

Title: The Measurement and Evaluation of Children's Singing Voice Development

Author(s): Joanne Rutkowski

Source: Rutkowski, J. (1990, Spring). The measurement and evaluation of children's singing voice development. *The Quarterly*, 1(1-2), pp. 81-95. (Reprinted with permission in *Visions of Research in Music Education*, 16(1), Summer, 2010). Retrieved from <http://www-usr.rider.edu/~vrme/>

It is with pleasure that we inaugurate the reprint of the entire seven volumes of The Quarterly Journal of Music Teaching and Learning. The journal began in 1990 as The Quarterly. In 1992, with volume 3, the name changed to The Quarterly Journal of Music Teaching and Learning and continued until 1997. The journal contained articles on issues that were timely when they appeared and are now important for their historical relevance. For many authors, it was their first major publication. Visions of Research in Music Education will publish facsimiles of each issue as it originally appeared. Each article will be a separate pdf file. Jason D. Vodicka has accepted my invitation to serve as guest editor for the reprint project and will compose a new editorial to introduce each volume. Chad Keilman is the production manager. I express deepest thanks to Richard Colwell for granting VRME permission to re-publish The Quarterly in online format. He has graciously prepared an introduction to the reprint series.

The Measurement and Evaluation of Children's Singing Voice Development

By Joanne Rutkowski

Pennsylvania State University

Singing as a principal activity in the general music classroom has been observed and confirmed (Atterbury, 1984a; 1984b; Britton, 1961; Brooks & Brown, 1946; Morgan, 1955; Sears, 1965). Though a principal activity and a trainable skill, 18 percent of today's students are "nonsingers" (Bentley, 1968; Davies & Roberts, 1975; Nye & Nye, 1970; Romaine, 1961). Since singing constitutes an important aspect of the elementary general music curriculum, nearly one-fifth of our students are unable to participate profitably in music class, and these students often form negative attitudes toward singing and toward the music class. "The longer the correction [of singing] is delayed, the more negative personality reactions develop and the more difficult it is to correct the problem" (Gordon, 1979, p. 56). Over the past 50 years, numerous studies have investigated topics related to the child's singing voice and its development.

Terminology used to describe the various stages of development of the child voice and/or the types of problem singers, however, has been inconsistent. According to Welch (1979), Anderson indicated that there were no "monotones"; instead, he labeled children deficient in pitch discrimination as "inaccurate singers" while Fieldhouse called those deficient in tonal memory "backward singers". McKenzie (1948) used the term "nonsinger" while Reuter (1956a, 1956b) used "inaccurate singer". Bentley (1968) disliked the term "monotone" since most of these children did produce more than one tone, but he used the term because he believed that it was less derogatory than other terms then in use.

Nye and Nye (1970) indicated two types of singers: "nonsingers", those who do not have use of the singing voice, and "problem singers", individuals who have a very limited range, usually not higher than E3 or F3. Hartzell (1949) established three classifications: children who can establish and maintain tonality, children who can establish tonality but do not maintain it, and children who can do neither. Kirkpatrick (1962) used Hartzell's classifications but relabeled each category using the terms, "singers", "partial singers", and "nonsingers".

Gaiser (1961) used the term "nonsingers" to refer to "children whose singing performance varies from the norm in that

"It seems logical that a child must gain use of the singing voice before intonation problems can be researched and evaluated. Surely an instrumentalist must know the fingering for a particular note and be able to produce a sound on the instrument before intonation problems become a concern."

they habitually sing several tones away from the group, usually below, or vary uncertainly from tone to tone" (p. 4.). She also employed three classifications of "nonsingers": "monotones" who only sing one tone; "near-singers" who sing multiple tones but lack control of tones; and "followers" who imitate a group but cannot sing alone.

Joyner (1969), while working with boys aged 11, employed four categories to describe their vocal achievement: "Normal singers" were those who could sing in a low and high key; "Grade A monotones" were tuneful in the low key but not in the higher key; "Grade B monotones" were those who were erratic at both pitch levels; "Grade C monotones" are always untuneful. Forcucci (1975) also used four categories to describe different types of singers: "independent singers" sing in tune without assistance; "dependent or lazy singers" sing in tune within a group; "uncertain singers" sing out of tune with or without a group; "restricted range singers" are those usually thought to be "monotones" even though they can actually produce more than one pitch.

Gordon (1971, 1979) also described problem singers: "nonsingers" attempt to sing either in the speaking-voice range or above the singing range. "Out-of-tune singers" either have a sense of melodic direction but lack a sense of pitch or lack both aspects.

Young (1971) observed several stages of voice-range development in kindergarten and first-grade children: D³ to F^{#3} or A² to E^{b3}; A² to F^{#3}; A² to C⁴; A² to D⁴ and above (see Figure 1). These four voice stages were not labeled. Young (1971) and Gordon (1971, 1979) both indicated the existence of a voice break or register lift in the child voice from approximately B^{b3} to D^{b4}. Children seem to have the most difficulty producing tones in this range.

Figure 1
Young (1971) voice stages



Purpose and Problem

Since standard terminology has not been in evidence for labeling the various stages through which a child's voice progresses, one purpose of this research was to establish a more consistent means of describing the various stages of child

singing-voice development. This new consistency would enable teachers and researchers to more accurately measure and describe the use a child has of his or her singing voice. Precise assessment is of concern when conducting research in which the singing voice is a factor and when designing music instruction for children.

It seems logical that a child must gain use of the singing voice before intonation problems can be researched and evaluated. Surely an instrumentalist must know the fingering for a particular note and be able to produce a sound on the instrument before intonation problems become a concern. Further, many children do not demonstrate accurate intonation within a phrase until about age 5.5 or 6 years (Davidson et al., 1981). Consequently, it seems that the use a child has of his or her singing voice may be a construct separate from and requisite to the ability to sing in tune. To investigate this hypothesis, the author 1) reviewed rating scales to determine their feasibility for measuring the use of singing voice but not the accuracy of intonation, 2) developed, piloted, and implemented a rating scale for measuring the use of the singing voice, and 3) formulated a consistent vocabulary by which one can describe the singing-voice development of children.

