimination power). The authors reported that there was statistically significant discrimination between the test scores of the normal singers and the monotones ($p < .001$). The comparative performance of all students, rather than "monotone" versus "normal" students in a

Table 2.1

Discrimination power of singing tasks (Roberts & Davies, 1975)

Test	Discrimination (Cj)	Split-half reliability
Single pitch	1.77*	0.88
Interval	1.62*	0.91
Melody	0.73*	0.76
Free song	1.01*	-

*$p < .001$

remedial situation, has not been studied in upper elementary students for solo or doubled singing.

Conclusion

The development of children's singing accuracy is a responsibility of music teachers, and it is important for teachers to have comprehensive measures of singing accuracy. When asked to report performance achievement to students, parents, or school leaders, teachers need to know which tasks provide the best "picture." The nature of specific singing tasks may reflect the ability of the student in a unique way, and students may respond differently to specific tasks. That is, one task may represent all students differently than another task, or students may respond differentially to certain tasks.

Generally, children sing accurately more often than not, especially if one diminishes the value of maintaining a pitch center (i.e., staying in the same key), which may prove to be a more advanced skill rather than one which occurs early in the developmental process. However, the literature indicates that as few as 13% of general music teachers formally assess singing in their students, usually using teacher-designed scales or rubrics (Salvador, 2010).

Singing research has employed pitch matching tasks such as single pitches, intervals, patterns, phrase singing, and song singing tasks. Roberts & Davies (1975) established that at least four of these—the single pitch, interval, pattern, and song singing tasks—are discriminators of accurate singing. Therefore, it is clear that these four tasks can be used as tasks in singing assessment. The relative difficulty of specific

pitches, intervals, and even patterns has been reported, but there is less research comparing these task types overall.

Demorest and Clements (2007) and Welch et al., (1997) have suggested that a) performance in singing accuracy is task-based, and b) future research is warranted because there are important classroom and research implications for this possibility. Singing accuracy assessment depends on a thorough understanding of how performance differs based on the task given and under which conditions students respond most accurately. Once this is understood, remediation for poor singers can be refined by the specific transfer of skills from easy tasks to more difficult tasks, especially since these varying tasks may be inter-correlated.

Performance on pitch matching and song singing has been found to be moderately correlated in a sample of kindergarteners (Demorest & Nichols, 2012) and highly correlated in a sample of adults (Pfordresher et al., 2010). Pfordresher and Brown (2007) suggested that single pitch, interval, and pattern tasks vary in terms of complexity because of the number of unique pitches they contain and varied in memory demand because of the total number of notes that they contain. For their study, they presented the three tasks using four notes for every task to equalize memory demands. The authors presented the single pitch task in four (identical) notes, the interval pitches in four notes (two for the first pitch and two for the second pitch), and the pattern in four unique pitches. Performance on the three pitch matching tasks varied for participants in this adult sample. For participants deemed “good” singers, performance improved across these three tasks. For “poor” singers, however, performance decreased across these tasks, suggesting that good singers may benefit from the added contextual

information, whereas poor singers become more and more overwhelmed with additional complexity. Contrastingly, Wise and Sloboda (2008) found performance of "tone deaf" and "not tone deaf" groups to differ significantly from each other, but the performance of both groups across the three tasks (single pitches, intervals, patterns) was similiar in that all adults' performance decreased across tasks of additional complexity. Their pitch matching items varied by number of total notes (they did not control for memory effect by presenting the same number of notes in each pitch matching task type).

Comparing multiple singing tasks in one sample involves lengthier testing sessions than is usually feasible in school settings or with young children. Indeed, a test that assesses every possible presentation or response mode would take much longer than is feasible in younger age groups. Additionally, singing accuracy has been explored in certain contextual conditions, doubling conditions, and with varied stimuli characteristics. Of these, however, the effect of doubling has not been clearly established.

Purpose

As previously stated, many measures of singing accuracy have been used; however, it is difficult to compare results from each of these because they vary in the nature of the tasks they include, the sample, and also in item difficulty within those tasks. In an effort towards feasibility, or practicality, researchers have not often compared a range of task conditions—including pitch matching and song singing—at the same time. Likewise, most music education researchers have not controlled for the effect of memory in tests using various task conditions. Further, many studies have used only one type of task to represent the complex skill of singing or have used two

tasks to create a composite singing score, but without examining the role of each of these tasks. The purpose of this study was to explore task-based variability in children's singing accuracy performance. The research questions were:

1. Does children's singing accuracy vary based on the nature of the singing task employed?
2. Is there a hierarchy of difficulty and discrimination ability among singing assessment tasks?
3. What is the interrelationship among different measures and how few tasks might be employed in a comprehensive measure of the skill of accurate singing?

Chapter 3: Method

The review of literature indicated that the use of various singing tasks may affect indicators of student performance, but that few studies have compared the results of multiple singing tasks in the same sample of children (e.g., Demorest & Nichols, 2012; Roberts & Davies, 1975) or in adults (e.g., Pfordresher & Brown, 2007; Wise & Sloboda, 2008). An understanding of how singing tasks affect assessment results is important for developing singing tests for music classrooms. The purpose of this study was to examine task-based variability in children's singing accuracy performance, and the following variables were included as singing task features: response mode (echo and doubling) and task type (single, interval, pattern pitch matching, and song singing). These tasks were combined in a single assessment to measure singing accuracy.

Measures

Participants' ability to match pitches or sing songs is dependent on their ability to discriminate pitches. However, no measure of pitch discrimination was included on this test in order to limit the duration to ten minutes. While Geringer (1983) found no difference in discrimination ability based on pitch matching ability in 4th graders (and kindergarteners), other studies have reported varying results. For example, in a sample of junior high boys, individuals identified as *certain* or *uncertain* singers did perform differently on pitch discrimination (Demorest, 2001). Still, the purpose of the present study was to compare task-based performance in both pitch matching *and* song singing, and the inclusion of a sufficient number of items was prioritized over the assessment of pitch discrimination. Only vocal pitch accuracy was tested in this assessment of singing and the participants were not pre-screened for any condition or ability.

Regarding contextual presentations, Demorest and Clements (2007) found a group of inconsistent singers in their sample to perform more accurately when pitches were presented in a context compared to no context. Fourth graders may or may not demonstrate stratified ability similar to the junior high students in the previous study, and contextually-presented stimuli would have been useful for comparison in the current study. However, context presentation adds time to the test duration and would have a varying effect on single pitch, interval, and pattern tasks. A three-pitch context presentation preceding a single pitch item would have resulted in participant exposure to up to four unique pitches and seven total notes. A three-note context preceding a pattern item, however, would have resulted in up to seven unique pitches and seven total notes. Since the design of the current study necessitated an equal number of pitch presentations, contextually-presented stimuli could not be used.

Elementary-aged students have been shown to perform better in groups than alone (Green, 1990). However, no difference between students singing alone or along with a researcher has been demonstrated (Cooper, 1995). Thus, previous research with children has been inconclusive regarding the impact of echo versus doubled singing. Singing along with peers may present different challenges than singing with a familiar or unfamiliar adult, perhaps due to social effects. The choice to use pre-recorded stimuli in the current study was based on a desire to standardize the experience for all participants, intending thus to increase internal reliability and to ensure that each participant heard identical pitches. Children were asked to sing with an unfamiliar voice, which differs from what would occur in a typical classroom setting. Additionally, students' confidence could be enhanced or diminished by the presence of

any peers. Despite these potential minor threats to external validity (it will be more difficult to generalize these findings to singing along with a teacher, a peer, or a group of peers), these decisions were made to increase internal reliability.

All four singing tasks were presented to all participants in two response conditions which reflect singing as it typically occurs in classrooms: echo and doubled singing. The first of these response modes represents the scenario in which a teacher presents a pitch or pitch sequence and evaluates student performance (whether for accuracy or other characteristics like tone) when the child sings it back alone. The song singing component is used as a solo condition in the current design, rather than being echoed after it was heard. In this case, children were asked to sing *Jingle Bells* from memory. The second condition represents occasions in which students are asked to respond at the same time as other students, or to sing along with the teacher. Referred to as doubled singing, this response condition also represents a common classroom activity.

Because some assessments incorporate pitch matching or song singing but not both, a song singing task was included in this instrument for a comparison to the pitch matching tasks. Each task was presented in a solo condition and presented again in a doubled condition. The song *Jingle Bells* was chosen for its familiarity to children and its use in and outside the music classroom as a secular song; moreover, it has been previously used in singing accuracy research with children (Apfelstadt, 1984) and adults (Berkowska & Dalla Bella, 2009; Pfordresher et al., 2010).

Sample

Participants ($N = 120$) were 4th grade students from public ($n = 2$) and private ($n = 4$) elementary schools in a metropolitan area in the northwestern United States. Fourth graders were sampled because they represent the later part of elementary voice development, and were expected to be able to perform well for a duration of ten minutes (Petzold, 1963). Fifth grade was not chosen because some students begin to experience voice change at this time (Green, 1990). These students came from different classrooms (i.e., they had different homeroom teachers) but shared the same music teacher within each school. The schools were chosen because the music teachers' curricula included singing and playing instruments and were otherwise representative of the typical elementary music setting, which included the criteria of once-weekly music classes.

Several documents were drafted for this study. An information sheet, designed for parents to keep at home, gave a brief description of the study and contact information (Appendix B). This sheet was distributed at the same time as the consent form when classes were introduced to the researcher. Additionally, a consent form (Appendix C) described:

1. The purpose of the study;
2. The procedures (described below);
3. The risks, stress, or discomfort potentially experienced by participants, which were explained to be similar to what students would normally experience in music class;
4. The benefits of the study;

5. An assurance of confidentiality;
6. The right to withdraw from participation at any time.

An assent form was used to remind students of the procedures and ability to withdraw at the time of data collection (Appendix D). These procedures were endorsed by the university's human subjects division and subsequently approved by the school districts where participants were sampled.

Background questions.

At the end of the consent form three background questions were provided. These three questions were prioritized based on how they were theorized to potentially affect singing instruction, and they were included on the consent form rather than a separate document to encourage completion. First, a question asked parents to record the child's age at the time of consent. Next, parents were asked to report whether "Your child has ever been diagnosed with any hearing difficulties." Hearing disorders, disabilities or other impairments could potentially affect a child's ability to match pitches or sing songs; therefore, it is important to consider the role of such variables.

Last, the form asked whether students had participated in any private music instruction and if so, for how long. Previous research supports a link between musical experience and instrumental tuning accuracy. Yarbrough, Morrison, and Karrick (1997) found that the only factor that significantly affected tuning performance in that sample of high school students was participation in private instruction. Possibly, private lessons have a more salient effect than other in- or out-of-school music participation. Or, participation in private instruction could be a moderating variable for other, unknown physical or psychological characteristics in children.

Stimuli

The design of this study involved the development of assessment items and corresponding pitch stimuli that allowed for a comparison of various singing tasks. The pitch stimuli were recorded by an adult female vocal model instructed to sing on an [u] vowel with minimal vibrato at prescribed pitches and tempi because these factors were suggested to maximize performance (Yarbrough et al., 1992). The model was a vocal music education student at a major university.

The acoustic profile of the first pitch from Item 1 of the vocal stimuli is reported here for two reasons:

1. To demonstrate that the /u/ vowel was used and at what mean frequency;
2. So that future research can replicate any parts of this assessment.

The /u/ vowel can be demonstrated by the mean frequency of the first and second formants (Figure 3.1). These descriptives refer to the approximate middle 50% of the sung sample. The term *approximate* is used because the exact onset and coda are difficult to determine and are chosen by the researcher. The mean of the first formant was 447.02 Hz and the mean of the second formant was 885.61 Hz. These means are consistent with the frequency range of the /u/ vowel, and the vocal model approximated this vowel for every pitch in each item. The pitch given to the model for recording was 440 Hz from an electronic pitch pipe and the mean of the stimulus pitch was calculated to be 438.89 Hz (*SD* = 1.74). The frequency range of the middle 50% of the sample was 435.31 to 441.87 Hz. The deviation of the stimulus from the given pitch was 4.37 cents, imperceptible to most individuals (Micheyl, Delhommeau, & Perrot, 2006). The model recorded the fifteen items, which were then used for both the solo and doubled

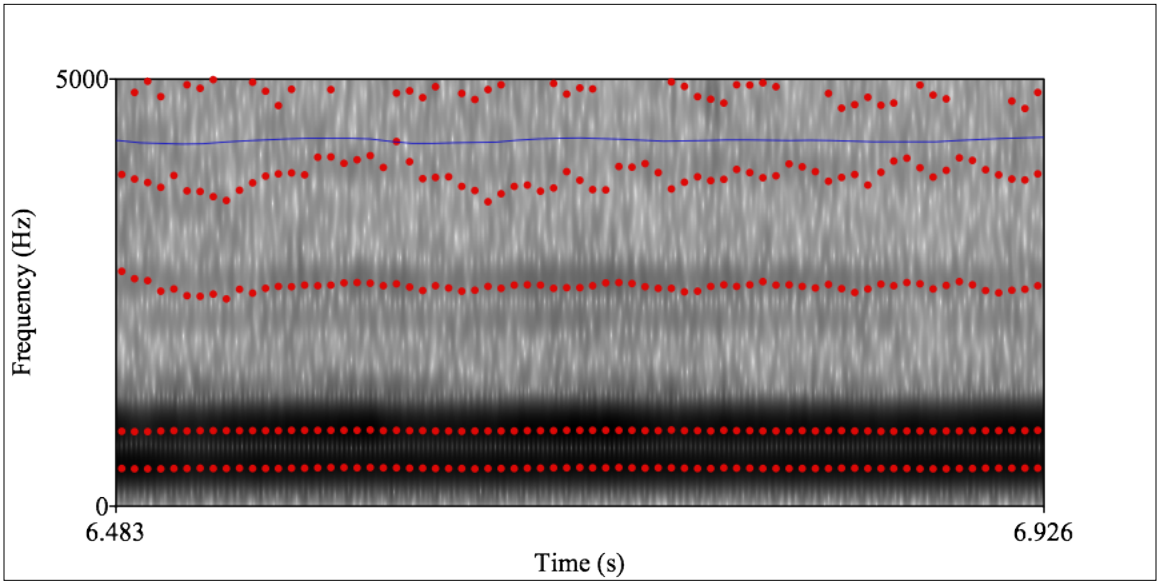


Figure 3.1. Spectrogram approximating A440. This figure demonstrates formants and pitch plotted using Autocorrelation Analysis Method (438.89 Hz).

conditions of the assessment. In the doubled condition, the same modeling was used for the participant to sing along with after hearing the stimuli.

The items were presented in the key of D (Goetze, 1985; Wolf, 2005) in this order: single pitches, intervals, and patterns (Table 3.1). All items were based on the same key center since this lessened the possibility that performance on one item would affect performance on the subsequent task. For example, if the assessment included varying key centers, some keys would be more and less related to the key center in the first item. Performance on subsequent items based on more distant key centers might provide more challenging transitions than to nearer key centers. It would be difficult to interpret difficulty and discrimination results with this additional factor.

Table 3.1

Pitch Matching Tasks (in the order they were presented for both the solo and doubled conditions)

	Single Pitch	**Interval**	**Pattern**
Single and Doubled Conditions	1. AAAA 2. DDDD 3. GGGG 4. F#F#F#F# 5. AAAA	6. AAF#F# 7. DDEE 8. F#F#DD 9. DDGG 10. AAF#F#	11. AF#AD 12. DEF#G, 13. F#ADE, 14. AGED 15. AF#AD

The pitches ranged from D4 to A4 and were presented at approximately 60 beats per minute using the syllable “doo” with no separation between notes. There were five 4-note items in each task condition, each task preceded by one 4-note practice item. The song task in the key of D was presented at the conclusion of the pre-recorded pitch matching stimuli; those participants who began the assessment with the echo pitch matching condition performed the solo song singing before advancing to the doubled condition, and vice versa. Students were randomly assigned to begin the assessment in the echo or doubled condition and to end the assessment with the other condition.

