This is to certify that the

thesis entitled

The Construction and Validation of a Method for
the Measurement of the Sight Singing Abilities
of High School and College Students

presented by

William Rodney Scofield

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Music

Major professor

Date December 5, 1979

THE CONSTRUCTION AND VALIDATION OF A METHOD FOR THE MEASUREMENT OF THE SIGHT-SINGING ABILITIES OF HIGH SCHOOL AND COLLEGE STUDENTS

By

William Rodney Scofield

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Music

1979

ABSTRACT

THE CONSTRUCTION AND VALIDATION OF A METHOD FOR THE MEASUREMENT OF THE SIGHT SINGING ABILITIES OF HIGH SCHOOL AND COLLEGE STUDENTS

By

William R. Scofield

The primary purpose of this study was. to develop a reliable and valid method for the measurement of the ability to sight-sing. Of specific importance was the practicality and usability of the test reflective of the method. A test was constructed using melodic exercises without rhythm and its reliability was determined in a pilot study involving 191 high school students attending Michigan State University Summer Youth Music. The test was found to be highly reliable (.969), possessed content validity as a result of the method of construction, and could be easily scored and administered.

Four data gathering instruments were used in the study: 1) a musical experience questionnaire; 2) a subjective rating - consisting of a numerical rating of the student's performance on a choral selection from standard literature; 3) a traditional standardized sight-singing test - the Otterstein-Mosher (O-M) test; and 4) the original unirhythmic sight-singing test. These four measurement devices were individually administered to the fifty-four high school and college students that volunteered for the study. The subjects were drawn from three distinct samples:

1) twenty high school students attending MSU Summer Youth Music 1979; 2) twenty-four students attending two small private religious colleges in the state of Michigan; and 3) ten students attending Michigan State University. The responses of the students were scored during the administration session and subsequently again after one week. Four additional music educators, previously unfamiliar with the study, were solicited to score the sample.

The results of the study indicated a high reliability (.974) for the test, high reliabilities for the twenty exercises used in the test, and extremely high scorer reliability. The coefficient for scorer reliability when scored by the researcher was r = .999. The coefficient of concordance between all scorers, the researcher and the four additional scorers, was found to be .995. Criterion-related validity was determined by comparing the performance on the standardized test with the unirhythmic test. A correlation coefficient of r = .926 was calculated establishing high criterion -related validity. Similarly high correlations were found between the subjective rating and both the O-M and unirhythmic tests. The reliability of the battery of three tests was computed and found to be .981 indicating a high degree of consistency between the three types of measurement devices.

Several other considerations were addressed; the reliability of the interval types used in the test, the relationship of experience to success on the test, and

the differences between the three groups as determined by analysis of variance and the Scheffe technique.

Two important conclusions were made from the results of the study. First, the method of measuring sight-singing performance using a test without rhythm was valid for the sample studied. The implication is that one need not measure all the elements of the skill to obtain a relatively accurate evaluation of the skill. Second, a sight-singing test is available that is not only reliable and valid, as are many others, but is highly usable as well. The usability is a result of its easy administration, immediate scoring capabilities, ease and speed of scoring, and the reliability of the scoring procedure.

Further study was recommended to determine if similar conclusions would be achieved using a much larger sample, equal sample sizes for the groups, and stricter controls throughout the study. In addition, the need for the unirhythmic test to be used in an actual classroom to insure its usability was noted.

To

Jason, Jennifer, and Jocelyn

ACKNOWLEDGEMENTS

My sincere gratitude is expressed to Dr. Robert Erbes, Chairman of the guidance committee and director of the dissertation. Appreciation is also extended to the members of the guidance committee: Dr. Merrell Sherburn, Dr. Daniel Russell, and Dr. Kazadi wa Mukuna; to Dr. Robert G. Sidnell whose guidance and inspiration were instrumental in the continuation of this process; to Dr. Albert LeBlanc and Ms. Suwatana Sookpokakit for their statistical assistance; and to both my parents and parents-in-law, Mr. and Mrs. M. Robert Scofield, and Mr. and Mrs. Danny D. Sain for their support and encouragement. A special debt of gratitude is expressed to a former committee member and cherished friend, Professor Richard Klausli, now deceased, for his contribution to my life and doctoral program.