Figure 2

Smith (1961) rating scale

- 4 complete accuracy in reproducing an interval
- 3 complete accuracy but with a tendency to slide into either the first or second interval tone
- 2 child was able to sing only one of the two tones
- 1 complete lack of tone matching ability

Previous Scales

Smith Scale

The rating scale constructed by Smith (1961) seems to be one of the first rating scales used to evaluate the young child's singing voice. Smith specified criteria for each of the four scoring levels he used to evaluate the performance of intervals (see Figure 2). The scale was intended to

measure intonation of intervals rather than use of the singing voice.

Boardman Scale

Boardman (1964) appears to have also been among the first to use an instrument to measure children's use of singing voice. Her instrument was described as a seven-point scale. One point was given if a student failed to respond at all, and seven points were scored for a perfect response. There is considerable subjectivity in this measure; additional criteria were not established for the intervening five scoring levels.

Dittemore Scale

The Dittemore (1969) scale also used seven points (Figure 3). Although criteria were established for each point level and interjudge reliability for two judges was reported as ranging from .61 to 1.00, the scale has serious limitations. It is only useful for songs containing four melodic patterns. Although points could be added for any additional patterns, this alteration prevents comparisons of scores if songs of different length were used. The phrase "comprehends tonality", which is part of the criteria for scoring levels 3 to 7, is ambiguous. Despite the initial reliability reported, it is highly unlikely that a rater could consistently determine from a student's singing if that student comprehended tonality. The two subdomains of interest to this research, use of the singing voice and intonation, are both included in this scale. Although these subdomains are presented in their hierarchical order, use of singing voice before intonation, it is not clear that both can be accurately measured with a single seven-point scale.

Figure 3

Dittemore (1969) rating scale

- 1 no correct response
- 2 student displays only a general sense of direction
- 3 student displays a general sense of tonal direction and comprehends tonality
- 4 . . . and correctly completes any one melodic pattern
- 5 . . . any two melodic patterns
- 6 . . . any three melodic patterns
- 7 . . . all four melodic patterns

DeYarman Scale

The DeYarman (1972) scale (see Figure 4) is a revision of the Dittemore scale. DeYarman used the phrase "sense of tonality", however, instead of "comprehends tonality". This rewording clarified the criterion. Since measurement of melodic patterns was not conducted, the rating scale could be used with any song materials. The interjudge reliability reported for this scale between two judges was $r = .85$ to $.96$. This same scale, when used in the Miller (1975) study, yielded reliability coefficients from $.91$ to $.98$. Both subdomains, use of the singing voice and intonation, are included in one measure. The DeYarman scale is an improvement on previous scales.

Figure 4

DeYarman (1972) rating scale

- 1 no correct response
- 2 no, or very poor sense of tonality, but general sense of direction
- 3 poor sense of tonality, general sense of direction
- 4 fair (moderately good) sense of tonality, good sense of direction
- 5 good sense of tonality, very good sense of direction
- 6 very good sense of tonality
- 7 excellent tonal performance

Young Scale

Young designed an instrument to document children's voice range characteristics (Young, 1971). Part One was a vocal singing-range test in which the children were asked to sing "Unfamiliar melodic segments", "Familiar song phrases in key of child's choice", and "Familiar song phrases in keys of examiner's choice" (p. 8). Part Two, a singing-ability and tonality preference test, was comprised of matched items in major and minor tonalities. This test was descriptive, rather than evaluative, in that voice graphs were constructed for each group of children based on the children's maximum voice range and accurate vocal range rather than scores being assigned for each child.

Roberts and Davies Scale

Measurement instruments constructed since 1975 seem to have abandoned the

seven-point scale in favor of a five-point scale. Roberts and Davies (1975), British researchers, used a scale similar in format to the Boardman (1964) scale (see Figure 5). Criteria were not established for the scoring levels and consequently considerable rater subjectivity was probable.

Figure 5

Roberts & Davies (1975) rating scale

- 0 tune completely unrecognizable
- 1 part of tune recognizable
- 2
- 3
- 4 correct performance

Hale (Runfola) Scale

The Hale (Runfola) (1977) scale used five specific hierarchical scoring levels to measure intonation (Figure 6). While a hierarchy had been empirically documented regarding difficulty levels for perception of patterns through listening (Gordon, 1976) at the time this scale was developed, no empirical research existed regarding difficulty levels of patterns for singing. Hale's assumption that matching difficulty levels to Gordon's hierarchy provided content validity was an untested assumption. Further, the scale was limited to song materials comprised only of tonic and dominant patterns. The interjudge reliabilities reported for two judges were slightly lower than those reported for previous studies of similar design: $r = .53$ to $.83$.

Figure 6

Hale [Runfola] (1977) rating scale

- 1 no sense of tonality, no accurate resting tone
- 2 accurate resting tone at beginning or end of song
- 3 tonic tonal patterns accurate
- 4 dominant-seventh tonal patterns accurate
- 5 all patterns accurate

Ramsey Scale

Ramsey constructed the Preschool Singing Ability Level Test (PSALT) to measure: "(1) preschool children's ability to reproduce pitches in a specified vocal range and (2) preschool children's ability to reproduce a song" (Ramsey, 1982,

p. 24). Part I centered on a story in which each animal from the story had a "sound" represented by a specific pitch (B \flat ² to C⁴). Each pitch was first played on a tone bell, then sung by the investigator, and finally echoed by the children. Each child had three opportunities to sing the correct "sound".

For Part II, each child was requested to sing a song of his or her choice. No beginning pitches were given for this section. All responses were tape recorded, notated, and rated according to the rating scale in Figure 7. The interjudge reliability reported was .99. Again, this scale attempted to simultaneously measure use of singing voice, correctness of melodic contour, maintenance of tonal center, and intonation. According to Gordon (1971), a rating scale that is most valid purports to measure only one aspect of performance.