The suggestion to equalize the presentations in number of notes to create greater uniformity in test items comes from the adult singing accuracy literature (e.g., Pfordresher & Brown, 2007), which compared four-note presentations of each of the single (e.g., *do-do-do-do*), interval (e.g., *do-do-re-re*) and pattern tasks (e.g., *do-re-mi-do*) in order to mediate memory and presentation effects. This strategy was used successfully with a sample of kindergarteners in demonstrating task performance variability in early elementary-aged students (Demorest & Nichols, 2012).

Each of the pitch matching tasks (single pitch, interval, pattern) were preceded by one practice example to familiarize participants with the next task and to provide an opportunity for participants to ask any procedural questions. The practice examples were intended to familiarize students with the task type prior to their performance on the first item in each task. Specific items were modeled after pitch sequences used in a previous singing accuracy study (Demorest & Nichols, 2012) and represent varying pitches and ascending and descending intervals within the range of a fifth (Table 3.1). Additionally, the fifth item in each task was designed to replicate the first (it contained an identical pitch set). This allowed for a comparison of these two identical items for a measure of internal reliability (i.e., how stable items in each task may be). This was important since the internal reliability of items within task conditions was unknown. Last, two of the pattern items were designed to replicate portions of the melody in the song-singing task. Item 13 was the pattern F#-A-D-E, which corresponds to the words in the song, *Jin-gle all the (way).* Item 14 was the pattern A-G-E-D, which corresponds to the words in the song, *(one) horse o-pen sleigh.* When presented as a pitch matching task, these pitches are removed from the song singing context; it was worthwhile to introduce this aspect because comparing pitch matching and song singing performance was an important part of the need for this study and because of the possibility of varying levels of association among tasks in the solo and doubled conditions.

The only order effect the design controlled for was the presentation of the solo and doubled conditions. A primary purpose of the study was to compare the solo and doubled response conditions and for this reason those order effects were prioritized. It

is possible that characteristics of a pitch level or specific intervals could affect performance on a subsequent item, but so too could any aspect of a student's performance on an item affect performance on the next item. These assessment stimuli were presented to children in such a way that every child heard the exact same stimuli in the same conditions, presented in one of two orders. Children were randomly assigned to a form so that a near-equal number of students were assigned to each order.

Song singing in the solo condition represented a song task similar to echoed pitch matching. The song was initiated by presenting the starting pitch, F-sharp, on a pre-recorded electronic pitch pipe and the recorded model saying, "Ready, set, sing." This song task in the solo condition is not echoed in the way that pitch matching exercises were since it was sung from memory. After performing the doubled pitch matching items, song singing in the doubled condition began similarly: the starting pitch was given and the recorded model spoke, "Ready, set, sing" before beginning to sing. Each pitch matching and song singing item was presented with adequate time for student response. The duration of this assessment was approximately ten minutes.

Procedure

Prior to data collection, the researcher was introduced to each class as a graduate student at the local university. The project and the stimuli were described and consent forms and an information sheet were distributed to each student. This usually occurred in the home classroom to enlist the support of the classroom teachers, who agreed to collect the forms. Returned consent forms were assigned a student

identification number to be used for coding data so that no identifying information was kept with the results of the assessment.

Participants were assessed individually during the school day and were asked to miss ten minutes of classroom activities. Some schools allowed testing throughout the school day, whereas others restricted participation to the time period for music class only. This restriction and classroom availability determined scheduling at each school. Empty classrooms, workrooms, and storage spaces were all used during data collection.

Assent procedures were followed for each participant, who had an opportunity to ask questions before beginning. The researcher tested all participants and made an effort to make them as comfortable as possible. The researcher explained that an identification number would be spoken at the beginning, and participants were shown that the recorder was in audio-only mode. During testing, stimuli were played on a Sony CD player in stereo placed three feet away and the responses were recorded as .WAV files with a Zoom stereo recorder placed between the participant and the CD player. Participants were recorded individually and an effort was made to minimize disruptions and distractions; participants were seated facing away from windows and doors.

Analysis

Previous studies have used of both computer scoring and human scoring; the latter was determined to be appropriate for this study since classroom teachers regularly must make decisions about student performance (ecological validity). Thus, human scoring was considered best for these tasks, which are similar to classroom tasks, and reflect how singing is judged in the classroom. In this study, the researcher

and an external evaluator, a graduate music education student with school teaching experience, scored 20% of the participants (n = 24). The use of the second evaluator addressed the problem of the researcher's potential familiarity with the students, which occurred during data collection and could have affected scoring. The second judge participated in a one-hour training session to practice the dichotomous scoring of pitch matching samples, which were selected to represent a range of ability. The second judge also practiced applying the eight-point scale to the song singing tasks. The researcher scored half the remaining participants (n = 48) and the second evaluator each scored the remaining participants (n = 48).

A student identification number was spoken at the beginning of each track to facilitate scoring and assure anonymity and ensure that filenames were accurately labeled. After recordings were made for each participant, the audio files were prepared for scoring by placing the echo condition first for those participants who were assigned Form B (doubled singing followed by solo singing) to facilitate use of the scoring form. All judges scored the solo condition followed by the doubled condition, regardless of presentation order. This was done to avoid scorer error in the event judges might accidentally assign doubled scores in the place of echo scores on the scoring form. All participants had their echo responses scored followed by their doubled responses. Last, the verbal instructions and practice examples were removed from each track to decrease scoring time. This shortened the length of each file by approximately three minutes.

Scoring.

Pitch matching was considered correct if the pitch was sung closer to the given pitch than a chromatically adjacent pitch, and each of the four pitches in each test item (15 items for the solo condition and the same 15 items for the doubled condition) were scored independently of the surrounding pitches. For example, Item One was A4-A4-A4-A4 (approximating 440 Hz) and each of the four pitches was scored independently, even for the single-pitch task mode. Every student sang 30 items (15 for the echo condition plus the same 15 items for the doubled condition) consisting of four pitches each for a total of 120 pitches for scoring. An "accurate" pitch was coded "1" and an inaccurate pitch was coded "0". Therefore, a singer who sang all pitches in Item One was coded 1-1-1-1, or a singer who sang only the first of the four pitches correctly was coded 1-0-0-0. Scoring continued in this way for each of the fifteen items in the solo condition and each of the fifteen items in the doubled condition.

Some students scoop when they sing, whether regularly or in some parts of their range. To mediate the effect of scooping on participants' pitch matching scores, participants were given credit if they sang the correct pitch during any part of their response. For example, if a student sang half the pitch too low and half the pitch accurately, a score of "1" was given. However, proportion did not affect the score. For example, if a student sang one-fourth of the pitch accurately then lost the pitch for the remaining three-fourths of the response, a score of "1" was given (reversed scooping). For a glide response, where the pitch began lower than the given pitch and ended higher than the given pitch (or vice versa), no credit was given.

Song singing on the song *Jingle Bells* was scored using the Singing Accuracy Scale (Wise & Sloboda, 2008). The scale contained eight levels, ranging from Level 1, little variation in pitch and singer may chant, to Level 8, all melody is in tune and singer maintains key. The song singing item was included in the assessment following the pitch matching items in each condition: solo pitch-matching was followed by solo song, and doubled pitch-matching was followed by doubled song.

Both song singing response modes, solo and doubled, were also given a second measure of tunefulness: whether the participant sang in the given key. This dichotomous measure was employed because it gave additional information about participants' in-tune or out-of-tune singing. It was useful for evaluating the varying ability of participants to sing in the given key.

Missing data.

Missing data occurred for two reasons: 1) Microphone interference from another electronic source made one item impossible to score; 2) Eleven students failed to sing one or more of the pitches presented. In the second case, participants occasionally would not make the onset of the first pitch coincide with the doubled stimulus and instead would wait for the second pitch to begin. These missing scores presented a unique challenge for interpretation. Several solutions exist and are discussed here. First, the participant could be left out of analysis entirely; however, this results in a smaller sample since a ten-minute test consisting of 30 items plus song singing presents opportunities for fatigue or distraction. Second, a zero could be substituted for the missing score, using the logic that no credit should be given for a pitch which is not accurately sung. However, for the student who sings the other three pitches in an item

correctly, (1-1-1-X), a false zero score where the student might rather have sung it correctly if it were sung by the student misrepresents the student's ability. A third solution is to replace the missing score with an accurate score, "1", where a student sings the other pitches accurately. This solution is especially logical when the student's overall score is near-perfect. For the student who sings mostly inaccurately, a score of 0-0-0-X could be changed to 0-0-0-0. More problematic are the cases where a missing score occurs in the pattern task: 1-1-X-1. How can we guess whether the student would have sung the third pitch accurately? Lastly, a fourth solution is to substitute the mean score, substituting the mean of all participants' scores on a specific pitch. This last method allows for the inclusion of every item for every participant, and does not raise or lower the means at the item level. It was decided to employ the fourth strategy, where item-level means were substituted for missing scores.

There were two exceptions for the handling of data. One participant had missing scores for Item 9 in the solo interval condition due to audio interference. This participant performed accurately on every other pitch except the fourth pitch of Item 13 in the solo condition and the fourth pitch of Item 13 in the doubled condition. Therefore the decision was made to substitute accurate scores for the missing data in Item 9. A second participant provided a song-singing sample that proved difficult to score. This participant sang the first phrase of the song in a way that would have been given a score in the middle of the eight-point scale. However, the student was unable to finish the first portion of the song and instead began again and sang in a more chant-like way. Since there was no plan for scoring such an inconsistent response, the item average was substituted for this participant's solo song-singing score.

Summary

The purpose of the study was to examine task variability in 4th graders' singing accuracy when singing alone and when singing with another voice, represented here by the reproduction of the stimuli presented a second time to the participant (the participant heard the stimulus, then it was played again for the child to sing along). The study employed a 2 X 4 factorial design in which an echo and doubled response condition were used for testing single pitch, interval, pattern, and song singing tasks. The familiar song *Jingle Bells* was chosen for song singing based on its use in a previous study and was presented to participants following the pitch matching tasks in each of the echo and doubled conditions. Pitch matching items were selected to represent pitches and intervals within the range of a fifth in the key of D and were preceded by a practice item in each task.

Participants were 120 4th grade students in elementary schools who completed the ten-minute assessment one-on-one with the researcher after voluntarily completing consent and assent procedures. School sites were chosen based on the presence of regular, weekly music instruction for students in the 4th grade. Once permission was received from the Human Subjects Division and the District and Building administrators, the researcher was introduced to each class in their home classroom. Data collection occurred between February and April 2013 using a Stereo CD player and a portable stereo recorder. Data were recorded using a student ID rather than students' names.

The data were scored by two judges because human scoring represents how scoring is usually done in classrooms. The pitch-matching items were scored dichotomously and the song singing was scored using a previously-designed scale

appropriate for the construct of singing accuracy. Additionally, a dichotomous measure for whether the song singing was performed in the given key was included. For pitch matching, student responses were considered accurate if they were sung closer to the given pitch than to an adjacent pitch. “Scooped” pitches were considered accurately sung if they led to the correct pitch, and a rationale was presented for substituting mean scores for missing data.

If researchers can evaluate task-based variability, they can provide better tools to teachers who test and remediate singing skills in their students. This test was set up to assess singing in two ways it occurs in classrooms, alone and doubled. The doubled condition was standardized for all students because a pre-recorded stimuli/doubling track was used, and future results may vary from those in the following chapter based on model characteristics, specific pitches used as stimuli, and stimuli presentation (live versus recorded modeling). The varying tasks were chosen based on their prevalent usage in previous studies as the most deserving of comparison. Further, they reflect actual singing activities in classrooms and have been shown to be good discriminators of singing accuracy.

Chapter 4: Results

The purpose of this study was to investigate the task-based variability in singing accuracy in a sample of 4th grade students. The variables included four types of singing tasks. Three were created by varying the number of unique pitches in four-note presentations: single pitch, interval, and pattern, in addition to singing *Jingle Bells* from memory. There was another variable in this design, the response mode, where students sang every task alone (solo) and along with a vocal model (doubled), who provided the stimuli for the assessment.

The research questions were:

1. Does children's singing accuracy performance vary based on the nature of the singing assessment employed?
2. What is the relative difficulty and discrimination of various singing assessment tasks?
3. What is the interrelationship among different measures and how few tasks might be employed in a comprehensive measure of singing accuracy?

Demographics

Data were collected at both public (n = 2) and private (n = 4) elementary schools in a large city in the northwestern United States between February and April of 2013. Three of the private schools were Catholic parish schools. Once-weekly music instruction was offered at all schools, and the entire 4th grade population at each school was recruited for the study by the music teacher and supported by classroom teachers (n = 15), who were willing to help collect the appropriate forms.

Consent forms including four background questions were returned by volunteer participants, who each completed the entire assessment (N = 120), except for one

student whose solo song singing score was removed based on an inconsistent response. The first background question asked the parent/guardian to indicate if "your child has ever been diagnosed with any hearing difficulties." It could be theorized that a history of hearing disorder would affect performance on an assessment that requires participants to sing newly-presented items or a memorized song. No participants indicated "yes." The second question asked if the child had ever participated in private instrument/voice lessons. Forty-five percent of participants indicated "no" (n = 54) and 55.0% indicated "yes" (n = 66). For those with a history of private lessons, the mean duration was 20.75 months (SD = 14.66) and ranged from .5 month to 54 months, with a skewness of .526 and kurtosis of -.782. The median reported lesson duration was 18 months and the mode was 12. The last question asked families to indicate the child's age at the time of consent. These 4th grade students ranged in age from 9 years, 6 months to 10 years, 10 months, with a skewness of .109 and kurtosis of -.874. The mean age was 10 years, 1.3 months (SD = 3.73 months). The median age was 10 years, and the mode was 10 years, 5 months.

Inter-rater reliability.

All data were scored either by the author or a second evaluator. To calculate inter-rater reliability, both raters independently scored 20% of the data chosen at random. Since the pitch matching was scored dichotomously and song singing component was scored using an eight-point scale, the reliability was calculated separately. Since the scores represented continuous data, Pearson's r was chosen for calculating reliability and the mean correlation across all pitch matching items was .88. Since the song singing data also represented continuous data, Pearson's r was chosen

and the the correlation of song singing scores was .90. Where raters' scores differed, they were averaged. These results were deemed acceptable to continue with scoring, and the remaining 96 subjects were randomly assigned to the researcher (n = 48) or the second judge (n = 48).

Transforming data.

Data were transformed in order to compare the pitch matching and song singing tasks in the same analysis. Pitch matching scores and song singing scores were transformed to a scale of 0-1 to express easily the difficulty and discrimination indices and to make scores from different measures comparable for conducting all other analyses. Mean pitch matching scores at the item level (x) were divided by four for this transformation (x') for a resulting scale in which the lowest possible score is "0" and the highest possible score is "1". This resulted in five possible item-level scores: 0 if none

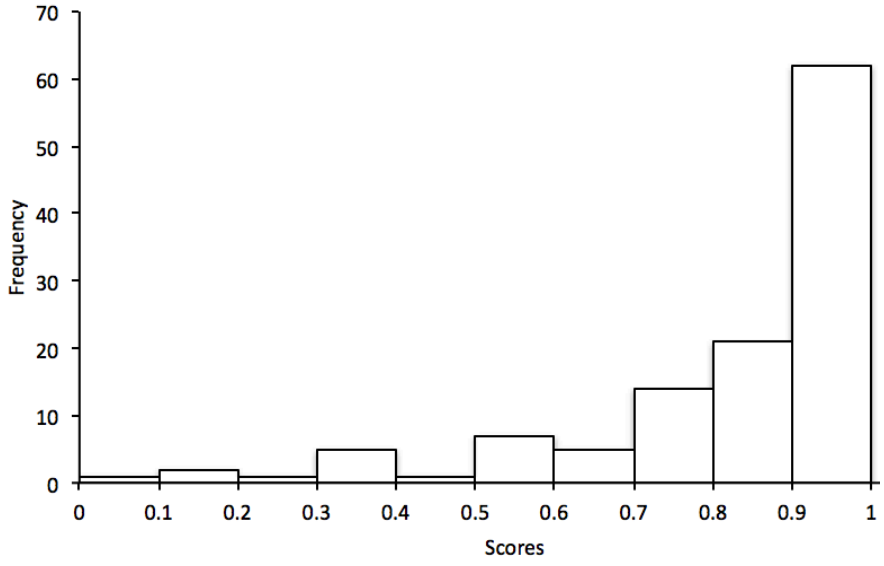


Figure 4.1. Histogram of overall performance. This chart shows performance on a total score averaging pitch matching and song singing items.

of the four pitches were sung accurately, .25 if only one of the four pitches were sung accurately, and so on. The song singing scale ranged from one to eight (no possibility of a "0" score), so scores (w) were transformed with the expression, w' = (w - 1)/7, for a resulting scale in which the lowest possible score is "0" and the highest possible score is "1". The distribution of overall scores is shown in Figure 4.1.