To my wife, Rebecca, I reserve the deepest expression of love and appreciation for her constant support, tolerance, personal sacrifice, and her gift of unconditional love. Without her, this process would have been impossible.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued . . .)

LIST OF TABLES

LIST OF TABLES (Continued . . .)

CHAPTER ONE

THE PROBLEM

Introduction

Throughout the history of music in Western civilization the capacity to read and vocally reproduce the notated musical symbols has been a skill sought after by musicians. Men from the time of the Greeks to Philippe de Vitry and beyond have attempted to devise a concise method of notation, providing verification that the need to interpret musical symbols was important. One could speculate whether the purpose was for instruction, posterity, or codification. The fact remained that written symbols needed to be interpretable, and as the notation of music became increasingly prevalent, greater emphasis was placed on the ability to read and understand the symbols. Guido of Arezzo was instrumental in this endeavor. His digital music reading system was one of the first attempts towards the instruction of music reading. During the Renaissance, music and singing were an integral part of the classical education as part of the quadrivium. In Elizabethan England, an individual's upbringing was suspect if he could not read music. In the Enlightenment period, singing and music reading were considered equally as important as reading and writing, especially in Germany. Dr. Burney writes, "even at the common boarding schools, children are taught to sing hymns in parts."[1]

At the same time in England and in early America, the state of music was less than desirable. This was primarily due to the neglect of singing by the Puritans. The result was that in the early 18th century the "singing school" emerged in America.

> From the church choir there developed the need for instruction in music reading, which accounts for the rise of the "singing school," and ultimately for many other important musical developments.[2]

This situation continued until the mid-19th century when the efforts of Lowell Mason, patterned after Pestalozzi in Switzerland, resulted in the teaching of singing becoming a regular subject taught in public schools.

From that time until the late 19th century, schools began to employ more music teachers and supervisors. The emphasis towards sight-singing and music-reading also increased until it reached its apex around the turn of the century. During the period of approximately 1885 - 1915, a multitude of music reading methods books were developed and distributed, frequently through the "normal schools"

[1] Cedric Howard Glover, Dr. Charles Burney's Continental Travels, 1770 - 1772. (London: Blackie & Son Limited, 1927), p. 200.

[2] Russell N. Squire, Introduction to Music Education (New York: The Ronald Press Company, 1952), p. 4.

and "summer schools."

Although the methods and techniques used in music education have undergone substantial changes as a result of the influence of Pestalozzi, Rousseau, Dewey and other educational reformers since that time, sight-singing and music-reading have remained an important part of the curriculum of music classrooms. Contemporary writers have reiterated the importance of instruction in sight-singing on numerous occasions. In 1957, Burmeister wrote:

> If the role of music in general education is to secure immediate and limited performance per se, rote learning will suffice. If, however, the general music program is designed to assist in growth toward musical maturity, musical independence, and self-motivation for continuance in musical activities, our duty is clear: We owe every child the right to develop as far as his abilities, needs, and interests will permit him. We owe every child the right to learn to read music.[3]

Later, in 1966, Thomson stated:

> The most fundamental requirement for musicianship is the ability to translate the symbols of music notation into the sounds the composer intended-the ability to read music. Evidence is plentiful that experience alone is not an adequate foundation for fluent reading; on the contrary, the reader must learn how to approach any piece of music with understanding before his experience with music can be meaningful.[4]

[3] C. A. Burmeister, "The Role of Music in General Education," Basic Concepts in Music Education. The Fifty-seventh Yearbook of the National Society for the Study of Education. (Chicago: The University of Chicago Press, 1958), p. 229.