Figure 7

Ramsey (1982) rating scale

- 0 The child made no response
- 1 The child used his speaking voice rather than singing the response
- 2 The child used his singing voice but sang incorrect general melodic contour, incorrect intervals, and exhibited no ability to establish or maintain a tonal center
- 3 The child maintained the general contour of the song but sang incorrect intervals and changed tonal center three or more times from that established at the beginning of the song
- 4 The child maintained the general contour of the song but sang incorrect intervals and changed tonal center two times from that established at the beginning of the song
- 5 The child maintained the general contour of the song but sang incorrect intervals and changed tonal center one time from that established at the beginning of the song
- 6 The child maintained the general contour and the beginning tonal center but sang incorrect intervals
- 7 The child sang accurately in regard to general melodic contour and correct intervals and maintained the beginning tonal center throughout the response

Rutkowski Scale

In order to investigate the use of magnitude estimation as an alternative

instrument for measuring singing voice achievement, Rutkowski (1983) developed a scale similar to the Hale (Runfola) scale (1977) but one not limited by specific song material (Figure 8). Scoring level 4, assigned to children who sang two or more tonal patterns accurately, was used to score performances that in fact may have had considerable variance if the song used had many tonal patterns. This scale, like several previous scales, measured children's use of singing voice as well as accuracy of intonation. Interjudge reliability reported was $r = .81$.

Figure 8

Rutkowski (1983) rating scale

- 1 not in singing voice: voice shows speech inflection but not melodic contour
- 2 voice shows pitch change and inflection (melodic contour) but no sense of resting tone or pattern accuracy
- 3 one tonal pattern sung accurately and/or a sense of resting tone exhibited
- 4 2 or more tonal patterns sung accurately
- 5 all tonal patterns sung accurately

As part of the same study, Rutkowski (1983) investigated the use of magnitude estimation as a means of measuring children's use of singing voice. Magnitude estimation does not provide scoring levels and criteria, but rather asks the rater to draw a line indicating performance achievement. Although interrater reliability for magnitude estimation was acceptable ($r = .61$), magnitude estimation did not correlate highly with the Rutkowski (1983) rating scale ($r = .15$). Several conclusions were drawn: (1) that the validity of magnitude estimation as an instrument to measure children's use of singing voice was suspect since the rating scale had a fairly high interjudge reliability of .81; (2) that the two scales may have been measuring different aspects of singing voice with magnitude estimation providing a Gestalt evaluation of the performances and the rating scale providing an Atomistic evaluation of the performance; and (3) that raters may have been more comfortable using the traditional rating scale format than with magnitude estimation and therefore additional practice

with magnitude estimation may have yielded different results.

Feierabend Scale

A five-point rating scale, constructed by Feierabend in 1984, used two- and three-tone major tonic patterns sung by a mezzo soprano as a model after which children were asked to reproduce each pattern (Figure 9). Interjudge reliabilities ranged from .72 to .86. The wording of each scoring level seemed sufficiently explicit for easy use. The instrument, however, focused upon intonation and correctness of melodic contour and not only on the use of the singing voice.

Figure 9

Feierabend (1984) rating scale

- 5 The tonal pattern is accurately reproduced with good intonation.
- 4 The tonal pattern is correctly reproduced but with some uncertainty.
- 3 Melodic direction is evident but some tones are incorrectly reproduced.
- 2 Melodic direction is evident but no tones are correctly reproduced.
- 1 Reproduction of the tonal pattern is not recognizable.

Summary

While several rating scales intended to measure various types of singing-voice achievement exist, none exclusively measures use of the singing voice. Although the quality of these scales has recently improved, most scales are designed to measure both use of singing voice and accuracy of intonation.

Singing Voice Development Measure

Use of the singing voice, as a domain separate from accuracy of intonation, does not appear to have been empirically investigated. Consultation with several elementary vocal music specialists as well as a compilation of results from previous studies were the means by which characteristic singing-voice behaviors were identified. These behavioral stages were assumed to be sequential and were initially classified as follows:

1. Children who use only speaking-voice inflection but do not sustain tones
2. Children who exhibit use of melodic contour and sustained tones, but use speaking range or a very high range
3. Children who use a very limited singing range, usually D³ to F^{#3}. This stage has been noted by several other researchers (Harkey, 1979; Joyner, 1971; Young, 1971)
4. Children who use initial singing range, usually D³ to A³; and
5. Children who are able to sing over the register lift, F^{b3} and above (Gordon, 1971; Smith, 1963; Young, 1971).

Pilot Study

Procedures. A rating scale, *The Singing Voice Development Measure* (SVDM), was designed based on the five stages. A pilot study was conducted to document the presence of these behavioral characteristics and to assess the feasibility, reliability, and validity of the instrument for measuring children's use of singing voice (Rutkowski, 1984). The 35 children in the pilot study were kindergarten students enrolled in a parochial school in the Buffalo, NY, metropolitan area. There was no reason to believe that the early music experiences of these children were any different from that of public school children residing in similar middle-class neighborhoods. These students had received music instruction from music specialists during the first five months of the school year, scheduled in two, 20-minute class periods each week. Since kindergarten children enrolled in public school usually also receive music instruction from a music specialist, the assumption was that these parochial school children were representative.

The song selected for the children's performances was a familiar song in harmonic minor mode (Figure 10). While pentatonic songs have often been used for the musical training of young children, Jarjisian (1981) and Michel (1973) concluded that children sing in major and minor modes just as readily as in pentatonic. In addition, several researchers have found minor songs easier for young children to sing than major songs (DeYarman, 1972; Dittmore, 1969). Furthermore, it seems

logical that assessment of young children's singing voices should involve the singing of material familiar to them. If children do not know the songs, test scores may reflect insecurity and unfamiliarity with the material. Any resulting measurement error would result in inappropriate conclusions about the children's use of their singing voices.

Figure 10
SVDM criterion song

BAKERMAN



Come along, come along, what shall we play? I'll be a bakerman, just for today.

The children's voices were tape recorded while they sang the familiar song as a group. The recording was immediately played back for the students. In other studies in which children's voices were tape recorded, test anxiety was diminished (Runfola, 1981; Rutkowski, 1983). The students were next individually sent to a familiar room for testing. Each student was asked to sing the same song that was recorded by the group. In order to procure a more natural portrayal of each child's use of the singing voice, tonality, starting pitch, and tempo were not established for the taped performance. If children were extremely familiar with a song, those with control of their singing voice will tend to reproduce the song at the pitch level that has been used in class (Zimmerman, 1981). Each child's voice was recorded again several days later following the same procedure. The children reported for the second testing in a different random order than for the first test. The testing procedure resulted in 70 recorded voices. Four raters, all elementary music specialists who are researchers in the discipline and practiced raters, were chosen to evaluate the recorded voices.