Since the song singing scores were based on an eight-point scale developed for singing accuracy (Figure 4.2), the untransformed scores are presented in Figure 4.3 and Figure 4.4 so that they can be interpreted according to the scale. The distribution of solo song singing scores indicated a skewness = -.858 and kurtosis of -.303. The distribution of doubled song singing scores had a skewness of -1.515 and kurtosis of

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

Figure 4.2. Song singing scale. This scale was developed by Wise & Sloboda (2008) and was used for evaluating the songs tasks in the current study.

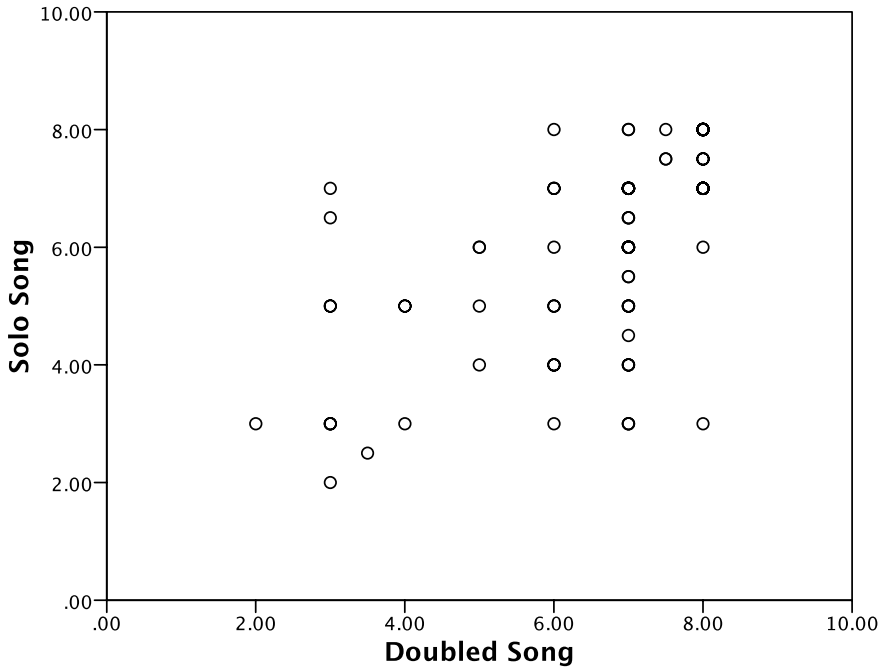


Figure 4.3. Plot of solo song singing and doubled song singing (n = 119). This figure plots doubled song singing against solo song singing using the 8-point scale (Wise & Sloboda, 2008).

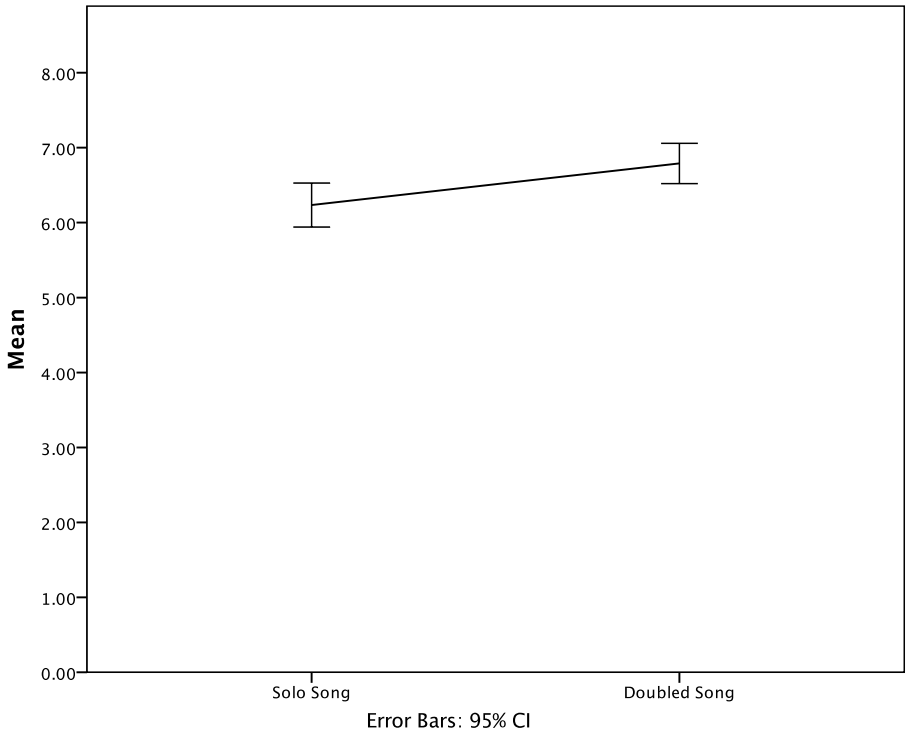


Figure 4.4. Line graph of song singing means. The songs were scored using an 8-point scale on the song Jingle Bells, *n* = 119).

1.525. A difference score for song singing in the two conditions was calculated by subtracting the solo condition score from the doubled condition score, and the distribution indicates that 50.4% evidenced higher doubled song singing (*n* = 60), 36.7% of participants had no increase or decrease in performance (*n* = 44), and 12.9% evidenced lower doubled song singing (*n* = 15). The distribution of these difference scores indicated a mean of .55 with a skewness of .208 and a kurtosis of 2.622. An additional measure of singing in the given key was used, and half the participants sang the solo song in the given key (*n* = 60) and half did not (*n* = 60). In the doubled condition, more students sang in the given key (*n* = 102) than did not (*n* = 18).

Order effects.

Half of the participants took the test beginning with the solo condition followed by the doubled condition (n = 60). The other half of participants took the test beginning with the doubled condition followed by the solo condition (n = 60). A one-way ANOVA to compare the four tasks (single pitch, interval, pattern, and song singing) was performed for the solo condition, indicating no main effect for presentation order (F(1,117) = 3.887,

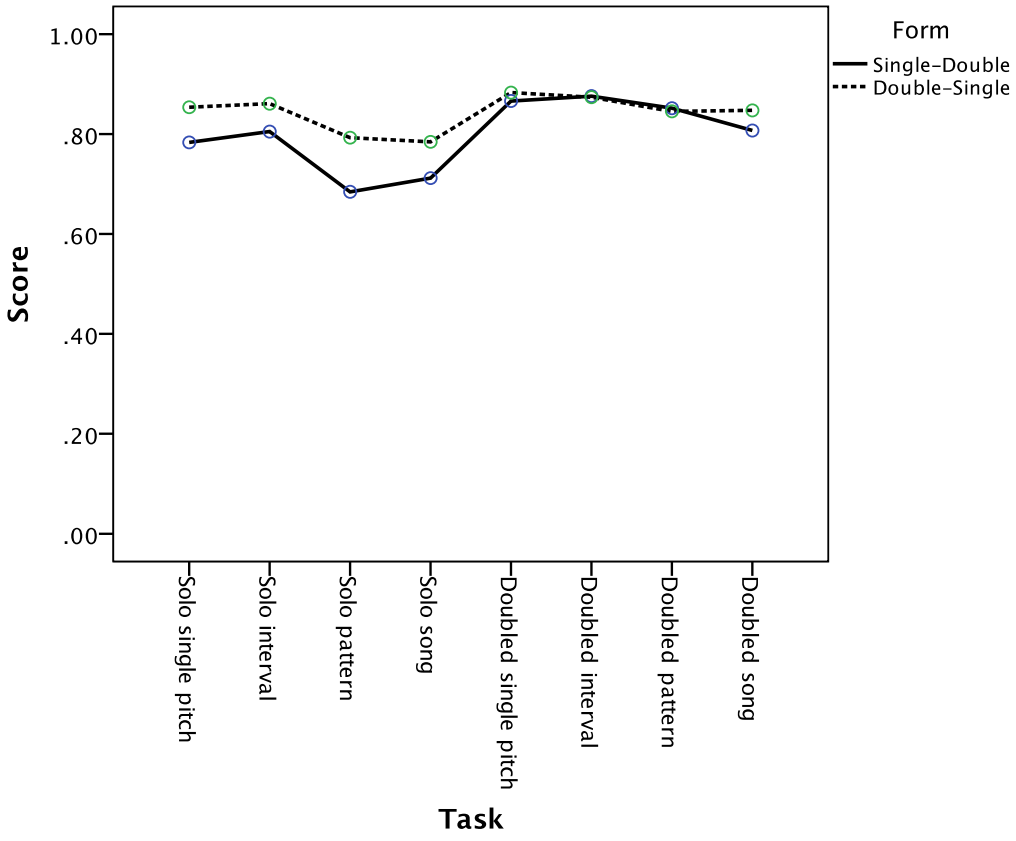


Figure 4.5. Task performance by presentation order. This figure illustrates that solo task performance is marginally higher when it was presented to subjects after the doubled condition.

p = .051). For the doubled condition, there was also no main effect for presentation order (F(1,118) = .017, p = .896). Thus, the following analyses were done with the groups combined. A plot of mean performance on each task by presentation order is given in Figure 4.5 because the result was nearly significant.

Question Number One: Does children's singing accuracy performance vary based on the nature of the singing assessment employed?

Previous results indicated that singing accuracy performance varied by the type of task chosen for assessment. In at least one adult sample, performance decreased across three types of pitch matching tasks (Wise & Sloboda, 2008). In another, performance increased across the three tasks for accurate singers, but decreased for poor singers (Pfordresher & Brown, 2007). A sample of kindergarten students performed better in the interval task rather than the single pitch or pattern task (Demorest & Nichols, 2012). For the current study, the question of task variability was answered primarily by an ANOVA with two within-subjects factors (response mode and task type) and one between-subjects factor (history of private music lessons). The transformed scores are displayed in Table 4.1. Performance across task types was

Table 4.1

Task Performance (SD)

	Single pitch	Interval	Pattern	Song
Solo condition	.82 (.29)	.83 (.23)	.74 (.24)	.75 (.23)
Doubled Condition	.87 (.25)	.88 (.21)	.85 (.24)	.82 (.22)

n = 120 for pitch matching tasks and n = 119 for song singing tasks

significantly different using a Greenhouse-Geisser correction ($F(2.421,280.822) = 3.984$, $p = .014$). Post hoc tests using the Bonferroni correction revealed that single pitch and interval task performance were significantly different from pattern and song task performance ($p < .0005$). Therefore we can conclude that performance on the first two tasks was significantly higher than the last two tasks (Figure 4.6).

There was a significant main effect for response mode, ($F(1,116) = 21.307$, $p < .0005$), with participants demonstrating significantly better performance in the doubled condition. There was also a significant task by response mode interaction, ($F(3,348) =$

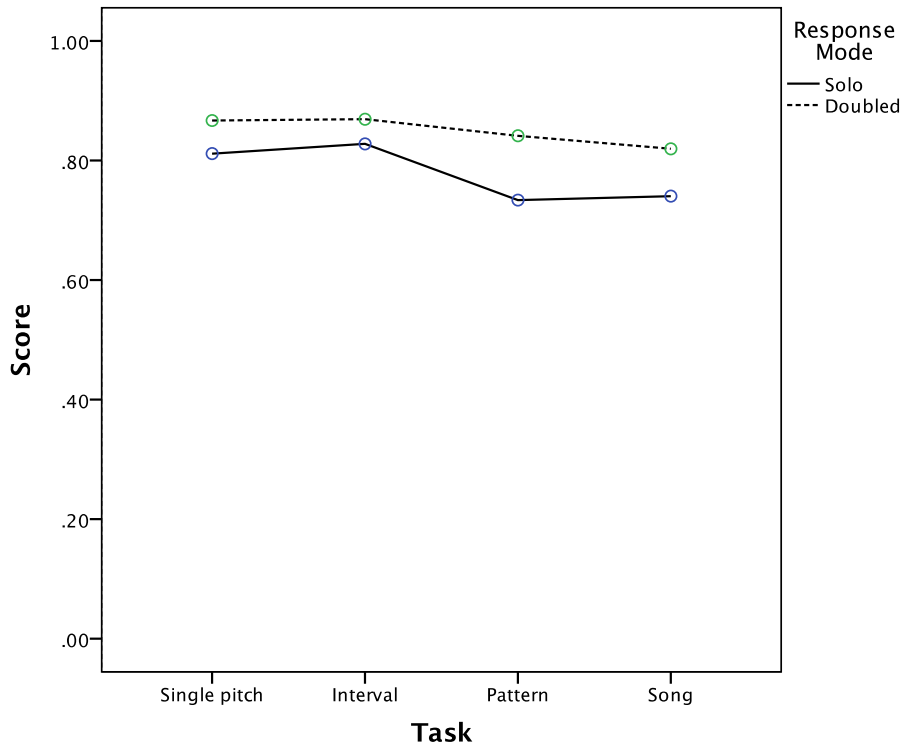


Figure 4.6. Task comparisons in solo and doubled conditions. This figure demonstrates the task by response mode interaction.

2.820, $p = .039$). Post hoc tests indicated a significant cubic relationship ($p = .009$). Therefore, we can conclude that task performance varies based on the response condition; specifically, solo interval performance was higher than single pitch; pattern performance decreased but song performance was slightly higher. Doubled singing was less accurate in the latter two tasks than the first two tasks.

There was a significant between-subjects effect for history of private music lessons ($F(1,116) = 10.718$, $p = .001$). Participants with a history of lessons performed

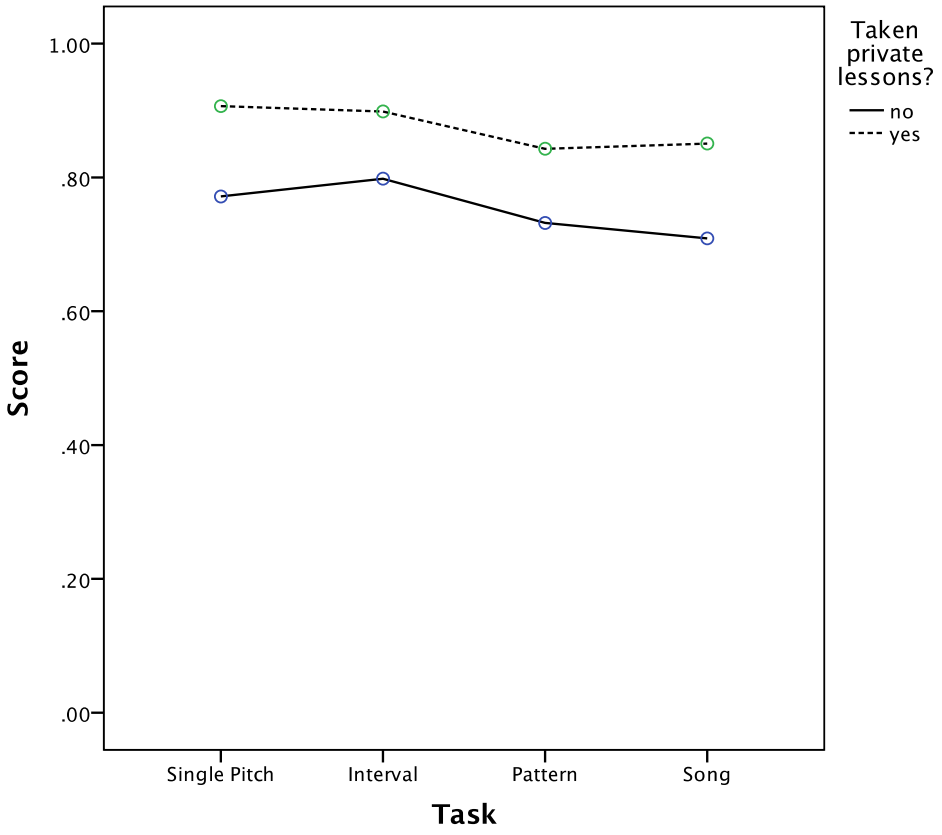


Figure 4.7. Task comparisons by history of private music lessons. This figure demonstrates participants' accuracy grouped by history of lessons.