[4] William Thomson, Introduction to Music Reading, Concepts and Applications, (Belmont, California: Wadsworth Publishing Company, Inc., 1966), p. vii (preface).

Anne Marie de Zeeuw, in 1975, further claimed:

> A necessary tool for every musician is the ability to reproduce vocally a given line of music, often at sight. This has been proven beneficial not only as an analytical tool, but also as an important and convenient means of musical communication, especially for the instructor in the classroom and for the choral or instrumental conductor. Sight-singing is also a complimentary aid in the development of listening skills.[5]

Recognizing the historical and contemporary concern for singing and particularly sight-singing, one could assume that the skill of sight-singing was and is considered rather important. Statements as "we owe every child the right to learn to read music," "the most fundamental requirement for musicianship," and "a necessary tool for every musician" imply that the skill is still considered extremely important. In addition, the latter portion of the above quotes by Thomson and de Zeeuw allude to the premise that sight-singing is a complimentary or "building block" skill from which other musical outcomes are derived.

The historical and present importance of sight-singing, and the foundational characteristics of the skill necessitate its identification as a fundamental educational goal for the music classrooms of contemporary public education. An evidence that this has occurred would be the the preponderance of textbooks, methods of instruction,

[5] Anne Marie de Zeeuw and Roger E. Foltz, Sight-Singing and Related Skills, (Manchaca, Texas: Sterling Swift Publishing Company, 1975), p. v (preface).

learning devices, and instructional games that have been devised for simplifying the task of improving the skill of music reading and sight-singing. Many authorities in the field have written volumes expressly for this purpose. This abundance of materials available for the improvement of skill in sight-singing provides evidence that this skill is indeed considered essential to musical growth.

Need for the Study

With the historical and contemporary emphasis upon the skill of sight-singing, it is eventually necessary to assess the progress of the student as well as the success of the instructional method or system being used by the teacher. Charles Leonhard supports this view by stating:

> The validity and reliability of judgments of performance depend upon the skill and experience of the evaluators and upon the number of evaluators. Training in the technique of evaluating musical performance should receive specific attention in the preparation of musicians and music teachers. Every teacher of music must constantly evaluate his own performance and that of his students...Such training should take the direction of specifying the important factors in musical performance, clarifying concepts of those factors, and providing guided experience in the formation and application of standards of excellence to specific performances.[6]

Leonhard further identified the following reasons for making evaluations: 1) appraisal of student progress;

[6] Charles Leonhard, "Evaluation in Music Education," Basic Concepts in Music Education, The Fifty-seventh Yearbook of the National Society for the Study of Education. (Chicago: The University of Chicago Press, 1958), p. 327.

2) guidance; 3) improvement of instruction; 4) motivation; 5) improvement of program; 6) student selection; 7) maintenance of standards; and 8) research.[7] The instruments designed to evaluate music performance for any of the above reasons must be reliable and valid, implying that the setting of the test must be the same as that of the day to day occurrence of the skill. It is possible for sight-singing to occur in two distinct environments; in a group or rehearsal, or individually. Excluding the practice of private voice students and the studio, since sight-singing and the evaluation of sight-singing skills are not primary objectives of private vocal instruction, sight-singing could occur individually when the singer rehearses his or her music outside of the rehearsal for the first time. The frequency of this phenomenon is considerably less than might be desired by choral conductors, particularly at the high school level.

Sight-singing in a group or the rehearsal might occur on several occasions. Although the term "rehearsal" is defined as "a time to perform for practice, or a time to drill," sight-singing might occur at the following times: 1) during specific lessons designed to improve or measure the skill; 2) during rehearsals designated specifically for reading new music; 3) at the sight-singing segment of an ensemble festival; or 4) when the entire ensemble

[7] Ibid., pp. 312 - 315.

participates in singing another vocal part of particular difficulty.