Three tapes were prepared from the two original tapes for the rating procedure. Tape A consisted of the first performance of each child; tape B1 the second

performance of each child; and tape B2 the same performances as tape B1, but with a different order of performance. Raters were assigned to rate two of the three tapes. Five practice examples were included at the beginning of each tape recording. Use of practice examples is recommended to familiarize raters with the rating scale they are to use (Brown, 1976; Fiske, 1979). Each rater was instructed to listen to 35 performances in one rating session and to assign a number from one to five for each performance corresponding to the singing voice scoring levels. The next 35 performances were similarly evaluated several days to one week later.

To assist the rater, a ready-sing prior to each performance was tested. Thirteen performances so arranged were included at the end of each tape for the raters to evaluate. Evaluator response was very positive to this addition. Specific questions were asked of the raters about the feasibility of the measurement instrument and their feelings toward the use of a "ready-sing" procedure.

Response frequency. The raters' use of the five singing-voice classification levels was of interest. If raters had not used all five scoring levels, one could argue that the criterion for each level was not presented clearly and revisions would need to be made to insure valid ratings. The response frequency for each rating is presented in Table 1. The raters did use all five scoring levels to evaluate the children's use of their singing voices.

Table 1
Raters' response frequency
of scoring levels

Rater	Scoring Levels				
	1	2	3	4	5
1	11	11	24	27	10
2	9	19	19	19	17
3	10	20	17	23	13
4	17	11	18	16	21
Total	47	61	78	85	61

Reliability. A correlation derived from a two-way mixed-model analysis of variance is the most appropriate estimate of reliability for a measurement instrument where

"n individuals drawn at random are being rated by k fixed raters by a fixed rating scale" (Bintig, 1980, p. 623).

All reliability coefficients for this rating scale were satisfactory. Interrater reliabilities ranged from .836 to .963 and performer consistency reliability was .918. An interrater reliability was also computed across all four raters for the 13 performances in which a "ready-sing" was given. This coefficient was .904. Upon comparison with the interrater reliability from the other performances, the "ready-sing" did not yield a more reliable measure (Table 2).

Table 2
Inter-rater reliability with and
without "ready-sing"

Setting	Reliability
Without "ready-sing"	.963
With "ready-sing"	.904

Since this scale was developed to evaluate the level of achievement, content validity was a concern. Initial steps taken to assure content validity included consultation with specialists in the field and comparisons with other scales. The raters were also queried. They indicated agreement with the content of the scale, although most noted an inclination to listen for intonation rather than just use of the voice.

Conclusions and revisions. Based upon frequency of responses, reliability, and validity for the instrument in the pilot study, the SVDM appeared to be an appropriate instrument for assessing the use a child has of his or her singing voice. A few modifications were made, however: The scoring level descriptions were revised for easier interpretation and evaluation (Figure 11). A "ready-sing" was included before all voices. Even though its use generated a slightly lower reliability, the raters overwhelmingly recommended its inclusion, as they felt that it greatly assisted with their evaluation of the voices.

Figure 11
Revised rating scale

- 1 "Pre-singer" does not sing but chants the song text
- 2 "Speaking range singer" sustains tones and exhibits some sensitivity to pitch but remains in the speaking voice range (usually A² to C³).
- 3 "Uncertain singer" wavers between speaking and singing voice, uses a limited range when in singing voice (usually up to F³).
- 4 "Initial range singer" exhibits use of initial singing range (usually D³ to A³).
- 5 "Singer" exhibits use of extended range (sings beyond the register lift: B^b³ and above).

Since the raters indicated difficulty in evaluating the children's voices based only upon performance of a song, the instrument was expanded to contain tonal patterns representative of each scoring level (Figure 12). Sims et. al (1982) found that shorter patterns were much easier for children to sing. While Jersild and Bienstock (1934) found no difference in children's ability to sing ascending and descending patterns and Sinor's (1985) research only partially supported this notion, Fox (1983) and McKernon (1979) documented that up to 82 percent of the patterns sung by "infants" are descending. Other researchers have noted the same phenomenon with young children as well (Bentley, 1973; Jersild & Bienstock, 1931; Pond, 1980; Sallstrom & Sallstrom, 1973; Vance & Grandprey, 1929). In addition, patterns that encompass pitches below and above the register lift should jump above the lift and then descend through it (Gordon, 1979). The patterns used for the revised SVDM were each comprised of three descending tones, with the exception of the fifth pattern, which ascended first in a jump over the register lift before descending. Therefore, the revised rating measure was comprised of two subtests: (1) performance of a song and (2) performance of five tonal patterns.

Main Study

The SVDM was implemented as an instrument to measure use of singing voice in a study conducted by Rutkowski in 1986. The instrument was used as a

pretest and a posttest measure. The study included 162 kindergarten children enrolled in three elementary schools in the Williamsport, PA Area School District (six classes) and in a parochial school located in the same geographic area (one class). The children enrolled in the Williamsport Area schools received music instruction from a music specialist for one 30-minute class per week. The children enrolled in the parochial school did not receive music instruction from a specialist; the classroom teacher periodically engaged the children in music activities.

Pretest preparation. An orientation period preceded the pretesting in order to allow the children an opportunity to become familiar with the songs and patterns of SVDM as well as with the investigator. One month prior to administration of the pretest, the investigator visited each participating kindergarten class, during which time she observed the

Figure 12
SVDM: tonal patterns



Text: 1: See the birds
2: in the tree
3: build a nest
4: with some twigs
5: and some leaves.

children during a music class, led some of the singing activities, and taught the criterion song, "Bakerman" (Figure 10), as well as the five tonal patterns comprising SVDM (Figure 12). The music teachers reviewed "Bakerman" with the children during subsequent music classes. Because the pattern section of SVDM is an echo activity and the children would be imitating the investigator's voice for the actual test, the teachers did not practice this activity with the children.