Table 4.2

Task Performance by History of Lessons (SD)

		Single pitch	Interval	Pattern	Song
Solo condition	**History of Lessons**	.88 (.24)	.87 (.17)	.77 (.22)	.81 (.19)
	No History	.74 (.33)	.78 (.28)	.68 (.26)	.67 (.25)
Doubled condition	**History of Lessons**	.94 (.16)	.92 (.15)	.91 (.18)	.89 (.15)
	No History	.79 (.32)	.81 (.25)	.77 (.28)	.74 (.26)

n = 120 for pitch matching tasks and *n* = 119 for song singing tasks

significantly better than those without. There were no significant interactions, meaning that task performance was not differentiated by history of lessons. Performance means by history are plotted in Figure 4.7 and described in Table 4.2.

Question Number Two: What is the difficulty and discrimination of the different singing assessment tasks?

From a test construction perspective, the mean scores on a particular item can be referred to as difficulty level, difficulty index, or pass rate. The overall mean score for all items, representing the overall difficulty across all items, is .82 (*SD* = .21). The task means were presented earlier in this chapter, and are presented here again in Table 4.3 in terms of task difficulty. Difficulty levels were more similar across items in the doubled condition, as discussed in the previous section. In the solo condition, single pitches and intervals were easier than patterns and songs.

Table 4.3

Task Difficulty (SD)

	Single pitch	Interval	Pattern	Song
Solo condition	.82 (.29)	.83 (.23)	.74 (.24)	.75 (.23)
Doubled Condition	.87 (.25)	.88 (.21)	.85 (.24)	.82 (.22)

n = 120 for pitch matching tasks and *n* = 119 for song singing tasks

Item difficulty demonstrates how useful a test item is based on the item's ability to test a construct in participants. For example, a test item on which every participant is scored accurately is too "easy" to be useful for separating students based on the measured construct. For that kind of item, all students are "accurate" and thus those items do not distinguish between students. Likewise, an item on which every participant is scored "inaccurate" does not distinguish students by singing accuracy: the item is a poor indicator of this construct since it shows all students to be inaccurate. To interpret difficulty at the item level (Table 4.4), the following scale was used: Very Difficult, .00-.49; Fairly Difficult, .50-.69, Moderately Easy, .70-.89; Very Easy, .90-1.00 (Allen & Yen, 2001).

In the solo condition, means ranged from .67 to .89, which was interpreted to indicate these items ranged from fairly difficult to moderately easy. The two easiest items were the descending minor third interval, which appeared twice, Item 10, AAF#F#, *M* = .89 (.27), which replicated Item 6, AAF#F#, *M* = .86 (.27), for item stability. The most difficult items were the two items selected from pitch sequences in the song *Jingle*

Table 4.4

Item Difficulty

Task	Solo Single pitch					Solo Interval					Solo Pattern					Song
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**-**
M	.81	.83	.79	.81	.82	.86	.84	.80	.76	.89	.79	.74	.69	.67	.79	.75
SD	.37	.35	.40	.38	.36	.27	.30	.31	.31	.27	.31	.35	.33	.28	.29	.23
Task	Doubled single pitch					Doubled interval					Doubled pattern					Song
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**-**
M	.85	.91	.87	.88	.85	.85	.91	.86	.88	.85	.83	.86	.87	.84	.84	.82
SD	.36	.24	.32	.30	.33	.31	.30	.25	.22	.33	.28	.25	.26	.26	.30	.22

n = 120 for pitch matching tasks and *n* = 119 for song singing tasks

Bells. They were Item 14 in the pattern task, AGED, *M* = .67 (.28), followed by Item 13, F#ADE, *M* = .69 (.33). The standard deviations ranged from .23 to .40 for all items in the solo condition.

In the doubled condition, the means appear very different in comparison. They ranged from .82 to .91. Most were considered moderately easy and some were interpreted to be very easy. The easiest items were Item 2 in the single pitch task, DDDD, *M* = .91 (.24), and Item 7 in the interval task, DDF#F#, *M* = .91 (.30). The most difficult items were Item 11 (pattern), AF#AD, *M* = .83 (.28) followed closely by Item 14 (pattern), AGF#D, *M* = .84 (.26) and Item 15 (pattern), AF#AD, *M* = .84 (.30). The standard deviations ranged from .22 to .36 for items in the doubled condition. The lowest mean across all tasks is found in the solo pattern task and the highest means were found in the doubled single pitch and interval tasks.

Task discrimination.

One discrimination analysis appropriate for these scores is the discrimination index. This index can be used to demonstrate how well performance on each task differentiates high performance overall from poor performance overall, and is expressed by the difference between the difficulty index of the test items for the upper- and lower-scoring sub-groups when divided into thirds. A composite score was calculated using the sum of the averaged task-level scores plus song singing scores, and participants were ordered from low to high based on these scores. Next, the participants were ordered by their overall total score and divided into three groups of 40 participants, and the mean performance of the lowest scoring third was subtracted from the mean performance of the highest scoring third for each item. The resulting number is the discrimination index, defined as each task's ability to discriminate participants' overall performance (Table 4.5).

Table 4.5

Discrimination Index by Task

		Single pitch	**Interval**	**Pattern**	**Song**
Solo	Low Mean	.50	.62	.48	.54
	High Mean	1.00	.96	.92	.94
	D Index	**.49**	**.34**	**.44**	**.40**
Doubled	Low Mean	.64	.66	.61	.61
	High Mean	1.00	.98	.98	.98
	D Index	**.36**	**.32**	**.37**	**.37**

n = 40 for low group and *n* = 40 for high group

To interpret the index, I used the following scale: Satisfactory, >=.40; Little/no revision: .30-.39; Revise: .20-.29; Completely revise/Drop: <=.19 (Allen & Yen, 2001). The discrimination index helps determine how useful these tasks were for evaluating the construct(s) being assessed. Generally, items and tasks that contribute to the overall accuracy score for many participants should be retained, and those which are not should be considered for removal. None of these tasks require revision.

Item discrimination.

Individual items were also examined for discrimination ability for overall performance. Since only two song singing items were used, discrimination index at the item-level was only calculated for pitch matching items (Table 4.6). For item analysis, the items were ordered based on the total score (not task-level scores), as they were for task analysis. The solo single pitch task (Items 1-5) yielded items that were all

Table 4.6

Discrimination Index by Item

	Task	Single pitch					Interval					Pattern				
	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Solo	Low Mean	.46	.57	.44	.48	.51	.65	.58	.55	.57	**.71**	.49	.42	.43	**.55**	.53
	High Mean	1.00	.98	1.00	1.00	1.00	.98	.99	.96	.94	**1.00**	1.00	.93	.89	**.83**	.97
	D Index	.54	.41	.56	.52	.49	.33	.41	.41	.37	**.29**	.51	.52	.45	**.28**	.44
	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Doubled	Low Mean	.56	**.77**	.62	.66	.59	.58	**.75**	.66	**.76**	.57	.59	.64	.65	.63	.58
	High Mean	1.00	**.99**	1.00	1.00	1.00	.99	**.99**	.99	**.98**	1.00	.98	.98	.99	.99	.99
	D Index	.44	**.23**	.38	.34	.41	.42	**.24**	.32	**.22**	.42	.39	.34	.35	.36	.41

n = 40 for low group and *n* = 40 for high group

satisfactory discriminators. The solo interval task (Items 6-10), however, yielded less satisfactory scores, using the above-cited criteria. Overall, the item with the lowest discriminatory power was the interval task in Item 10, with an index of .29. The solo pattern task yielded mostly satisfactory scores but included one index that was lower than the others, Item 14. The doubled single pitch task (Items 1-5) showed even more contrasting indices. Item 2 yielded a low index of .23 because nearly everyone performed accurately, and because performance on Item 2 was less highly related to overall performance. Items 7 and 9 indicated similar descriptives.

Since some items warranted further consideration for discrimination ability, a second analysis was performed. As another measure of discrimination, I calculated the corrected item-total correlation of the item score to task-level score, rather than the overall total score, which indicates how well each item predicted task performance. The

Table 4.7

Corrected Item-Task Correlations

	Task	Single pitch					Interval					Pattern				
	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Solo	Mean	.81	.83	.79	.81	.82	.86	.84	.80	.76	.89	.79	.74	.69	.67	.79
	SD	.37	.35	.40	.38	.36	.27	.30	.31	.31	.27	.31	.35	.33	.28	.29
	Item-Task R	**.72**	**.49**	**.75**	**.61**	**.68**	**.67**	**.64**	**.68**	**.58**	**.65**	**.77**	**.57**	**.67**	**.45**	**.67**
	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Doubled	Mean	.85	.91	.87	.88	.85	.85	.91	.86	.88	.85	.83	.86	.87	.84	.84
	SD	.36	.24	.32	.30	.33	.31	.30	.25	.22	.33	.28	.25	.26	.26	.30
	Item-Task R	**.75**	**.47**	**.82**	**.84**	**.78**	**.78**	**.72**	**.54**	**.60**	**.69**	**.76**	**.82**	**.82**	**.84**	**.86**

n = 120

following scale was used for interpretation: Satisfactory, >=.30; Little/no revision: .19-.29; Revise: .00-.19; Completely revise/Drop: 0 or negative (Allen & Yen, 2001). All items in the solo response mode and all items in the doubled response mode were considered satisfactory discriminators based on this discrimination analysis (Table 4.7).

Question Number Three: What is the inter-relationship among different measures and how few tasks might be employed in a comprehensive measure of the skill of accurate singing?

To answer the third question regarding the interrelationship among measures and how few tasks might be employed in a comprehensive measure of accurate singing, I will attempt an optimum design of a comprehensive singing accuracy assessment by investigating which tasks highly correlate with each other and might be considered redundant in a testing situation, and to test the possibility of unidimensionality in singing accuracy. Additionally, I will demonstrate how many or how few items are required for a reliable assessment interpretation based on the analysis of the internal consistency of item scores.

Item stability.

Each set of five items was preceded by a practice example to familiarize the student with the task; these practice items were not scored. Next, the stimuli for the first item was introduced. The stimuli for the first item was replicated in the fifth item to examine how consistently students performed on an item when presented a second time (Table 4.8). There is a relationship among mean performance levels in each condition and the strength of association between the test and retested items. The

Table 4.8

Stability of Replicated Items (Pearson's r)

	Single pitch	Interval	Pattern
Solo condition	.59	.59	.69
Doubled Condition	.67	.79	.77

correlation is relatively higher in the doubled condition, where participants scored higher overall than in the solo condition.

Unidimensionality of tasks.

To test the unidimensionality of these pitch-matching and song singing tasks, an exploratory factor analysis was performed using the Principal Components Method with the task scores. A one factor model explains the variation in these scores, as demonstrated by the factor loadings displayed in Table 4.9. The exploratory factor analysis was an important part of this study because it suggested each task measures the construct of singing accuracy. Each of the four task types in both response modes was entered for a total of eight variables. All factors loaded at or above .72, which was the lowest load for solo song singing. Like the remaining factors, this was considered "high." As an indicator of reliability, the communality was expressed for each factor. The communality measures the percent of variance in any one variable that is explained by all the variables jointly. The communality for solo song singing was .52, and although this variable loaded highly in the factor analysis, the following should be stated: All these task and response combinations seem to represent a single construct called

Table 4.9

Factor Analysis of Tasks

Factor	Load	Communality
Doubled Single Pitch	.92	.84
Doubled Interval	.92	.85
Doubled Pattern	.92	.85
Doubled Song	.88	.78
Solo Interval	.86	.74
Solo Single Pitch	.85	.72
Solo Pattern	.83	.70
Solo Song	.72	.52

singing accuracy, and each can be said to indicate singing accuracy in individuals. However reliable, solo song singing may be the least representative of one's overall ability compared to other accuracy tasks.

Relationship among tasks.

The strength of association can demonstrate how close tasks are to evaluating singing accuracy in the same way. The correlation of pitch matching tasks and song singing tasks in both the solo and doubled response modes are presented in Table 4.10. Consistent with the results of the principle components analysis, the strength of these associations were all significant ($p < .001$) and ranged from .52 to .92. There is a pattern of higher correlation between the same tasks when presented in the solo and doubled response modes, as highlighted in the table. Also, within the doubled condition, there are higher correlations between all tasks, also highlighted.

Table 4.10

Correlations of All Tasks

		Solo				Doubled			
		Single	**Interval**	**Pattern**	**Song**	**Single**	**Interval**	**Pattern**	**Song**
Solo	**Single**	-	.73	.72	.61	.70	.72	.72	.68
	Interval		-	.74	.55	.77	.77	74	.66
	Pattern			-	.65	.69	.68	.65	.66
	Song				-	.52	.53	.57	.64
Doubled	**Single**					-	.90	.89	.82
	Interval						-	.92	.79
	Pattern							-	.84
	Song								-

Note: $p < .001$

Two of the five pattern items were modeled after pitch sequences found in the song task, *Jingle Bells,* so they were further compared. Item 13 was the pattern F#-A-D-E, which corresponds to the words in the song, *Jin-gle all the (way).* Item 14 was the pattern A-G-E-D, which corresponds to the second occurrence of the words in the song, *(one) horse o-pen sleigh.* In the solo condition, Items 13 and 14 are significantly correlated to solo song singing, $r = .57$ and $r = .37$, respectively ($p < .01$). However, regarding the correlation with the song singing, Item 13 did not statistically differ from Item 11, which was not from the song, and had the strongest association within the pattern task to song singing ($r = .58$, z = -.07, $p > .05$). Item 14 was less strongly associated with the song singing than Item 11 (z = -2.37, $p = .024$). In the doubled condition, Items 13 and 14 are significantly correlated to doubled song singing, $r = .70$

and $r = .68$, respectively ($p < .01$). For Item 13, this association was less strong than the strongest association within the pattern task to song singing, which again was Item 11, $r = .81$ ($z = -1.96$, $p = .025$). Item 14 was also less strongly associated with the song singing than Item 11 ($z = -2.36$, $p = .009$).

Item inclusion.

Last, it was important to determine how many items should be required for singing accuracy assessment using these types of tasks. It was theorized that five items per task would be sufficient based on task difficulty and sample size used in a previous study (Demorest & Nichols, 2012). Using fewer tasks, where possible, shortens the test duration and this is an important consideration in test design. Cronbach's *alpha* was calculated for each task as a measure of internal consistency and is shown in the shaded cells of Table 4.11. For solo single pitch matching, the removal of Item 2 increases the *alpha* coefficient only marginally. No removal increases *alpha* for the solo interval task. For the solo pattern task, *alpha* increases incrementally if Item 14 is removed. For the doubled single pitch task, *alpha* increases from .89 to .91 if Item 2 is removed. The *alpha* does not increase for the doubled interval or pattern tasks. Thus, the removal of items does not increase the *alpha* coefficients for some of the tasks; for other tasks, the removal of items only marginally improves the *alpha* coefficients.