Considering the relative infrequency of the individual sight-singing of the average choir member, it would seem that the majority of the sight-singing experienced by the student is in the rehearsal. If this assumption is true, several problems arise for the evaluator. To replicate the environment where most sight-singing occurs would imply that the students must be tested in a group setting. The inherent paradox here is that while sight-singing frequently occurs within the group, it is an individual skill. The ability of the student to sight-sing in the group could be substantially different from his ability to sight-sing alone. The possibility exists that when a student loses his/her place in the music, it can be found again simply by listening to the other performers. In addition, if there was more than one singer per part, the mistakes made by any student would be somewhat compensated by the others. The implication is that sight-singing would be extremely difficult to evaluate in a group performance, particularly for one evaluator. A multitude of evaluators would be necessary to accomplish the task in a group.

Despite the above difficulties, attempts have been made to measure the sight-singing ability of vocal students using group methods without performance. The first of the tests using these methods was written in 1924 by Raymond

Mosher.[8] shortly thereafter by William Knuth,[9] and more recently in a 1963 study by Adolph White.[10] In all three instances the results of the group tests were correlated with an individual sight-singing criterion to determine the validity of the methods. These tests will be discussed in greater detail in Chapter Two. It would seem that these tests, no matter how usable, cannot be definitive in evaluating the students's abilities to sight-sing when no actual vocal performance is required of the student.

While the above group tests seem less than totally successful for evaluating sight-singing ability, the intent of each was to provide a method for measuring sight-singing ability that would be practical, feasible, and usable. Many of the individual sight-singing tests currently available are both reliable and valid, yet are less usable because of the amount of time required to administer the tests and score the responses of the students. These tests

[8] Raymond Mosher, A Study of the Group Method of Measurement of Sight-Singing, (New York: Teachers College, Columbia University, 1925).

[9] William Edward Knuth, Knuth Achievement Tests in Music, (Philadelphia: Education Test Bureau, n.d.).

[10] Adolph Peter White, "The Construction and Validation of a Group Test in Music Reading for Intermediate Grades," (unpublished Ph.D. dissertation, The University of Minnesota, 1963).

were designed to allow educators to periodically assess the abilities of their students. From a purely educational standpoint, this periodic assessment should not be based upon the consideration of time. The musical development and growth of the student should far outweigh considerations of convenience for the teacher. Realistically, however, few public school teachers or college conductors have used or would use the published tests for precisely those reasons.

Before one can identify the type of test that would be reliable, valid, and usable, the current type of traditional tests must be analyzed.

The format of most traditional standardized sight-singing tests is quite similar. These tests are usually comprised of several melodic exercises, varying in length, but encompassing an entire musical thought or phrase. The number of melodic exercises varies, usually more than ten, involving various styles, keys or modes, and levels of difficulty. The rhythmic figures and melodic sequences include those that would normally be encountered by the singer in traditional choral repertoire. Some tests, particularly instrumental performance tests, include dynamic variations, articulation changes, and tone in the overall consideration of the ability of the performer. Sight-singing tests generally consider only two elements; pitch and rhythm.

Once the test is constructed, the administration procedure is relatively simple. Common testing procedures include setting the tempo, establishing the tonality and starting pitch, followed by the performance of the exercises while the investigator records the responses. The teacher or investigator must then spend the time necessary to score the tests according to the method indicated or adopted for the specific test. With several tests of this nature, this task, when completed, may have taken up to two hours or more per subject. This was experienced by Cooper,[11] Thostenson,[12] and this writer[13] in previously constructed sight-singing tests. The result is that few teachers ever use standardized tests, similar in their design to the aforementioned, that are carefully constructed, well-written, standardized and available. Recalling the reasons given by Leonhard for evaluation, whatever reason(s) is uppermost in the mind of the teacher, he usually relies upon an informal method of evaluating sight-singing ability. While these methods may indeed be reliable and valid, they have not been subjected to the careful study of

[11] John J. Cooper, "The Development of a Sight-Singing Achievement Test," (unpublished Ph.D. dissertation, University of Colorado, 1965).