Singing-voice data were collected using procedures similar to those utilized in the pilot study (Rutkowski, 1984). The children's voices were tape recorded twice, several days to one week apart. Immediately prior to individual testing, the

investigator rehearsed the song and patterns with the children as a group. A "ready-sing" was given in the same mode and meter as the song (minor, duple) on the pitches of the tonic chord of the song (D³ F³ A³). These pitches were first played on tone bells and then sung by the investigator with the words "ready-sing". The patterns were also first played on tone bells, sung by the investigator, and then echoed by the children. A tape recording was made of the group performance and played for the children. This procedure was followed for the first administration of the test only. As a technique to motivate enthusiasm for individual performance, the investigator commented positively on the children's singing, but noted that individual voices could not be heard within the group. Therefore it was explained that each child was going to be given the opportunity to hold the microphone and sing without the rest of the class. At no time was the situation referred to as a test. Without exception, the children appeared enthusiastic and welcomed the opportunity to sing individually.

The children were taken to a familiar room in the school where their individual performances could be recorded in privacy. The children reported to the testing room in groups of two; one child waited outside the room while the other sang. Then the first child went back to his regular classroom and summoned the next child while the second child sang, and so on. No importance was placed on the order the children were dismissed, but the teachers were requested to use a different order at the time the second tape recording was made. To offset any unintended order effects, the children were randomly assigned to one of two "groups". The children in Group 1 sang the patterns first for the first tape recording and the song first for the second tape recording. Children in Group 2 performed the materials in the reverse order.

Although time-consuming, the testing procedures were smoothly conducted. The children responded positively and seemed comfortable both with the individual singing and with using the tape recording.

Posttest procedures. Several weeks prior to administration of the posttest, the music teachers were asked to review the song "Bakerman" with the children. On the actual testing day, the investigator rehearsed both the song and the selected patterns with the children. A group tape recording was not made prior to the posttest: The children indicated enthusiasm for singing alone as soon as the investigator began to review the materials. No motivation activity was necessary. The pretest procedures were followed for the posttest. This time, children in Group 1 sang the song first for the initial tape recording and the patterns first for the second tape recording. The order was reversed for children in Group 2.

Scoring procedures. Raters 2 and 4, who participated in the pilot study (Rutkowski, 1984), were chosen as raters for the main study. These raters had the highest inter-rater reliabilities and the most previous experience rating children's singing voices. Both raters are researchers in elementary general and early childhood music and have extensive experience with and knowledge of children's musical and vocal capabilities. For rating purposes, scoring tapes were constructed from the original recordings of the children's performances.

In order to eliminate rater bias for the second section of the test (song or pattern) arising from the children's performance on the first section of the test (song or pattern), song performances were separated from pattern performances. As a result, one recording contained only song performances, and the other only pattern performances. Although the classes remained intact on the evaluation tapes, the order of the classes was randomly assigned.

In addition, one class was randomly selected to be included twice on the song tapes, and another class was randomly selected to be included twice on the song tapes and another class was randomly selected to be included twice on the pattern tapes. This arrangement allowed for investigation of internal consistency of each rater. Since each child performed twice for the pretest and twice for the posttest, each set of evaluation tapes

contained 15 groups of performances. While evaluation of pretest and posttest performances randomly mixed would have helped insure against biased ratings, the large number of performances to be evaluated made this impractical. The quantity of performances to be rated also rendered one rating session unrealistic. Considering rater fatigue, raters were encouraged to rate as many performances as possible in a sitting without jeopardizing the quality of their ratings.

For the evaluation procedures, each rater was presented with a packet of materials including the song and pattern performance by the children, the rating scale to be used, and rating forms on which to record scores (singing-voice categories) for each performance. Scores were then derived by adding the rater's scores for both pretest performances and both posttest performances. Therefore, subjects were each assigned three pretest and three posttest scores: a pattern score (4-20 points possible), a song score (4-20 points possible), and a composite score (8-40 points possible.).

Means, standard deviations, reliability coefficients, and standard errors of measurement were computed across all subjects for SVDM on the Pattern, Song, and Composite tests. Reliability coefficients were computed for both the pretest and posttest through the two-way mixed model as in the pilot study. Since two raters were employed for the main study rather than four as in the pilot study, only reliabilities for the consistency of children's performances over time, within the pretest and posttest and intra-rater stability were computed. Although reliabilities were reported for all tests within SVDM (Patterns, Song, and Composite), only coefficients for the Song test were compared with corresponding coefficients from the pilot study because the pilot study measure contained only that test.

Results. Means, standard deviations, standard errors of measurement, and reliability coefficients for the SVDM pretests are presented in Table 3. Means, standard deviations, and standard errors of measurement for the Pattern and Song subtests were similar, and the children's performances were consistent. The raters

were similarly consistent within each subtest: $r = .74$ for the Patterns and $r = .93$ to $.94$ for the Song. As can be seen, however, these raters exhibited a higher internal consistency when rating the song performances than when rating the pattern performances.

Table 3
SVDM: Pretest

Test	\bar{X}	SD	SEM	* r^1	r^2	r^3
Patterns	11.2	4.30	0.38	.96	.74	.74
Song	11.3	4.15	0.33	.95	.93	.94
Composite	22.5	8.03	0.63	.97	—	—

* r^1 = consistency of children's performances

r^2 = intrarater: rater 1; r^3 = intrarater: rater 2

Means, standard deviations, standard errors of measurement, and reliability coefficients for the SVDM posttests are presented in Table 4. Again, the means, standard deviations, and standard errors of measurement for the Pattern and Song subtests are similar. The raters exhibited better internal consistency for the posttest than the pretest. Both raters were more reliable when rating the pattern performances than the song performances.

Table 4
SVDM: Posttest

Test	\bar{X}	SD	SEM	* r^1	r^2	r^3
Patterns	12.2	4.27	0.34	.94	.97	.91
Song	12.0	4.31	0.34	.95	.92	.88
Composite	24.2	8.20	0.65	.96	—	—

Coefficients for the Song test were compared with corresponding coefficients for the pilot study: The pilot test did not contain patterns (Table 5). The children in the main study were more consistent than were the children who participated in the pilot study. Since the coefficients are similar, the higher reliabilities in the main study may have been a function of either the larger sample size used for the main study ($n=162$) than for the pilot study ($n=70$) or the reduced number of raters used for the main study. Conversely, the intra-rater reliability coefficients for the pilot study were slightly higher than for the main study. Rater fatigue, due to the larger number of performances to be rated for the main study, may be one reason for these lower reliabilities.