The reader will recall from the difficulty and discrimination analyses that Item 2 was not a very useful item, but that it could be helpful to retain it for establishing singing accuracy ability (the student who does not sing this item accurately may be unlikely to

Table 4.11

Cronbach's alpha for Items Deleted

	Task	Single pitch					Interval					Pattern				
Solo	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
	Mean	.81	.83	.79	.81	.82	.86	.84	.80	.76	.89	.79	.74	.69	.67	.79
	SD	.37	.35	.40	.38	.36	.27	.30	.31	.31	.27	.31	.35	.33	.28	.29
	Alpha for each task			.84					.84					.83		
	Alpha if item deleted	.79	.85	.78	.82	.80	.80	.81	.80	.82	.81	.75	.81	.78	.84	.78
	Alpha if 3 items are used			.76					.76					.75		
Doubled	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
	Mean	.85	.91	.87	.88	.85	.85	.91	.86	.88	.85	.83	.86	.87	.84	.84
	SD	.36	.24	.32	.30	.33	.31	.30	.25	.22	.33	.28	.25	.26	.26	.30
	Alpha for each task			.89					.84					.93		
	Alpha if item deleted	.86	.91	.84	.84	.85	.77	.80	.34	.83	.81	.93	.92	.92	.91	.91
	Alpha if 3 items are used			.83					.76					.89		

sing others accurately). For the doubled interval task, all items contribute to task consistency (another way of explaining Cronbach's *alpha*).

The Spearman-Brown formula was used to calculate the reliability coefficient if less than five items were used for each task. A minimum of .75 was deemed acceptable (Allen & Yen, 2001), and all pitch matching tasks meet this criteria if a minimum of three items are used for testing.

Conclusion

A sample of 4th grade students ($N = 120$) completed an assessment of singing accuracy using three pitch matching tasks (single pitch, interval, pattern) plus song singing (*Jingle Bells*) in two response modes: solo and doubled. Students ranged in age from 9 years, 6 months to 10 years, 10 months, with a mean age of 10 years, 1.3 months. No participants indicated a history of hearing difficulty diagnosis. Fifty-five percent of the sample had exposure to private music lessons, and those participants evidenced higher accuracy scores than those who reported no history of private lessons.

Scoring was done by the researcher and one other evaluator and inter-rater reliability using a randomly selected 20% of participants from all participating schools was found to be very strong. Next, data were transformed to a scale of 0-1 for analysis. The results indicate children's singing accuracy performance varies significantly based on the nature of the singing task and the response mode used for assessment. Overall, students performed better in the doubled condition than the solo condition in both pitch matching and song singing tasks. For pitch matching, participants performed the solo interval task significantly better than the other solo single pitch or pattern tasks. However, the interaction contrasts with performance in the doubled condition. In that condition, performance across the three tasks did not vary. There was no main effect for Form A versus Form B, accounting for order effects of the solo and doubled conditions.

For song singing, an equal number of participants sang the solo song condition in or out of the given key. However, more participants sang the doubled condition in the

the given key as used by the model (n = 102). For pitch matching, the results indicated a hierarchy of difficulty and discrimination ability among singing assessment tasks. The discussion which follows this chapter will explore the value of items with low item-total correlations and discrimination ability. Each of the tasks used were examined and loaded onto one factor. This finding confirms construct validity by suggesting that these tasks each measure the same ability in 4th grade students.

Chapter 5: Discussion

The purpose of this study was to compare task-based variability in 4th graders' singing accuracy while singing alone or doubled by another voice. For this study, singing accuracy is synonymous with vocal pitch accuracy and was defined as one's ability to sing closer to the given pitch than an adjacent pitch. Participants were given pitch stimuli that were categorized by the number of unique pitches in a four-note sequence: single pitch, interval, and pattern, followed by a song singing task, *Jingle Bells*. Participants sang all tasks in two response modes referred to here as solo singing and doubled singing. For the pitch matching exercises, the solo singing was conceptualized as an echo response mode, whereas song singing was conceptualized as singing a song from memory.

The results indicate that singing accuracy varied by task because some tasks were more difficult than others; overall, these tasks as a group are representative of a single construct called singing accuracy. Further, the analysis of these tasks confirmed that each task alone is a satisfactory discriminator of singing performance. For certain purposes of assessment, it may not matter which task is chosen, but that all children are assessed within the same task. Since difficulty levels vary across items and tasks, a student's performance on an interval task should not be compared to another student's performance singing a song.

This discussion will begin with the main findings as they related to the three research questions. First, the variability of task and response mode will be discussed, followed by test construction and scoring concerns. Within each section, these findings are discussed from the perspective of the role of instruction in singing accuracy,

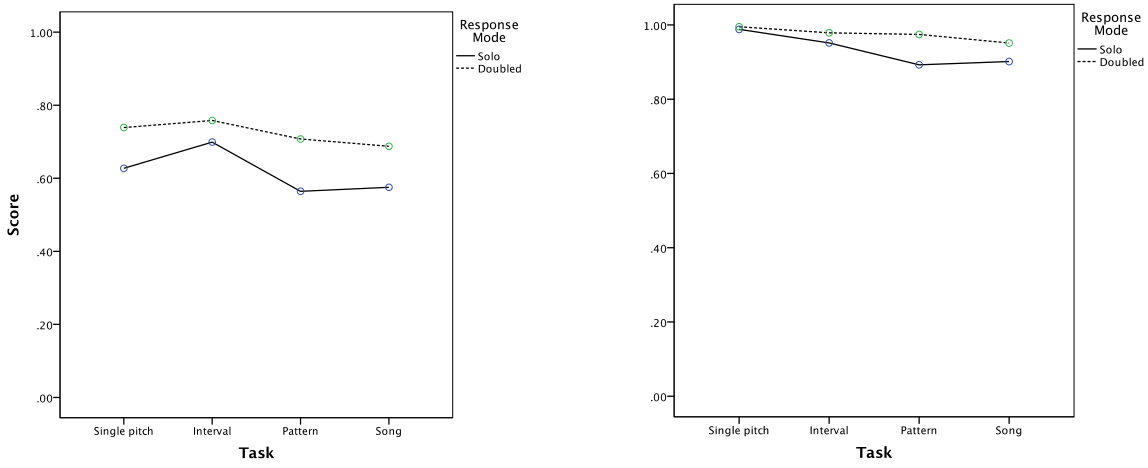
including implications for future research. Finally, a section describing implications for teachers is used to summarize the application of these findings for classroom use.

Variability by Task

The results of this study indicated that 4th grade students' performance did vary significantly based on the task they were given to sing. Specifically, they performed better on single pitch and interval tasks than they did on pattern tasks or singing a song from memory. This result would suggest that complexity, and to a certain extent memory, may play a role in singing accuracy development. The fundamental aspects of pitch matching must be in place before one can successfully sing a song, and students may acquire proficiency in simple tasks before they can be proficient in increasingly complex tasks. These results are similar to studies that compared multiple tasks in children (Demorest & Nichols, 2012; Welch et al., 1995; Welch et al., 1997) and adults (Pfordresher et al., 2010).

Several authors have hypothesized that tonal memory may play a role in a person's ability to sing accurately (Jones, 1971; Joyner, 1969; Petzold, 1963; Pfordresher & Brown, 2007), specifically suggesting that accurate tonal memory is required for the reproduction of longer or more complex sequences. The results here support that view. While all of the tasks were highly correlated, and all load on a one-factor model of singing accuracy, comparisons of student performance should be done within task (e.g., single pitch to single pitch) not between tasks (e.g., single pitch to song singing). Singing is a complex task and parsing possible components of the singing accuracy construct are important for advancing an understanding the development of this skill in children.

Task variability itself may be variable. In a previous study (Pfordresher & Brown, 2007), the "good" singers in one adult sample decreased in accuracy across the same three pitch matching tasks as used in this study, and the poor singers increased in accuracy across tasks. For comparison, the performance of 4th graders in this study was graphed separately for those who performed overall less than 90% accuracy and those who performed overall above 90% accuracy (Figure 5.1). Both groups evidenced stability across tasks in the doubled condition, but more variable performance in the solo condition. The implications for a memory effect are important since the doubled condition may have less memory demands than the solo condition. In the solo response mode, the higher achieving group had decreasing performance across pitch matching tasks, similar to the good singers in the previous adult sample. The lower



Low group High group

Figure 5.1. Performance of low and high groups. Mean performance plotted separately for high achieving and low achieving students.

achieving group got worse on patterns and songs, which suggests they were aided more by the cues provided with the additional pitch in the interval task.

Wise and Sloboda (2008) found “tone-deaf” and “non-tone-deaf” singers to decrease in accuracy across the same three tasks, which is similar to the high performing 4th graders in the present study. The low performers were more impaired in the solo single pitch and pattern condition, similar to a previous sample of kindergarteners (Demorest & Nichols, 2012). They seem to respond best to the interval task, which may represent a pitch sequencing compromise between complexity (not too long) and context (the two pitches in the interval task may have provided contextual cues). The lower group may have evidenced task performance more similar to the kindergarteners in a previous study (Demorest & Nichols, 2012) because their skills may be less developed than the higher group.

Neither group increased in performance across pitch matching tasks, as did the poor singers in the previous adult sample (Pfordresher & Brown, 2007). Doubled performance was relatively stable across tasks, but in the solo response mode, these 4th graders were like younger children in previous studies who performed single pitches better than some other tasks (Demorest & Nichols, 2012; Welch et al., 1995; Welch et al., 1997). Solo tasks were all higher for the high group than the interval task in the low group, which was the most accurately performed of the low group tasks. In other words, the most accurately performed solo task was not as accurately sung as the poorest doubled task.

Variability by Response Mode

Participants in this study performed more accurately in the doubled condition

than in the solo condition. This result differs from some of the previous research on doubling. Additionally, there was a significant task by response mode interaction. A comparison of the task means indicated that doubled singing results in more stable scores across tasks, whereas solo singing evidenced varying degrees of accuracy. Specifically, there was a cubic interaction since scores were higher for intervals than single pitches, but lower than both for patterns and higher again for song singing.

Researchers have compared students singing alone to singing along with a pre-recorded child model (Cooper, 1995), the researcher (Smale, 1987), with other students and the researcher (Goetze, 1985; Goetze & Horii, 1989), and in small groups of students (Green, 1994; Smith, 1973). The current study used a pre-recorded adult female voice and it is possible that the specific singing task used in previous studies may have influenced their results. Doubled singing as represented in this study is ecologically valid because it represents those occasions where the student sings along with one other individual, either the teacher or another student. Doubled singing in groups, as done in other studies, would represent those occasions where students sing along with other children, which is common in general music and choral music settings.

The results from this 4th grade sample are similar to those found by Green (1994) in grades 1, 2, 3, and 5, who sang the song *Bow Wow Wow*. In that study, students were asked to sing alone and to sing in groups of eight, and they sang better in the latter (doubled) condition. The participants in the current study performed along with a pre-recorded adult female voice and these results and those reported by Green (1994) contrast those from a 1st through 5th grade sample who evidenced no different singing when singing alone or along with the researcher on four-note patterns (Cooper, 1995) or

with 5th and 6th graders on the song *America* (Smith, 1973).

In other studies in which the student sang along with the researcher, individual singing was shown to be better than "unison" (doubled) singing (Goetze, 1985; Smale, 1987). One difference between those designs and the current study is that both of those studies used song phrases, rather than short patterns or whole songs. In the case of Goetze (1985), students were taught two melodic phrases and were recorded as they imitated the melody in response to the researcher's voice. The current study, though, emphasized the students' ability to sing whole songs and thus, students were asked to recall a song from memory. Additionally, the decision was made for song singing to be presented in solo and doubled conditions, but not in echo and doubled conditions. Students were not primed by vocal modeling prior to performance.

In comparison to the solo condition, some students sang the song task better in the doubled condition ($n = 60$); some performed the same ($n = 44$); and some performed worse ($n = 15$). Students who performed the doubled condition in a key other than the given key were more likely to evidence lower singing performance, reinforcing the suggestion that key and range selection is a critical component of test design because students exhibit varying degrees of voice control (Rutkowski, 1990). While solo singing performance may be predictive of doubled singing performance, and vice versa, neither is representative of the other in terms of difficulty. For pitch matching, accuracy was more stable when doubled (performance across tasks was similar in difficulty), and less stable in the solo response mode because pattern performance decreased more in the solo condition than it decreased in the doubled condition.

The comparative performance of students' song singing in the two response conditions may have differed from some previous studies which used a song task due to the specific nature of the song singing task. Green's (1994) participants sang a song like was used in the current study. Cooper (1995) also used a song and found no difference, although older students (5th and 6th grade) were sampled in that study, which may account for the contrasting finding. When phrases were used instead of whole songs (Goetze, 1985; Smale, 1987), individual singing was shown to be better, but songs were presented in call and response fashion after they were taught by the researchers. Green's (1994) participants sang in groups of eight peers, whereas the other studies asked students to sing with the researcher, which may have been more like singing along with the CD in the current study, with the exception of Goetze (1985) who had students sing with two peers.

The varying results in previous research have shown doubled singing to be significantly better than solo singing (Green, 1994) but also that solo singing is significantly better (Goetze, 1985; Goetze & Horii, 1987; Smale 1987). There are also studies which demonstrate no significant difference between these two response modes (Cooper, 1995; Smith, 1973) and the cause of these differences may be due to slight changes in the many variables used for testing singing. The stimuli model and presentation of the stimuli is different in each study, sometimes using an adult or child model and either pre-recorded or live. Sometimes children were doubled by a few or many of their peers or with the researcher, and though the ranges were fairly similar and always within the range of a fourth, fifth or sixth, the tasks varied in difficulty across the studies. Further, the age and experience of the participants ranged from pre-school to

grade 6. Last, these studies used differing scoring techniques. Thus, research in doubling, whether referred to as unison singing, or "accompaniment" (Wise & Sloboda, 2008), or "auditory feedback" (Pfordresher & Brown, 2007) is deserving of systematic study across all grades.

The potential for order effect.

In this study, there was no significant main effect for the order of response mode presentation. Overall, participants sang the solo condition with somewhat higher ability when they performed the doubled condition first, although not statistically significantly (p = .051). Performance on the doubled condition did not vary based on which order the test began. For those who sang the solo condition after the doubled condition, the results suggest that they could have been primed for improved performance by singing along with another voice. Since overall doubled performance was significantly higher (i.e., this condition was easier), these students could have performed marginally better because they were aided by the doubling experience prior to solo singing. Students who are still developing their singing may improve their ability to follow contours and stay in the given key by first practicing along with an accurate model. Then, their skills could be improved in solo singing. Successful solo singing could be seen as the ultimate singing goal for every student, and if so, could be expected to be assessed in every student in the music classroom.

Solo song singing may be a more advanced skill than doubled song singing since overall, participants performed better in the doubled condition. This suggests students are aided by an external stimulus while they sing, rather than inhibited by its presence. Perhaps doubled singing provides a memory aid which could be useful since

participants were not primed for singing *Jingle Bells*, at least for the participants who performed the solo condition prior to the doubled condition. Participants sang the solo condition better when they sang the doubled condition first (doubled performance varied only slightly based on order). Thus, it is possible that doubled singing primes students for improved solo singing. These effects may be different for the varied task types, which included echo (pitch matching) and memory (song singing) tasks. In future research, these effects should be further explored in this and other age levels, perhaps using pitch sequences of systematically increasing complexity.

Private Lesson History

Students in this study who had a history of private lessons performed with significantly greater accuracy than those who did not. In this case, the presence of this history may simply represent more musical experience which led to greater performance, or students who were provided private lessons were better. The effect of instruction in classrooms is deserving of exhaustive study, and future research should include lesson history as a variable in studies of instruction or remediation.