[12] M. S. Thostenson, "The Study and Evaluation of Certain Problems in Ear-Training Related to Achievement in Sight-Singing and Music Dictation," Council for Research in Music Education, No. 11, (1967), pp. 14 - 35.

[13] A small pilot study conducted in 1977 in which a sight-singing test was constructed and validated. The scoring time per subject ranged from 1.5 hours to 2.5 hours.

music researchers to verify the premise of reliability and validity. In addition, because each method is essentially unique to the individual teacher, no norms are available precluding standardization or generalizations to the larger population of sight-singers, or comparisons between different groups of students.

Recognizing the need for some type of periodic assessment; the necessity of individual testing; the type of standardized tests currently available; and the limited usability of these tests because of the difficulties involved in administration and scoring, particularly the latter, the present study was conceived and undertaken.

Purpose of the Study

The purpose of this study was to devise a method of evaluating sight-singing ability such that a test reflective of this method could be easily administered and scored, yet be definitive regarding the abilities of the students. Only increased practice and experience can improve the administration and scoring time of the sight-singing tests currently available. Many of these tests cannot be scored during the administration session which would substantially increase the time necessary to obtain the results. The method to be devised should be definitive regarding the abilities of the students, provide for immediate scoring, and be consistent when scored by different persons or at different times.

There are two alternatives available that would reduce the scoring time and the difficulty of the scoring procedure for the teacher using a sight-singing test. First, the measure could be used as the unit of error, rather than the note. This alternative has been used by several researchers in the past. Kwalwasser stated, "the measure should unquestionably be the unit of error."[14] This method was also used in the individual tests by Mosher,[15] Otterstein and Mosher,[16] and for the instrumental performance scale by Watkins.[17] The deficiency inherent in this system is the loss of the ability to differentiate between the performer who makes a slight error and the performer who is totally incapable of interpreting the written musical symbols at the most rudimentary level. Both are awarded the same score on any given measure. The student who is unable to interpret the symbols would presumably perform consistently fewer correct measures, although successive slight errors would result in a similar

[14] Jacob Kwalwasser, Tests and Measurements in Music, (Boston, New York: C. C. Birchard and Company, 1927), p. 106.

[15] Mosher, op. cit.

[16] Adolph Otterstein and Raymond Mosher, O-M Sight-Singing Test; Manual of Directions, (Stanford, California: Stanford University Press, 1932).

[17] John Goodrich Watkins, Objective Measurement of Instrumental Performance, (New York: Teachers College, Columbia University, 1942).

score on an exercise. This method, however, substantially reduces the difficulty of the scoring and considerably decreases the time necessary to score the students.

The second alternative for reducing the difficulty in scoring and the scoring time would be to eliminate either pitch or rhythm from consideration in a sight-singing test. If pitch were eliminated, the result would be a test requiring no singing, similar to the group tests mentioned earlier. This would hardly be an acceptable alternative as the performer would not need any singing skills to achieve a relatively high score. Conversely, if rhythm were eliminated from consideration, the scoring time would be greatly reduced, the singing element would still be present, and perhaps this method would differentiate between students in the same manner as a traditional standardized test.

While a test reflective of this alternative would only be measuring pitch reproduction abilities, a logical inference could be made if the definition of a sight-singing test is one which measures the achievement or predicts the success of the student with regard to the singing of music at sight. If it can be shown that the singer who performs rhythmless, or unirhythmic, exercises well is also the singer who performs traditional sight-singing exercises with the same ability, logical conclusions could be made concerning the ability to sight-sing from the results of the unirhythmic test. If the above can be shown, a pitch reproduction test of this nature could then be classified as a sight-singing test. The following syllogism may prove helpful for clarification.

Table 1-1

IF	Sight-singing ability and achievement *equals* performance on a traditional standardized sight-singing test,
AND	Performance on a traditional standardized sight-singing test *equals* performance on a unirhythmic sight-singing test,
THEN	Performance on a unirhythmic sight-singing test *equals* sight-singing ability and achievement.