Table 5
SVDM: Comparison with
pilot study group

	Group		
	Pilot	Implementation	
Reliability		Pretest	Posttest
Children's consistency	.92	.95	.95
Intra-rater	.97	.93-.94	.88-.92

Conclusions. The high reliability coefficients for SVDM supported the consistency and stability of the measure (Isaac & Michael, 1982). The lower coefficients for intra-rater reliability on the Patterns subtest for the pretest and the higher coefficients for this same subtest for the posttest suggest that the raters were more comfortable with rating the pattern performances after experience with the rating process. The standard deviations and standard errors of measurement were similar for the pretest and posttest, a further indication of the raters' stability and the children's consistency of performance. Furthermore, a comparison of the Song subtest reliability from the main study with those of the pilot study, comprised of the same types of performances, revealed that the measure functioned similarly in both situations. The SVDM should be an appropriate measure of use of singing voice.

Conclusions and Recommendations

The problems of this study were: 1) to review rating scales to determine their feasibility for measuring the use of singing voice but not the accuracy of intonation, 2) to develop, pilot, and implement a rating scale for measuring use of the singing voice, and 3) to formulate a consistent vocabulary by which one can describe the singing voice development of children. Results of each problem, and conclusions and recommendations based on those results, include the following.

Problem 1

Several rating scales have been used to assess children's singing-voice achievement.

However, none exclusively measured use of singing voice: Melodic contour and intonation were of primary concern. Since these existing scales were not appropriate for measuring use of singing voice, a rating scale to measure this domain needed to be designed, piloted, and implemented.

Problem 2


The Singing Voice Development Measure (SVDM) was designed, piloted, and implemented. It was shown to be a valid instrument to measure children's use of singing voice (Rutkowski, 1984, 1986).

Even though SVDM is a valid measure for children's use of singing voice, several further revisions are recommended. The testing procedure, while successful, was very time-consuming. Since the Patterns and Song subtests were highly correlated on both the pretest and posttest (Rutkowski, 1986), it seems that one subtest would be sufficient for measuring children's use of singing voice. The children's mean scores on both subtests were similar. Their gain scores, however, were slightly better for the patterns. Although reliability coefficients were also similar for both subtests, the raters indicated that once they had become familiar with rating the pattern performances, they found these performances easier to rate than the song performances. Furthermore, when using a song for evaluation, a lengthy orientation period is required in order to familiarize the children with the criterion song. While an orientation period is also necessary when using patterns for evaluation, this period need not be as long since the pattern performances are an echo activity. Although singing a song, rather than patterns, is generally considered singing, it seems that performing a song involves aspects other than use of singing voice. These include memorization of text, rhythm patterns, and tonal patterns. A child may not be employing singing voice simply because she cannot remember some of these other components. Therefore, measuring use of singing through performance of a song may not yield a valid score. Since the singing of patterns involves echoing, rather than memorization, the role of

these other components would be diminished.

Upon consultation with the raters, the belief was that the first and second patterns should be eliminated. Since the first pattern was chanted, rather than sung, it may have encouraged children to use their speaking voice rather than singing voice on the subsequent patterns. Similarly, the second pattern was sung in a speaking voice range and may have encouraged children to use their speaking voice range on the subsequent patterns. Furthermore, while the text for the pattern performances seems interesting and easily remembered by the children, the vowel and consonant combinations were not easily sung. The use of a neutral syllable, rather than words, would alleviate this problem. This investigator has observed that children often experience difficulty singing on a neutral syllable. Research studies investigating the use of a neutral syllable for singing with children have yielded contradictory results (Goetze, 1985; Levinowitz, 1989; Smale, 1987). Until more information is available regarding this matter, it is recommended that new, more singable, texts be identified and used for the pattern performances. A revised set of patterns, based on these recommendations has been identified and comprises SVDM as used in a study by Rutkowski (1989) (Figure 13).

Figure 13
SVDM patterns: recent revisions



#1 #2 #3 #4

Text: 1: See the bird
 2: in the tree
 3: See it fly
 4: over me.

Finally, it is recommended that the SVDM be used to evaluate use of singing voice for research purposes as well as in a classroom setting to assist the music teacher in providing more appropriate instruction.

Problem 3

Since content validity was established for SVDM for measuring children's use of singing voice (Rutkowski, 1984, 1986), it may be inferred that the categories established for use of singing voice do exist. They are defined as follows.

1. Pre-singers: Children who do not sustain tones; their singing response resembles chanting in the speaking voice range. Since "nonsinger" is a derogatory term often used to describe these children, and might be misinterpreted as referring to those who do not participate in singing activities, "pre-singer" was chosen as a more accurate term to describe these children.
2. Speaking-range singers: Children who sustain tones and exhibit some sensitivity to pitch but remain within the speaking-voice range, usually A² to C³.
3. Uncertain singers: Children who sustain tones but often waver between a speaking-voice range and a singing-voice range. When in singing voice, they utilize a range up to approximately F^{#3} and seem to have difficulty in lifting the voice above this pitch. This stage has been noted by several other researchers (Harkey, 1979; Joyner, 1971; Young, 1971).
4. Initial-range singers: Children who have use of the singing-voice range up to the register lift, usually to A³. At this stage, the children rarely drop back into speaking-voice range.
5. Singers: Children who are able to sing over the register lift, B^{b3} and above, and have full use of their singing voices.

It should be re-emphasized that these categories are not concerned with accuracy of intonation or melodic contour. Some children in stages 2 to 5 will sing in tune within the limits of their voice range. Many children in these stages, however, will be out-of-tune singers.