Murry and Zwirner (1991) studied experience in six adult female singers with a range of 2-12 years of voice training. There was no relationship between years of experience and singing accuracy using five attempts at a single pitch. This kind of task could be said to be simple because of its low memory demands (one pitch in a one-note presentation). However, in the current study, participants performed better on single pitches and intervals than patterns or the song. So the single note task mentioned above could be expected to be easier than other more lengthy tasks. It should be noted

that in the adult female study, singers with more experience improved more across the five attempts than those with less experience.

Previous research supports a link between musical experience and instrumental tuning accuracy. Yarbrough et al. (1997) replicated previous results with middle school students showing that tuning accuracy improved with additional years of ensemble performance experience for high school players. The authors note that the finding could be due to an improvement in student's tuning abilities or to student attrition where less successful students leave the ensemble setting over time. The only factor that significantly affected tuning performance in their sample of high school students was participation in private instruction. Similar results are reported in the current study, where fourth graders who indicated a history of private music lessons sang more accurately than those who did not. Like the previous authors, the current results can be used to suggest that students who take private lessons have more musical experience - that is, they have ensemble playing time and experience plus lesson playing time and experience. Their repertoire experience may be higher. It is even possible that private lesson experience contributes more to pitch accuracy than does ensemble participation. Still, the possibility exists that students who exhibited greater potential were more likely to be afforded private lessons in the first place, and their higher scores could be due to inherent differences in those students rather than due to a history of private lessons.

The authors of the previous tuning accuracy study suggest the possibility of a relationship between tuning accuracy and overall achievement, which is strengthened by the fact that high-performing students in tuning accuracy also received distinguished ratings at the solo and ensemble festival. Perhaps students who receive private

lessons evidence higher achievement overall, which could be due to extra time spent playing their instrument, extra tuition (one-on-one teaching), or additional factors like motivation or self-efficacy.

Task and Item Analyses

The specific items used in this assessment can be compared to previously reported hierarchies of tonal sequences. Sinor (1984) suggested that several elements may contribute to the difficulty of an item. She included eight factors for her four-note sequences including:

1. Half-steps (have been theorized to increase difficulty)
2. Contour (notes move down, up, or down and up)
3. Range
4. Repeated notes (number of unique pitches versus repeated pitches)
5. Stepwise only (does not include skips)
6. Successive skips (includes more than one skip in a row)
7. Number of notes (total length of sequence)
8. Location of skips (skips are located between which notes in the sequence)

While only one item was an exact match for the patterns in Sinor's study (Item 12 DEF#G), all of the items can be evaluated using Sinor's factors.

Item 1 (AAAA)

Half-steps	no
Contour	none
Range	unison
Repeated notes	yes
Stepwise only	n/a

Successive skips	no
Number of pitches	1
Location of skips	n/a

Item 2 (DDDD)

Half-steps	no
Contour	none
Range	unison
Repeated notes	yes
Stepwise only	n/a
Successive skips	no
Number of pitches	1
Location of skips	n/a

Item 3 (GGGG)

Half-steps	no
Contour	none
Range	unison
Repeated notes	yes
Stepwise only	n/a
Successive skips	no
Number of pitches	1
Location of skips	n/a

Item 4 (F#F#F#F#)

Half-steps	no
Contour	none
Range	unison
Repeated notes	yes
Stepwise only	n/a
Successive skips	no
Number of pitches	1
Location of skips	n/a

Item 5 (AAAA)

Half-steps	no
Contour	none
Range	unison
Repeated notes	yes
Stepwise only	n/a
Successive skips	no
Number of pitches	1
Location of skips	n/a

Item 6 (AAF#F#)

Half-steps	no

Contour	down
Range	minor 3rd
Repeated notes	yes
Stepwise only	no
Successive skips	no
Number of pitches	2
Location of skips	2-3

Item 7 (DDEE)

Half-steps	no
Contour	up
Range	major 2nd
Repeated notes	yes
Stepwise only	yes
Successive skips	no
Number of pitches	2
Location of skips	2-3

Item 8 (F#F#DD)

Half-steps	no
Contour	down
Range	major 3rd
Repeated notes	yes
Stepwise only	no
Successive skips	no
Number of pitches	2
Location of skips	2-3

Item 9 (DDGG)

Half-steps	no
Contour	up
Range	major 4th
Repeated notes	yes
Stepwise only	no
Successive skips	no
Number of pitches	2
Location of skips	2-3

Item 10 (AAF#F#)

Half-steps	no
Contour	down
Range	minor 3rd
Repeated notes	yes
Stepwise only	no
Successive skips	no
Number of pitches	2

Location of skips	2-3

Item 11 (AF#AD)

Half-steps	no
Contour	down-up-down
Range	major 5th
Repeated notes	no
Stepwise only	no
Successive skips	yes
Number of pitches	3
Location of skips	1-2, 2-3, 3-4

Item 12 (DEF#G)

Half-steps	yes
Contour	up
Range	major 4th
Repeated notes	no
Stepwise only	yes
Successive skips	no
Number of pitches	4
Location of skips	none

Item 13 (F#ADE)

Half-steps	no
Contour	up-down-up
Range	major 5th
Repeated notes	no
Stepwise only	no
Successive skips	yes
Number of pitches	4
Location of skips	1-2, 2-3

Item 14 (AGED)

Half-steps	no
Contour	down
Range	major 5th
Repeated notes	no
Stepwise only	no
Successive skips	no
Number of pitches	4
Location of skips	2-3

Item 15 (AF#AD)

Half-steps	no
Contour	down-up-down
Range	major 5th

Repeated notes	no
Stepwise only	no
Successive skips	yes
Number of pitches	3
Location of skips	1-2, 2-3, 3-4

Only one item incorporated half steps, and contour varied from unison to up or down, including directional changes. As explained in Chapter 3, these items were designed to be limited to a range of a fifth using both stepwise and non-stepwise sequences. There were skips located in all possible intervals and four items had successive skips. The pitch matching items are presented here from easiest to hardest in the solo condition:

1. Item 10: AAF#F# (sol-sol-mi-mi)
2. Item 6: AAF#F# (sol-sol-mi-mi)
3. Item 7: DDEE (do-do-re-re)
4. Item 2: DDDD (do-do-do-do)
5. Item 5: AAAA (sol-sol-sol-sol)
6. Item 4: F#F#F#F# (mi-mi-mi-mi) and Item 1: AAAA (sol-sol-sol-sol)
7. Item 8: F#F#DD (mi-mi-do-do)
8. Item 3: GGGG (sol-sol-sol-sol) and Items 11 and 15: AF#AD (sol-mi-sol-do)
9. Item 9: DDGG (do-do-fa-fa)
10. Item 12: DEF#G (do-re-mi-fa)
11. Item 13: F#ADE (mi-sol-do-re)
12. Item 14: AGED (sol-fa-re-do)

Next, the pitch matching items are presented here from easiest to hardest in the doubled condition:

1. Item 2: DDDD (do-do-do-do) and Item 7: DDEE (do-do-re-re)
2. Item 4: F#F#F#F# (mi-mi-mi-mi) and Item 9: DDGG (do-do-fa-fa)
3. Item 3: GGGG (sol-sol-sol-sol) and Item 13: F#ADE (mi-sol-do-re)
4. Item 8: F#F#DD (mi-mi-do-do) and Item 12: DEF#G (do-re-mi-fa)
5. Items 1 and 5: AAAA (sol-sol-sol-sol) and Items 6 and 10: AAF#F# (sol-sol-mi-mi)
6. Item 14: AGED (sol-fa-re-do) and Item 15: AF#AD (sol-mi-sol-do)
7. Item 11: AF#AD (sol-mi-sol-do)

The order of difficulty varied between the two response modes. In the item analyses in the previous chapter, a commonly-accepted cut-off was used to categorize the difficulty level of items in this assessment. As stated, the scale was: Very Difficult, .00-.49; Fairly Difficult, .50-.69, Moderately Easy, .70-.89; Very Easy, .90-1.00 (Allen & Yen, 2001). The items in the solo condition ranged from .67 to .89 (fairly difficult to moderately easy) and the items in the doubled condition ranged from .82 to .91 (moderately easy to very easy). Although these single pitches, intervals and patterns were all presented in four-note sequences, two items can be compared to to sequences used in Wolf's (2005) study of 560 students in grades K-2. The first sequence in common between the current study and that previous study are sol-sol-mi-mi compared to Wolf's sol-mi. This sequence was interpreted in the current study to be moderately easy (the third of four categories) and it was interpreted in Wolf's study to be

“moderate” (the second of three categories). The second sequence was mi-mi-do-do compared to Wolf’s mi-do. This sequence was interpreted in the current study to be moderately easy (the third of four categories) and it was interpreted in Wolf’s study to be easy (the first of three categories).

Wolf (2005) based her difficulty elements on different criteria than Sinor (1984). Sinor tested 40 tonal sequences and determined that difficulty level was influenced by by:

1. Number of notes (2 or 3)
2. Intervallic relationships
3. Range (low, middle, high)

The following factors did not seem to affect item difficulty:

1. Modality (major/minor)
2. Melodic contour (ascending/descending)
3. Harmonic function (tonic/dominant)

As said, these two sequences are difficult to compare because the current study used four-note presentations, whereas Wolf’s design used one note per unique pitch. She designated difficulty levels based on relative performance. Difficult items were designated as such if the mean item score was greater than one standard deviation away from the overall mean. Similarly, easy items were those items with a mean score greater than one standard deviation away from the overall mean. The remaining items were within one standard deviation of mean performance overall, and they were deemed to be moderate items. In the current study, a conventional scale was used to

label difficulty level based on its common use in the measurement literature (Allen & Yen, 2001) and was not based on relative means.

Task correlation.

The sample used in this study sang songs as accurately as they matched patterns, which differs from a previous sample of 5-year-olds that performed better on single pitches, glides, and patterns than two songs called *Fox in a Box* and *Five Green Frogs* (Welch et al., 1995). Notably, the participants in that study performed lower on the latter song, which reinforces the idea that songs vary in difficult like pitch matching items vary in difficulty. Two years later, using an age range of 5-7 year-olds, patterns were again found to be easier than two songs, which were taught to students as a part of the study (Welch et al., 1997). Their designed differed from the design contained herein because these 4th graders were not primed for performance on *Jingle Bells* prior to testing.

Comparing pitch matching tasks to song singing is more difficult than comparing two songs. It is generally not possible or desirable to match echo tasks and song singing based on length, use of text, and other features, which affect difficulty and discrimination. However, one way to examine the relationship between pitch matching and song singing is by the correlation found between performances of these two variables. Van Zee (1984) suggested that first graders do not always sing songs as well as patterns selected from those songs, but the present results indicate pattern performance and song performance to be significantly related in 4th grade students, but not based on whether the pattern was taken from the song. Also, the tasks were shown to represent one same construct in the exploratory factor analysis. Pitch matching and

song singing have also been found to be moderately correlated in kindergarteners (Demorest & Nichols, 2012), which suggests this relationship may be defined well before 4th grade.

Two of the patterns used in this assessment were designed to match patterns taken from the song *Jingle Bells*. The results suggested these items were not more strongly correlated with song singing than other patterns. Since all patterns were correlated to song singing, and the two patterns were taken directly from *Jingle Bells*, it is possible the patterns from the song were not recognizable as song extracts, or that patterns generally were not perceived as song fragments. This should be the case for the first half of the test, in which students were not yet primed by singing the song from which the two patterns came. It is notable that these phrases were considered difficult and they were also incomplete samples of the actual phrases (they contained just four notes). Welch et al. (1995) incorporated each of six patterns in two songs composed for the study, but they did not report the relationship between pattern and song performance beyond stating superior song performance.

Discriminability.

In this study, the single pitch task in the solo response mode discriminated overall performance best, followed by patterns, followed by the song task, and lastly, intervals. Within the doubled condition, intervals had the least discrimination ability (the other task types were similar). Roberts and Davies (1975) compared groups of “normal” and “monotone” participants and also reported satisfactory discrimination between the two groups using these tasks. They found single pitches to discriminate best, followed by intervals, free song (any song from memory), and melody (echo a given “tune”). The

difference in discrimination hierarchy could be due to sample differences: the previous study sampled “normal” and “monotone” singers for comparison, whereas participants in the current sample were not pre-screened for ability. The measure in the current study was the discrimination index, which expresses the discrimination of each item and task for overall performance, not for comparing groups of participants.

In the current study, ability was stratified by history of lessons. Thus, the discrimination indices of pitch matching tasks were further analyzed by lesson history and are presented in Table 5.1. The order of task discrimination varies for students with a history of lessons and those without. For participants without a history of lessons, the order of discriminability is solo single pitch, doubled single pitch and pattern, solo interval and pattern, and doubled interval. For participants with a history of lessons, the order of discriminability is solo pattern, solo single pitch, solo interval, doubled pattern, and doubled single pitch and interval, demonstrating yet another way that students with a history of lessons are different from those without.

Table 5.1

Discrimination Index by Task

		Single pitch	**Interval**	**Pattern**
Solo	**D Index - No Lessons**	.62	.50	.50
	D Index - Lesson HIstory	.35	.25	.39
Doubled	**D Index - No Lessons**	.54	.48	.54
	D Index - Lesson History	.18	.18	.22

n = 40 for high group and *n* = 40 for low group

Specific items were satisfactory discriminators of singing performance at the task level and for performance overall, indicating these items were useful for separating students based on ability. The item-total correlations were also used to evaluate how well performance on an item related to performance on the given task, with satisfactory results. Item 2 (D-D-D-D) in the doubled condition had a low discrimination index of .23, relative to other items. This item was the least difficult item for participants, which can be used to explain why it did not discriminate between low- and high-performing participants: nearly everyone did well on this item. Still, performance on Item 2 was moderately correlated to the task score, $r = .47$. However, when used in the solo condition, this item was useful because its discrimination index was satisfactory (.41) and was moderately correlated to interval task performance, $r = .49$. The discrepancy in difficulty between these identical items suggests participants do not behave similarly in the two response modes (solo and doubled). Indeed, all items' discrimination ability varied between the two response modes.

Some items with lower discrimination indices were also low in range, which could indicate that range may affect singing accuracy performance. These items could be retained, though, because students who can perform these items may be more likely to perform others. The dominant functioning pitch (a fifth interval above the tonic pitch) could be expected to be easier, but was perhaps moderated by participants' range development, which could explain why the mediant pitch was represented in an item with low discrimination ability.

Challenges in assessment.

More students sang *Jingle Bells* in the given key when they sang in the doubled

condition than in the solo condition. The actual key chosen by the participant was not recorded by the scorer; only a dichotomous measure of singing in the given key was used. Some sang the solo condition higher than the given key, but the scorers noted that more sang lower than the given key. In this study, the beginning pitch was the third scale degree, F-sharp, and was given to participants prior to singing. The song was presented in the key of D, like the pitch matching tasks, and because of this, students were primed to sing in the key of D. However, at the time of reference pitch presentation for song singing, tonic was not re-established nor was an arpeggio or other chord presentation used. Further, a synthesized pitch pipe was used to generate the reference tone. The synthesized tone was used to avoid priming students for a song singing task that was designed to be sung from memory. Specifically, the researcher wished to present only the starting pitch, absent of vocal cues, as sometimes occurs in music settings (i.e., “Sing back this phrase” and teacher plays starting note on piano). That said, some students may have demonstrated a presumably higher ability to sing in the given key given additional cues or support, normally in the form of vocal modeling, text modeling, or modeling of the first phrase.

A second assessment consideration was the scoring treatment of vocal scooping, which was prevalent among participants in this study. Students were not penalized for scooping, which was more frequently found in some students than others. If a student sang closer to the given pitch by the end of the response time for each pitch, it was scored as accurate. In other words, raters were not asked to “average” the pitch but to identify if the singer did or did not sing close enough to the given pitch even if they spent part of the pitch duration scooping. A computerized measurement of cent deviation

might necessarily have averaged the sung frequency and indicated a lower mean frequency score than was dichotomously indicated by the human judges in this study (Pfordresher & Brown, 2007). Scooping was not operationally defined as an accuracy feature and did not affect the accurate scoring of participants, but if computerized scoring was used, averaging frequencies may have shown certain participants to have lower accuracy. Researchers employing deviation scoring usually mitigate the scooping phenomenon by using a middle portion of the sung response, but since the pitch durations in these stimuli were approximately one second, participants who scooped sometimes spent half or even more of the pitch duration doing so.