It is the intent of the investigator to make use of unirhythmic melodic exercises and compare the performance of the singers on these exercises with their performance on the exercises of a traditional standardized sight-singing test. This "short-cut" approach using unirhythmic techniques, if shown to be reliable and valid, may provide a usable measure for teachers in the field. The measures that are available are excellent devices for determining sight-singing ability and achievement, but are not as usable as might be desired in the opinion of some of the writers themselves, and as illustrated by the infrequency of their use.

Definitions for the Study

1. *Sight-reading* - the visual recognition of musical symbols accompanied by performance on a given instrument. It is further the translation of these notated

symbols to physical sound at first sight.

2. Sight-singing - the visual recognition of musical symbols accompanied by performance on the vocal instrument. Again it is the translation of these symbols at first sight.

3. Unirhythmic - having no specifically designated rhythmic duration for the written notes. The student should perform all pitches at approximately one note per second.

4. Pitch exercises (exercises) - the sequences of pitches that comprise the original unirhythmic test in this study.

5. Melodic exercises - the sequences of pitch with rhythm that comprise the criterion for this study and all traditional standardized tests.

6. Battery of tests - the three measurement devices used in this study.

Research Questions to be Answered

1. Are there significant differences between the performances of the three samples on the tests used in the study and between the musical experiences of the groups?

2. Is the original unirhythmic test reliable in terms of the test, the items and the scoring procedure?

3. Does the unirhythmic test display content validity and criterion-related validity when correlated to the standardized traditional sight-singing test?

Description of the Population

A sample of fifty-four students participated in the study, ranging from first-year high school students to graduate university students. All students were or had been choral students during their scholastic experience in either high school or college. All participating students were volunteers and were not remunerated in any way. The subjects were from three distinct samples: 1) high school, 2) small college, and 3) university students. Twenty high school students attending Michigan State University Summer Youth Music comprised the first sample. Twenty-four students attending either John Wesley College or Jordan College, two small private religious institutions in the state of Michigan, comprised the second sample. The third consisted of ten undergraduate and graduate students attending Michigan State University during the summer of 1979.

The results of the study will be limited to the stated sample. No attempt will be made to generalize the results to the universal population of singers, of which the present sample is a subset.

Design of the Study

The fifty-four subjects participating in the study were tested using four data-gathering instruments: a subjective rating; the Otterstein-Mosher (O-M) Sight-Singing Test; a unirhythmic pitch reproduction test (U-test); and a musical experience questionnaire. All tests were individually administered in the above order following the completion of the musical experience questionnaire. The responses were tape-recorded for scoring at a later date by the investigator and by the four music teachers in the field. The responses were also scored during the administration session and correlated with the second scoring by the investigator to determine scorer reliability. The data were analyzed at the Computer Center on the campus of Michigan State University using SPSS (Statistical Package for the Social Sciences), version 7.0. Reliability for the test, items, and intervals was established using Subprogram - RELIABILITY; Cronbach's Alpha. Item Analysis was determined using Pearson r's and Net D. Other correlation coefficients were found using Pearson r, and scorer reliability using Kendall's Coefficient of Concordance (W) and Cronbach's Alpha. Validity coefficients were calculated between the unirhythmic and O-M tests using Pearson r and Cronbach's Alpha.

Assumptions for the Study

1. The traditional standardized sight-singing tests reviewed in Chapter Two and the specific test used as the criterion for this study (Otterstein-Mosher) are generally accurate in their assessment of the sight-singing abilities of students. Specific considerations resulted in the selection of the Otterstein-Mosher Sight-Singing Test.

2. The student who sings a correct relative

interval at the wrong pitch level has performed an incorrect interval. (i.e. if within any exercise one or both of any two consecutive notes are incorrectly pitched, the interval encompassed by those notes is incorrect.) This would result from singing two consecutively wrong notes, although the distance between them is identical to the desired distance of the intended response.