In order to further understand kindergarten children's use of singing voice, the mean category (scoring level) for the kindergarten children who participated in the Rutkowski (1986) study on each subtest of SVDM are presented in Table 6. As can be seen, the average kindergarten child in this study was an uncertain singer. This result indicates that

these children wavered between a speaking and singing voice and, when using a singing voice, they had a very limited range of D³ to F³. The uncertainty of these children regarding their use of singing voice may be a result of several phenomena. It is possible that they do not understand or cannot hear the difference between a speaking and singing voice. It may be that they understand the difference, but cannot demonstrate the difference. If this is the case, perhaps a physiological problem exists, or they do not have the physical readiness. The problem may also be a function of the extended practice they receive with using a speaking voice as opposed to a singing voice. Speaking skills have been developing for five years, whereas use of singing voice may not have been similarly reinforced or encouraged during this time. As several studies recommend, perhaps very restrictive song ranges should be used for this age group (Buckton, 1977; Cleall, 1970; Drexler, 1938; Hattwick, 1933; Kirkpatrick, 1962; Rutkowski, 1986; Smith, 1963; Smith, 1974, Vaughan, 1981; Wilson, 1971). Studies investigating these possible scenarios are recommended.

Table 6
Average voice categories

Test	Pretest	Posttest
Patterns	2.80	3.05
Song	2.82	3.00
Composite	2.81	3.03

In conclusion, it appears that the hypothesis regarding use of singing voice as a separate but requisite behavior to the ability to sing with accurate intonation is well-founded. Consequently, it seems appropriate for this domain to be considered when singing-voice development or achievement is being evaluated. Since the SVDM appears to be a valid measurement instrument to assess use of singing voice, its use is recommended. Further study to investigate the validity of this hypothesis and the SVDM is encouraged. □

References

- Atterbury, B. W. (1984a, April). Are you really teaching children how to sing? *Music Educators Journal*, pp. 43-45.

- Atterbury, B. W. (1984b). children's singing voices: A review of selected research. *Bulletin of the Council for Research in Music Education*, 80, 51-63.
- Bentley, A. (1968). Monotones: a comparison with normal singers in terms of incidence and musical abilities. *Music Education Research Papers No. 1*. London: Novello & Co., Ltd.
- Bentley, A. (1973). Technical problems in group measurement of pitch discrimination and an apparent subjective preference for downward tonal movement. *Psychology of Music*, 1(2), 31-38.
- Bintig, A. (1980). The efficiency of various estimations of reliability of rating scales. *Educational and Psychological Measurement*, 40, 619-643.
- Boardman, E. (1964). An investigation of the effect of preschool training on the development of vocal accuracy of young children. *Dissertation Abstracts International*, 25, 1245A. (University Microfilms No. 64-8354).
- Britton, A. (1961). *Music education: An American specialty*. In P. H. Lang (Ed.), *One hundred years of music in America* (pp. 211-229). New York: G. Schirmer.
- Brooks, B. M., & Brown, H. A. (1946). *Music education in the elementary school*. New York: American Book Company.
- Brown, F. G. (1976). *Principles of educational and psychological testing* (2nd ed.). NY: Holt, Rinehart and Winston.
- Buckton, R. M. (1977). A comparison of the effects of vocal and instrumental instruction on the development of melodic and vocal abilities in young children. *Psychology of Music*, 5(1), 36-47.
- Cleall, C. (1970). *Voice production in choral technique*. London: Novello.
- Davidson, L., McKernon P., & Garner, H. (1981). The acquisition of song: A developmental approach in *Documentary Report of the Ann Arbor Symposium* (pp. 301-315). Reston, VA: Music Educators National Conference.
- Davies, A. D. M., & Roberts, E. (1975). Poor pitch singing: A survey of its incidence in school children. *Psychology of Music*, 3(2), 24-36.
- DeYarman, R. M. (1972). An experimental analysis of the development of rhythmic and tonal capabilities of kindergarten and first grade children. In E. E. Gordon (Ed.), *Experimental research in the psychology of music*: 8 (pp. 1-44). Iowa City, IA: University of Iowa Press.
- Dittemore, E. E. (1969). An investigation of some musical capabilities of elementary school students. *Dissertation Abstracts International*, 29, 4516A. (University Microfilms No. DA8410243).

- Drexler, E. N. (1938). A study of the development of the ability to carry a melody at the preschool level. *Child Development*, 9(3), 319-331.
- Feierabend, J. M. (1984). The effects of specific tonal pattern training on singing and aural discrimination abilities of first grade children. *Dissertation Abstracts International*, 45, 110A. (University Microfilms No. DA8410243)
- Fiske, H. E. (1979). Music performance evaluation ability: Toward a model of specificity. In E. Asmus (Ed.), *The psychology and acoustics of music: A collection of papers* (pp. 147-169). Lawrence, KS: University of Kansas Press.
- Forcucci, S. L. (1975, October). Help for inaccurate singers. *Music Educators Journal*, pp. 57-61.
- Fox, D. B. (1983). The pitch range and contour of infant vocalizations. *Dissertation Abstracts International*, 43, 2588A. (University Microfilms No. 8300247)
- Gaiser, P. E. (1961). *A study of tone-matching techniques as remedial instruction for non-singers*. Unpublished doctoral dissertation, University of Oregon, Eugene, OR.
- Goetze, M. (1985). Factors affecting currency in children's singing. *Dissertation Abstracts International*, 46, 2955A. (University Microfilms, No. DA8528488)
- Gordon, E. E. (1971). *The psychology of music teaching*. Englewood Cliffs, NJ: Prentice Hall, Inc.
- Gordon, E. E. (1976). *Tonal and rhythm patterns: An objective analysis*. Albany, NY: State University of New York Press.
- Gordon, E. E. (1979). *Primary measures of music audition*. Chicago: G. I. A. Publications, Inc.
- Gordon, E. E. (1984). *Learning sequences in music: Skill, content, and patterns*. Chicago: G. I. A. Publications, Inc.
- Hale, M. [Runfola] (1977). An experimental study of the comparative effectiveness of harmonic and melodic accompaniment in singing as it relates to the development of a sense of tonality. *Bulletin of the Council for Research in Music Education*, 53, 23-30.
- Harkey, B. L. (1979). The identification of and the training of the vocal range of three-year-old preschool children. *Dissertation Abstracts International*, 39, 6618A. (University Microfilms No. 7911572.)
- Hartzell, R. E. (1949). *An exploratory study of tonality apprehension and tonal memory in young children*. Unpublished doctoral dissertation, University of Cincinnati, Cincinnati, OH.
- Hattwick, M. S. (1933). The role of pitch level and pitch range in the singing of school children. *Child Development*, 4(4), 281-291.
- Issac, S., & Michael, W. B. (1982). *Handbook in research and evaluation*. San Diego, CA: Edits Publishers.
- Jarjisian, C. S. (1981). The effects of pentatonic and/or diatonic pitch pattern instruction on the rote-singing achievement of young children. *Dissertation Abstracts International*, 42, 2015A. (University Microfilms No. 8124581)
- Jersild, A. T., & Bienstock, S. F. (1931). The influence of training on the vocal ability of three-year-old children. *Child Development*, 2(4), 272-291.
- Jersild, A., & Bienstock, S. (1934). A study of the development of children's ability to sing. *Journal of Education Psychology*, 25, 481-503.
- Joyner, D. R. (1969). The monotone problem. *Journal of Research in Music Education*, 17(1), 115-124.
- Joyner, D. R., (1971). *Pitch discrimination and tonal memory and their association with singing and the larynx*. Unpublished masters thesis, University of Reading, England.
- Kirkpatrick, W. C. (1962). Relationships between the singing ability of pre-kindergarten children and their home musical environment. *Dissertation Abstracts*, 23, 886. (University Microfilms No. 62-3736)
- Levinowitz, L. M. (1989). An investigation of preschool children's comparative capability to sing songs with and without words. *Bulletin of the Council for Research in Music Education*, 100, pp. 14-22.
- McKenzie, J. S. (1948, March-April). New methods for nonsingers. *Educational Music Magazine*, pp. 20-21, 52-54.
- McKernon, P. E. (1979). The development of first songs in young children. In H. Gardner & D. Wolf (Eds). *Early symbolization: New directions for child development* (pp. 43-58). San Francisco: Jossey-Bass.
- Michel, P. (1973). The optimum development of musical abilities in the first years of life. *Psychology of Music*, 1, 14-20.
- Miller, P. H. (1975). An experimental analysis of the development of tonal capabilities of first grade children. In E. E. Gordon (Ed.), *Experimental research in the psychology of music*: 10 (pp. 77-97). Iowa City, IA: University of Iowa Press.
- Morgan, H. N. (Ed.). (1944). *Music in American education*. Washington, DC: Music Educators National Conference.
- Nye, R. E., & Nye, V. T. (1970). *Music in the elementary school* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Pond, D. (1980, March). The young child's playful world of sound. *Music Educators Journal*, pp. 39-41.
- Ramsey, J. H. (1982). An investigation of the effects of age, singing ability, and experience