Thirdly, the song singing of these participants was rated using an eight-point scale (Wise & Sloboda, 2008). The scale was selected because it had been shown to be useful for singing accuracy evaluation in younger children (Demorest & Nichols, 2012) and adults (Wise & Sloboda, 2008). Problematically, one student sang the song in an inconsistent way: the first part of the sung response would have been given a score in the middle of the scale, but the second part of the sung response was chanted. Scored solely on the first half, the participant would have scored a level of five or six. Scored solely on the second half, the participant would have scored a level of one. For this student, the item mean was substituted for the solo singing score. For future studies, this type of inconsistency should be planned for. The student could be a Level 6 singer, since (s)he briefly demonstrated that level, but the student finished by chanting. Several possibilities exist to explain this change halfway through singing. The range of the song could have affected the participant's ability; when the range went slightly higher, the student chanted rather than singing lower than the accurate notes

(e.g., Rutkowski, 1990). Next, the student may have lacked an advanced ability to recall a succession of pitches beyond what (s)he demonstrated (e.g., Joyner, 1969). Last, the student could have lacked the motivation to continue singing and chose to continue by chanting instead (e.g., Goetze et al., 1990).

Internal consistency.

Cronbach's *alpha* was used as a measure of internal consistency at the task level, and each task was found to be satisfactory (*alpha* > .80). Very high reliabilities (> .95) would have indicated that these items are highly redundant, which would have been plausible given the one-factor model represented in the factor analysis (these items all measure the same construct and are similar in that way). In this case, the items did function similarly to one another, but not identically. The Spearman-Brown formula was used to determine that the internal consistency would remain at or above 0.75 for each task in both response modes if three items in each task were used. Internal consistency would have been above .80 for each task if four items were used, rather than the five items that were used in this design. Future researchers could be assured satisfactory internal consistency using four items per task, and if time constraints exist, an argument can be made for using three items per task.

For the development of this instrument, each pitch matching task was conceptualized as different from the others because they had to be: there was no previous evidence otherwise. For example, Pfordresher & Brown (2007) found performance to increase across the three pitch matching tasks for accurate singers and to decrease across the tasks for poor singers. Demorest & Nichols (2012) found performance to be significantly better in the interval task compared to single pitch and

pattern tasks. As in the current study, Demorest & Nichols (2012) preceded the items within each task with a practice item. The satisfactory internal consistency coefficients indicate that in a future assessment, the order of items across tasks possibly could be varied, at least in cases where the pitch sequence is constant in number (all pitch matching tasks were four notes in length). Varying the item order, in which items are followed by items from other tasks, could allow for further testing of order effects. For this study, all tasks were given in the same order, varied by order of response mode. Some participants sang in the doubled condition first and others sang in the solo condition first.

The results of this study confirm previous findings that these tasks are useful discriminators of singing accuracy. Roberts & Davies (1975) reported those data based on the comparison of a sample of "normal" singers (non-monotones) and a sample of "monotone" singers (their sample was screened for ability). This study replicates their discrimination ability finding in a sample that was not screened for any ability, and using different stimuli. Generally, singing accuracy studies have lacked in reporting the results of item analyses. In future studies, researchers might consider reporting the difficulty and discrimination indices as well as reliability because the expression of those statistics indicates that a researcher has been attentive to item-level test construction.

Implications for Teachers

Teachers are charged with improving musical skills like singing accuracy in children and there are several implications for teachers found in these results. The main findings of this study are that:

1. Doubled singing was easier than solo singing;

2. Performance in one task should only be compared to performances using the same task;
3. Summative assessment should include varied tasks, and not use only solo song singing, which was least representative of a children's performance across all the tasks used in this study.

Solo versus doubled singing.

Doubled singing was more accurate than solo singing across all tasks. For teachers this means that in the upper grades, students may be more likely to perform well when singing along with an accurate adult model. The stimuli in this study was provided by a pre-recorded track, which was used to ensure identical stimuli presentations for every child. It should be noted that vocal performance doubled by an accurate model might be very different from singing that is doubled by an inaccurate model. This is important to note since children are often asked to sing with their peers in school, and can often be found singing and chanting games with peers outside of school. Some peers may sing accurately but others may sing less well, and singing doubled by a poor singer would be theorized to negatively impact a student's performance.

For example, Demorest and Clements (2007) called some participants in their sample of junior high boys "inconsistent singers." The inconsistent singers performed better in a condition where they were given a tonal context prior to pitch matching, and the authors suggest those singers made use of their perceptual skills to better guide their pitch matching. Doubled singing in the current study could have helped some students to sing better in the same way. Overall, these fourth graders performed better

in the doubled condition, which may mean they used their perceptual skills to adapt their singing to the accurately-sung stimuli. Given a less accurate model, like a poor-singing peer, doubled singing would be much less likely to aid singing accuracy.

This kind of doubling—or feedback—variability has been studied before. Pfordresher and Brown (2007) asked adults to sing in three feedback conditions. First, participants sang pitch matching tasks normally, where they could hear themselves singing - even if more quietly since headphones were covering their ears (their study used headphones to play the stimuli). Second, they sang tasks in an "augmented" feedback condition, where the pre-recorded synthesized stimuli was played again while participants sang (as in the current study), though it was played more quietly than when it was first sounded as stimuli. Last, they sang tasks in a "masked" feedback condition, where pink noise was sounded at approximately 80 dB during singing.

The augmented feedback condition above could be conceptualized as doubling by a synthesized (non-human) sound. Like in the current study, those adult participants heard an accurate, pre-recorded representation of the pitches as they sang. The high group in this fourth grade sample performed like the "good" singers in that adult study because they sang less well as complexity increased. However, in this study, the high group's difficulty with increased complexity was attenuated in the doubled condition, very much like the adults in Pfordresher and Brown's "augmented" feedback condition.

Next, for purposes of comparing an individual's accuracy to prior performance or to another student's performance, the same task(s) must be used. In this study, singing accuracy was variable by task which means that performance on one task cannot be compared to performance on another task - they are expected to be different. For

teachers documenting formative assessment of student progress, employing the same tasks and possibly identical items is important for comparing performance over time or between students. Teachers might choose tasks that will work throughout the year, though the teacher must make note of register concerns, as children older than these fourth graders may experience voice change. The matter of register is important because a teacher could inadvertently be testing voice development rather than singing accuracy if a singer does less well on a test of singing accuracy because some notes fall outside the singing range of the student.

In the current study, the low group performed better on the interval task than the others, in both the solo or doubled condition. The results of these low group fourth graders is more similar to the kindergarteners in a previous study (Demorest & Nichols, 2012) where children performed best on the interval condition. For lower performing fourth graders, and for developing singers like kindergarteners, the interval task may indicate a complexity/contextual compromise between the low complexity of the single pitch condition and the context established by the addition of pitches. They may have been aided by their perceptual skills like the inconsistent singers who performed better in a contextual condition in the previous study of junior high boys (Demorest & Clements, 2007), and teacher may need to look for that level of complexity compromise in each individual student.

The third finding was that summative assessment should include varied tasks and not use solo song singing only, which was least representative of a children's performance on all the tasks used in this study. Summative assessment, which is used to summarize and often report student development at a point in time, will most often be

used to describe student's general or specific accuracy development. Occasionally, music teachers may use tests that report only ability on solo or doubled song singing, but more often teachers will be evaluating progress using varied tasks like single pitch, interval, and pattern pitch matching. A test that incorporates only interval pitch matching items may show lower performing students to perform more closely to the upper performing students. A test that incorporates only doubled singing may indicate stable task-based (non)variability that exists differently when the children are singing solo. Therefore the most descriptive summative assessment would use a range of task types and response modes to report student ability.

The results of this study indicate a hierarchy of task difficulty that teachers may find useful for developing or remediating singing skills in their students. For this fourth-grade sample, the order of difficulty from easiest to hardest:

1. Doubled intervals
2. Doubled single pitches
3. Doubled patterns
4. Solo intervals
5. Solo single pitches and Doubled song (*Jingle Bells*)
6. Solo song (*Jingle Bells*)
7. Solo pattern

The results suggest that any of these four task types can be used to measure the construct of singing accuracy. Additionally, due to the significant correlations found among tasks, a student with high achievement on one task is more likely to do well on others. Students may have the most success transferring task-based skill in this order:

doubled intervals, single pitches, patterns and songs, then solo intervals, single pitches, songs and patterns (see Table 4.3). Wolf (2005) compared pitch sequences and determined all of her two-note "patterns" were categorized as "easy." She also determined all easy patterns were low in range. Teachers could substitute song singing for patterns, if it is more ecologically valid, to approximate similar results, and the same can be said for using single pitches and intervals interchangeably. The benefit of using patterns for assessment is that teachers may find the use of short patterns to be briefer than song singing and that patterns may be easier to score. However, teachers may find song singing to be more ecologically valid since song singing is the ultimate application of tuneful singing.

Teachers may wish to know that of these tasks, the least predictive way to evidence singing accuracy may be to ask students to sing a song alone, since the task sharing the least variance in the factor analysis was solo song singing. Another recommendation is that teachers include easier items, like Item 2 in this test (D-D-D-D), if they are useful for transferring skill to other, more difficult items. Alternatively, teachers may want to determine whether students can perform easier items on a test that also includes more difficult items, identifying what the student can and cannot do.

Since pitch matching skill may improve over time, teachers should teach for transfer of easier pitch matching skills to more difficult song singing tasks. Van Zee (1984) found several specific problems to which the vocal music teacher may be able to relate: 1) Confuse speaking and singing in low register, 2) Don't hear direction correctly, 3) Confuse loud and high, 4) Switch back and forth between speaking and singing (poor vocal control), 5) Little vocal flexibility - no head tones, 6) Speaking voice low pitched,

heavy quality. These problems are listed here because Van Zee grouped students by problem so that each could be specifically remediated. In other words, the researcher found it possible and useful to differentiate instruction for students in an effort to remediate singing skills. Phillips and Aitchison (1997) demonstrated that certain component skills of singing like breath support, breath control, and range extension can be improved in grades 4-6 (but without improving singing accuracy) and teachers-in-training should be given practice in how to remediate singing skills.

Given these and previous findings, teachers may want to offer as much individual singing and instrument instruction as possible. This kind of instruction and assessment could be tailored to the changing needs of groupings of students (Van Zee, 1984) following an understanding that singing accuracy may develop out of an awareness of pitch contour followed by pitch matching (Welch et al. 1997) in singers who are developing use of the singing voice (Rutkowski, 1996).

It is important to note that specific tasks vary in difficulty, as do items within those tasks, and teachers should be aware of that variability when making instructional decisions based on formative assessment. For example, teachers who use long pitch sequences should understand that these may be more difficult than shorter sequences, like intervals, and choose sequences based on the needs of their students. Some students may have greater success singing with others (doubling), either with the teacher or one or more other students. Teachers can identify which students evidence differentiated performance between the two conditions and target instruction for them. For others, teachers can design singing tasks at a more advanced solo singing level. In this way, teachers can use discriminating singing tasks to differentiate instruction for

students, who have varying needs.

Conclusion

Research in children's singing accuracy is important because tuneful singing is a fundamental skill for basic musicianship. Whether or not students continue musical participation after the elementary school years, they benefit from the general ability to sing in either formal or informal settings. Children are expected to learn to sing well, per the national, state, and local teaching standards in many places. Since children are engaged in singing as part of music class, it can be assumed they will be assessed in singing, and an understanding of task-based variability in singing accuracy is important for the many purposes of testing in the classroom. The tasks used in this study are all significantly correlated, but the picture they paint of singers varies by difficulty of task: performance on any one task can be used to predict ability on another task, but not perfectly.

It seems that singing accuracy in these fourth grade children was better on less complex tasks compared to more complex tasks and most, but not all, participants performed better in the doubled condition. Some participants evidenced doubled singing ability closer to their (lower) solo singing ability. Each student's singing profile varies, and children in classrooms should be considered unique from one another, and assessed accordingly. Some students habitually scooped pitches, which may be a characteristic borrowed from popular music styles or may represent a student's still-developing singing skill (advanced students may have better control of pitch onset), which has implications for testing. Scooping does not seem to represent a singing accuracy deficit, but rather a singing development identifier.

In research terms, performance on one task is predictive of performance on others. From a curriculum and instruction perspective, if the teacher improves a student's ability using one task, the student's ability may improve on other types of tasks. When ability does not develop on other tasks, teachers can use tasks where students excel for transferring skill to other areas. To build confidence in children, teachers can build on doubled singing success by progressing to solo singing exercises. Likewise, teachers can advance from single pitches and intervals to patterns, which were shown here to be more difficult for this sample, but student performance should always be compared to performance on the same task.

The participants in this study who had a history of private music lessons demonstrated significantly higher performance than those who did not, and this finding represents a causality dilemma for future research. One explanation is that these students may have performed better because they had more musical history, some taking voice lessons or lessons on instruments. The other explanation is that these students previously displayed higher skill, aptitude, or enjoyment, and as a consequence they pursued—or were afforded—private instruction. The rationale in the first case would be based on a student and parent's desire to improve musical skill, likely perceived by the student or parent to be valuable. In the second case, a student's "talent" is identified and so strengthened by lessons. Component skills for singing have been improved in children (e.g., Apfelstadt, 1984; Phillips & Aitchison, 1997; Phillips & Aitchison, 1999; Porter, 1977; Roberts & Davies, 1975; Rutkowski, 1996), but the actual effect of private instruction in these participants is unknown.

The proportion of time spent singing intervals and patterns in schools compared

to time spent song singing is difficult to measure because of the song singing that may occur in the general classroom and other domains (i.e., student-initiated informal singing). This proportion likely depends on the individual music teacher and on the methodologies and instructional systems used in the music classroom. The participants in this study came from various schools and perhaps represent abilities developed by various music teachers. That said, these tasks might evince ability differently in students with more or less experience singing each task type and response mode, as suggested by the superior performance of students with a history of lessons.

Singing is a complex human ability, and one that develops differently in each individual. The act of singing requires much more than accurate pitch matching in order to be successful because singers must also execute timely rhythms and use appropriate timbre, diction, and posture. These elements of singing were not studied as a part of the current design, but are important considerations for teaching singing to children. Developing tuneful singing is important as one of several aspects of good singing, and teachers should strive to develop all the component skills of singing, like breath support (Phillips & Aitchinson, 1997) as well as the other elements of singing cited above.

The easiest item in this test was lowest in range, giving credit to the suggestion that some students' voice development must be facilitated in terms of singing range, which is much larger than typical speaking voice usage. This skill is challenging to assess because students vary in ability, testing is time-consuming for classes that meet once per week, and it presents some real-time scoring challenges. Some teachers deem singing important enough to assess regularly, and to do this, they routinely design their own assessments. These teachers need valid and reliable tools for developing

and assessing these and other musical skills so that their students can enjoy the benefits of singing later in life. Teachers are encouraged to compare student performance within-task, to use summative assessments that incorporate as many tasks as is feasible, and to consider solo and doubled song singing as separate skills that must be developed in every child.

References

Allen, M., J. & Yen, W. M. (2001). *Introduction to Measurement Theory.* Long Grove: Waveland Press.

Apfelstadt, H. (1984). Effects of melodic perception instruction on pitch discrimination and vocal accuracy of kindergarten children. *Journal of Research in Music Education*, *32*(1), 15–24. doi:10.2307/3345277

Atterbury, B. W., & Silcox, L. (1993). The effect of piano accompaniment of kindergartners' developmental singing ability. *Journal of Research in Music Education, 41*(1), 40-47. doi: 10.2307.3345478

Bergee, M. (2012). *An application of rasch modeling to the development of a rhythm-reading measure*. Unpublished manuscript.