- with pitched instruments on the melodic perception of preschool children. *Dissertation Abstracts International*, 42, 3053A. (University Microfilms No. 8128451)
- Reuter, G. S. (1956a, January-February). Remedial treatment of inaccurate singers. *Educational Music Magazine*, pp. 19-20, 56-61.
- Reuter, G. S. (1956b, March-April). Remedial treatment of inaccurate singers: Part II. *Educational Music Magazine*, pp. 41-45.
- Roberts, E., & Davies, A. D. M. (1975). The response of "monotones" to a programme of remedial training. *Journal of Research in Music Education*, 23(4), 227-239.
- Romaine, W. B. (1961). *Developing singers from non-singers*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York, NY.
- Runfola, M. (1981, April). *An investigation of a technique for identifying early childhood uncertain singers*. Paper presented at the national meeting of the Music Educators National Conference, Minneapolis, MN.
- Rutkowski, J. (1983, February). An investigation of the reliability of magnitude estimation as an instrument for evaluating kindergarten children's use of singing voice. Paper presented at the Eastern regional meeting of the Music Educators National Conference, Boston, MA.
- Rutkowski, J. (1984). Development of a rating scale to assess individual children's use of the vocal instrument. Unpublished manuscript, State University of New York at Buffalo, Buffalo, NY.
- Rutkowski, J. (1986). The effect of restricted song range on kindergarten children's use of singing voice and developmental music aptitude. *Dissertation Abstracts International*, 47, 2072A. (University Microfilms No. 8619357)
- Rutkowski, J. (1989). *The comparative effectiveness of individual and group singing activities on kindergarten children's use of singing voice and developmental music aptitude*. Unpublished manuscript, funded by a Research Initiation Grant, Penn State University.
- Sallstrom, G. & Sallstrom, J. (1973), February-March). Singing exercises that develop and liberate the child voice. *National Association of Teachers of Singing Bulletin*, pp. 22-24.
- Sears, M. F. (1965). The tape recorder employed in the development of children's singing: An experimental study. *Colorado Journal of Research in Music Education*, 2, 8-12.
- 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, 104-108.
- Sinor, E. (1985). The singing of selected tonal patterns by preschool children. *Dissertation Abstracts International*, 45, 3299A. (University Microfilms, No. DA8723851)
- Smale, M. JH. (1987). An investigation of pitch accuracy of four- and five-year-old singers. *Dissertation Abstracts International*, 48, 2013A. (University Microfilms No. DA8723851).
- Smith, R. B. (1963). The effect of group vocal training on the singing ability of nursery school children. *Journal of Research in Music Education*, 11(2), 137-141.
- Smith, R. S. (1974). Factors related to children's in-tune singing abilities. *Dissertation Abstracts International*, 34, 7271A-7272A. (University Microfilms No. 74-11, 404)
- Vance, T. F., & Grandprey, M. B. (1929). Objective methods of ranking nursery school children on certain aspects of musical capacity. *Journal of Educational Psychology*, 22, 577-585.
- Vaughan, M. M. (1981). Intercultural studies in children's natural singing pitch and walking tempo. *Bulletin of the Council for Research in Music Education*, 66-67, 96-101.
- Welch, G. F. (1979). Poor pitch singing: A review of the literature. *Psychology of Music*, 7(1), 50-58.
- Wilson, D. S. (1971). A study of the child voice from six to twelve. *Dissertation Abstracts International*, 31, 5453A-5454A. (University Microfilms No 71-10, 796)
- Young, W. T. (1971). *An investigation of the singing abilities of kindergarten and first-grade children in East Texas* (Report No. PS-006-178). Nagodoches, TX: Stephen F. Austin State University.
- Zimmerman, M. P. (1981). Response. In *Documentary Report of the Ann Arbor Symposium* (pp. 315-317). Reston, VA: Music Educators National Conference.