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

Boardman, E. L. (1964). *An investigation of the effect of pre-school training on the development of vocal accuracy in young children*. Retrieved from ProQuest Digital Dissertations. (AAT 6408354)

Bradshaw, E., & McHenry M. A. (2005). Pitch discrimination and pitch matching abilities of adults who sing inaccurately. *Journal of Voice 19*(3), 431-439. doi:10.1016/ j.jvoice.2004.07.010

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

Clements, A. (2002). *The importance of selected variables in predicting student participation in junior high choir*. Retrieved from ProQuest Digital Dissertations. (AAT 3062930)

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*(3), 222–231. doi:10.2307/3345637

Cuddy, L. L. (2005). Musical difficulties are rare: A study of “tone deafness” among university students. *Annals of the New York Academy of Sciences, 1060*(1), 311-324. doi: 10.1196/annals.1360.026

Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority. *Annals of the New York Academy of Sciences*, *1169*(1), 99–107. doi:10.1111/j. 1749-6632.2009.04558.x

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *The Journal of the Acoustical Society of America*, *121*(2), 1182–1189. doi:10.1121/1.2427111

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

Demorest, S. M., & Clements, A. (2007). Factors influencing the pitch matching of junior high boys. *Journal of Research in Music Education*, *55*(3), 190–203. doi: 10.1177/002242940705500302

Demorest, S. M., & Nichols, B. E. (2012). *The impact of focused instruction on kindergartener’s singing accuracy*. Unpublished manuscript.

Flowers, P. J., & Dunne-Sousa, D. (1990). Pitch-pattern accuracy, tonality, and vocal range in preschool children's singing. *Journal of Research in Music Education 38*(2), 102-114. doi: 10/2307/3344930

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. (1983). The relationship of pitch matching and pitch-discrimination abilities of preschool and fourth-grade students. *Journal of Research in Music Education*, *31*(2), 93–99. doi:10.2307/3345213

Goetze, M. (1985). *Factors affecting accuracy in children's singing*. Retrieved from ProQuest Digital Dissertations. (AAT 8528488)

Green, G. A. (1990). The effect of vocal modeling on pitch matching accuracy of elementary schoolchildren. *Journal of Research in Music Education*, *38*(3), 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*(2), 105–114. doi: 10.2307/3345186

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

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*(1), 64–76. doi:10.2307/3345525

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

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*(2), 183–194. doi:10.2307/3343822

Joyner, D. R. (1969). The monotone problem. *Journal of Research in Music Education*, *17*(1), 115–124. doi:10.2307/3344198

Klemish, J. (1974). Treating the uncertain singer through the use of the tape recorder. *Bulletin of the Council for Research in Music Education*, *37*, 36–45.

Kramer, S. J. (1986). The effects of two different music programs on third and fourth grade children's ability to match pitches vocally. Retrieved from ProQuest Digital Dissertations. (AAT 8524224)

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. Retrieved from ProQuest Digital Dissertations. (AAT 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*(1), 35–47. doi: 10.2307/3345758

Martinez-Castilla, P., & Sotillo, M. (2008). Singing abilities in Williams Syndrome. *Music Perception, 25*(5), 449-469. doi: 10.1525/mp.2008.25.5.449

McCoy, C. (1997). Factors relating to pitch matching skills of elementary education majors. *Journal of Research in Music Education*, *45*(3), 356–366. doi: 10.2307/3345531

Moore, R. S. (1991). Comparison of children's and adults' vocal ranges and preferred tessituras in singing familiar songs. *Bulletin of the Council for Research in Music Edcuation, 107*, 13-22.

Moore, R., Fyk, J., Frega, A., & Brotons, M. (1995). 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 for Research in Music Education*, *127*, 127–135.

Murry, T., & Zwirner, P. (1991). Pitch matching ability of experienced and inexperienced singers. *Journal of Voice*, *5*(3), 197–202. doi:10.1016/S0892-1997(05)80187-0

Muse, M. B. (1993). *A comparison of two methods of teaching singing to primary children: An attempt to determine which of two approaches to teaching singing is more effective*. Retrieved from ProQuest Digital Dissertations. (AAT 1358052)

Petzold, R. G. (1963). The development of auditory perception of musical sounds by children in the first six grades. *Journal of Research in Music Education*, *11*(1), 21–43. doi:10.2307/3344529

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness." *Music Perception*, *25*(2), 95–115. doi:10.1525/mp.2007.25.2.95

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *The Journal of the Acoustical Society of America*, *128*(4), 2182–2190. doi:10.1121/1.3478782

Phillips, K., & Aitchison, R. (1997). Effects of psychomotor instruction on elementary general music students' singing performance. *Journal of Research in Music Education*, *45*(2), 185–196. doi: 10.2307/3345579

Phillips, K., & Aitchison, R. (1999). Second-year results of a longitudinal study of the relationship of singing instruction, pitch accuracy, and gender to aural acuity, vocal achievement, musical knowledge, and attitude towards singing among general music students. *Contributions to Music Education*, *26*, 67–85.

Phillips, K., & Doneski, S. (2011). Research on elementary and secondary school singing. In R. Colwell & P. Webster (Eds.), *MENC Handbook of Research on Music Learning, Volume 2: Applications*. New York, NY: Oxford University Press.

Phillips, K., & Vispoel, W. (1990). The effects of class voice and breath-management instruction of vocal knowledge, attitudes, and vocal performance among elementary education majors. *The Quarterly Journal of Music Teaching and Learning*, *1*(1-2), 96–105.

Phillips, K., Aitchison, R., & Nompula, Y. (2002). The relationship of music aptitude to singing achievement among fifth grade students. *Contributions to Music Education*, *29*, 47–58.

Porter, S. Y. (1977). The effect of multiple discrimination training on pitch matching behaviors of uncertain singers. *Journal of Research in Music Education*, *25*(1), 68–82. doi:10.2307/3344846

Price, H.E, Yarbrough, C. Jones, M., & Moore, R. S. (1994). Effects of male timbre, falsetto, and sine-wave models on interval matching by inaccurate singers. *Journal of Research in Music Education, 42*(4), 269-284. doi:10.2307/3345736

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*(4), 227–239. doi:10.2307/3344852

Rutkowski, J. (1983). *Development of a rating scale to assess individual children's use of the vocal instrument*. Unpublished manuscript.

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

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*(4), 353–368. doi:10.2307/3345447

Rutkowski, J., & Barnes, P. J. (2000, March). *Validation of the Singer Accuracy Measure: Versions 2.1 and 2.2.* Paper presented at the Music Educators National Conference, Washington, DC.

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

Salvador, K. (2010). How can elementary teachers measure singing voice achievement? A critical review of assessments, 1994-2009. *Update: Applications of Research in Music Education*, *29*(1), 40–47. doi:10.1177/8755123310378454

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

Sinor, E. (1984). *The singing of selected tonal patterns by preschool children.* Retrieved from ProQuest Digital Dissertations. (AAT 8501456)

Sloboda, J. A. (2005). Quantifying tone deafness in the general population. *Annals of the New York Academy of Sciences, 1060*(1), 255-261. doi: 10.1196/annals.1360.018

Smale, M. J. (1987). *An investigation of pitch accuracy of four-and five-year-old singers*. Retrieved from ProQuest Digital Dissertations. (AAT 8723851)

Smith, R. (1973). *Factors related to children's in-tune singing abilities*. Retrieved from ProQuest Digital Dissertations. (AAT 7411404)

Tsang, C. D., Friendly, R. H., & Trainor, L. J. (2011). Singing development as a sensorimotor interaction problem. *Psychomusicology: Music, Mind & Brain*, *21*(1-2). doi:10.1037/h0094002

Van Zee, N. (1984). An investigation study of young children's vocal problems and remedial needs. *Missouri Journal of Research in Music Education*, *5*(2), 55–71.

Watts, C., Moore, R., & McCaghren, K. (2005). The relationship between vocal pitch matching skills and pitch discrimination skills in untrained accurate and inaccurate singers. *Journal of Voice*, *19*(4), 534–543. doi:10.1016/j.jvoice.2004.09.001

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

Welch, G., Himonides, E., Saunders, J., Papageorgi, I., Rinta, T., Stewart, C., Preti, C. et al. (2008). *The national singing programme for primary schools in England: An initial baseline study overview, February 2008*. Institute of Education, University of London. Retrieved November 2, 2011, from http://www.imerc.org/papers/nsp/nspfeb08.pdf

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

Western, B. (2002). *Fundamental frequency and pitch matching accuracy characteristics of first grade general music students.* Retrieved from ProQuest Digital Dissertations (AAT 3073408)

Wise, K. J., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined “tone deafness”: Perception, singing performance and self-assessment. *Musicae scientiae*, *12*(1), 3–26. doi:10.1177/102986490801200102

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

Wurgler, P. S. (1990). *A perceptual study of vocal registers in the singing voices of children*. Retrieved from ProQuest Digital Dissertations. (AAT 8501456)

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*(1), 30–38. doi:10.2307/3345772

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

Yarbrough, C., Morrison, S. J., & Karrick, B. (1997). The effect of experience, private instruction, and knowledge of directional mistunings on the tuning performance and perception of high school wind players. *Bulletin of the Council for Research in Music Education, 134*, 31-42.

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

Young, W. (1971, September 27). *An investigation of the singing abilities of kindergarten and first grade children in east Texas*. [Washington, D.C.] : Distributed by ERIC Clearinghouse, http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED069431

Appendix A - Singing Accuracy Assessment

Equipment: recording stimuli

Professional recording

Equipment: recording responses

Zoom recorder, 1½ feet in front of participant, operated by researcher

Equipment: stimuli playback

Sony stereo, 3 feet in front of participant, operated by researcher

Stimuli parameters

1. Range of D-A (limited range appropriate initial testing for K-5 age range)
2. Note durations and inter-onset intervals both approx. one second
3. Echo on “doo"
4. Recorded with female voice to best match child’s voice

Test Items (in order for Form A)

1. Single pitch: A, D G, F#, A
2. Interval pitches doubled in repetitions: AAF#F#, DDEE, F#F#DD, DDGG, AAF#F#
3. 4-note pattern: AF#AD, DEF#G, F#ADE, AGED, AF#AD
4. Single pitch doubled: A, D G, F#, A
5. Interval pitches doubled: AAF#F#, DDEE, F#F#DD, DDGG, AAF#F#

6. 4-note pattern doubled: AF#AD, DEF#G, F#ADE, AGED, AF#AD

7. Song Singing from F# starting pitch

Assessment Script for stimuli recording*

Use electronic pitch pipe (A = 440) to establish pitch reference.

Begin recording:

“Today I will sing some notes for you to sing back to me. If you have any questions, just stop and ask them. Sing loudly and strongly. It’s okay if you don’t do every one right, just try your best.”

“Here’s a practice example.” (sing DDDD)

1. [sing AAAA for one second then pause three seconds]

2. [sing DDDD for one second then pause three seconds]

3. [sing GGGG for one second then pause three seconds]

4. [sing F#F#F#F# for one second then pause three seconds]

5. [sing AAAA for one second then pause three seconds]

“Now we’ll try a different kind. Here’s a practice example.” (sing DDF#F#)

6. [sing AAF#F# for one second then pause six seconds]

7. [sing DDF#F# for one second then pause six seconds]

8. [sing F#F#DD for one second then pause six seconds]

9. [sing DDGG for one second then pause six seconds]

10. [sing AAF#F# for one second then pause six seconds]

“Now we’ll try a different kind. Here’s a practice example.” (sing DEF#G)

11. [sing AF#AD for one second then pause six seconds]

12. [sing DEF#G for one second then pause six seconds]

13. [sing F#ADE for one second then pause six seconds]

14. [sing AGED for one second then pause six seconds]

15. [sing AF#AD for one second then pause six seconds]

*Prior to beginning the CD, I will read from the Assent Form script (pending future IRB approval, and not provided here), which will remind children their participation is voluntary, that they may stop at any time, and that it’s okay if they unsure about any or all of the test items. No one will be mad at them if they do not do as well as they wish.

Singing Accuracy Assessment - Information Sheet (Do not return)

Students' singing skill will be assessed in an 8-minute researcher-led individual testing environment. Why? To learn more about how to test for singing ability, just like schools test for reading and math ability, etc. Participation will not affect your child's grade or your child's status at the school. You do not have to participate if you don't want to, and you can change your mind at any time, even after signing the consent form.

Students will be asked to:

Complete a brief singing accuracy assessment that asks them to: Sing back some notes and sing some songs from memory.

For questions about this study, you may contact:

Bryan Nichols, University of Washington (Teaching Assistant, Music Education)
phone: XXX-XXX-XXXX, email; _____

This research project is directed by:

Professor Steven Demorest, University of Washington (Professor, Music Education)
phone: XXX-XXX-XXXX, email: _____

Appendix C - Consent Form

UNIVERSITY OF WASHINGTON
CONSENT FORM
"SINGING ACCURACY ASSESSMENT"

Bryan Nichols, Teaching Assistant, Music Education, phone: XXX-XXX-XXXX, email: _____

This project is being directed by Professor Steven Demorest, phone: XXX-XXX-XXXX, email: _____

We are asking you to let your child be in a research study. The purpose of this consent form is to give you and your child the information needed to help you decide whether to be in the study or not. Please read the form carefully. You may ask questions about the purpose of the research, what we will ask your child to do, the possible risks and benefits, your child's rights as a volunteer, and anything else about the research or this form that is not clear. When we have answered all your questions, you can decide if you want your child to participate in the study or not. This process is called "informed consent." We will give you a copy of this form for your records.

PURPOSE OF THE STUDY

The purpose of this study is to learn better ways to test singing skills.

STUDY PROCEDURES

Children will be asked to sing back ("echo") notes and to sing songs for no longer than eight (8) minutes. This one-on-one session will occur during regular classes or during music class, outside of the music room. We will make an audio recording of your child's sung responses. We will be sure that the audio recordings don't contain any information that would identify your child by name.

Your child may refuse to participate in the individual singing assessment at any point.

RISKS, STRESS, OR DISCOMFORT

Learning and assessment experiences can involve some stress or other discomfort; however, your child will not be asked to demonstrate skills differently than would normally occur during the school day, including singing alone or with others. Students will miss twenty minutes of instruction from their teacher.

ALTERNATIVES TO TAKING PART IN THIS STUDY

The alternative is to continue participating normally in class.

BENEFITS OF THE STUDY

Students who participate will receive some extra singing practice, which might benefit their skill development. No individual results will be shared with anyone outside of the research team without the parents' permission.

OTHER INFORMATION

You or your child may refuse to participate and you are free to withdraw from this study at any time without penalty or loss of benefits to which you are otherwise entitled. All of the information you provide will be confidential - no identifying information will be maintained.

Participant's statement

This study has been explained to me. I volunteer my child to take part in this research. I have had a chance to ask questions. If I have questions later about the research, I can ask the researcher listed above. If I have questions about my rights as a research participant, I can call the Human Subjects Division at (206) 543-0098. I will receive a copy of this consent form.

_____
Printed name of child

_____ _____ _____
Printed name of parent/representative Signature Date

Has your child ever been diagnosed with any hearing difficulties? ____ No ____Yes

Has your child had private instrument/voice lessons? ____ No ____Yes

Which and for how long? (e.g., piano - 1 year, 6 months)_____

Your child's age as of today: _____years _____months

Appendix D - Assent Form

ASSENT FORM

Singing Accuracy Assessment

Investigator:

Bryan Nichols, University of Washington (Music Education)

PURPOSE AND BENEFITS

This is a study about music class. We want to learn more about how people learn to sing.

PROCEDURES

If it's okay with you, I'll ask you to sing certain notes and a song that you already know.

OTHER INFORMATION

You don't have to take part in this study if you don't want to. No one will be mad at you. We will give you a copy of this paper to keep.

_____

Student Name

_____ _____

Researcher Signature Date

Copies to:

___ Student

___ Research folder