Overview

In the ensuing chapters the main body of this study and the conclusions drawn from the statistical results will be presented. A review of the related literature, in particular the studies conducted in the area of the measurement of sight-singing during the past fifty years and pertinent to this investigation, comprises Chapter II. Chapter III is divided into two section: The first includes the presentation of the methodology and the results of the pilot studies while the second contains the methodology for the present study. The presentation of the statistical data, the analysis, and interpretation of that data comprises Chapter IV, followed by a summation, conclusions drawn from the statistics, and their implications for use in subsequent research presented in Chapter V. The bibliography and appendices conclude the document.

CHAPTER TWO

REVIEW OF LITERATURE

Introduction

Music reading, or more specifically sight-singing, is a subject that has received much attention during the recent decades. The instruction and achievement of the skill of sight-singing have been important areas of discussion in many books and articles concerning music achievement. As indicated in the preceeding chapter, many texts have been written specifically to aid in improving the ability to sight-sing. Few of these volumes contain any procedure for measuring or evaluating this ability. The studies that do confront the topic of measurement and evaluation of sight-singing ability fall into several categories: 1) studies of the factors that influence achievement in the skill; 2) group sight-singing methods and tests; 3) studies emphasizing some particular aspect of sight-singing or its assessment; and 4) studies which include the construction of a valid and reliable sight-singing or music-reading measurement device. The following material will be reviewed in the above order.

Factors That Influence Sight-Singing

Several studies have been conducted that have attempted to identify the various musical and nonmusical factors that contribute to achievement in sight-singing. The musical factors range from musical experiences, such as classes and private lessons, to musical abilities, such as pitch discrimination, tonal imagery, and other aptitude characteristics. The nonmusical factors include home environment, grades in school, reading ability, and others.

Frank Salisbury and Harold Smith[1] [1926] devised a sight-singing test "for predicting in advance the ability of the students to profit by the instruction given [sight-singing training and instruction]."[2] The test was designed for entering students at the State Normal School in Bellingham, Washington. The test was used for placement in the music classes. It consisted of four short songs each sung individually at sight. Against this criterion was correlated the six individual Seashore tests, the Thorndike Examination for High School Graduates, a test in melodic dictation, and an achievement test which the authors did not describe. Salisbury and Smith found the strongest predictors of sight-singing skill to be: 1) melodic dictation (r = .785); 2) the tonal memory test (r = .656; and 3) the pitch test (r = .645). They concluded the most effective combination for predicting sight-singing success was by

[1] Frank S. Salisbury, and Harold B. Smith, "Prognosis of Sight Singing Ability of Normal School Students," Journal of Applied Psychology, XIII, No. 5, (1929).

[2] Ibid., p. 425.

weighting the standard scores of these three factors resulting in a multiple correlation coefficient of r = .84. The three factors were weighted using the following factors: 1) pitch - .22; 2) tonal memory - .22; and 3) melodic dictation - .54.

Dean[3] achieved similar findings in 1937 when he determined the value of using the Seashore Measure of Musical Talent for predicting success in sight-singing. He found that the Seashore pitch test and the tonal memory test were the most accurate in predicting success in sight-singing.

Robert Ottman[4] [1956] conducted an experimental study at North Texas State University to investigate the influence of selected factors on the skill of sight-singing. The subjects for his study consisted of fifty-two second year music theory students at that institution. The author indicated that because of the sample, the study should be considered descriptive and the statistics valuable for that population only.

Ottman employed several measurement devices: 1) a

[3] Charles D. Dean, "Predicting Sight Singing Ability in Teacher Education," Journal of Educational Psychology, XXVIII (November, 1937).

[4] Robert W. Ottman, "A Statistical Investigation of the Influence of Selected Factors on the Skill of Sight-Singing," (unpublished Ph.D. dissertation, North Texas State College, 1956).

series of standardized published tests; 2) several original tests; and 3) a sight-singing criterion. The standardized published tests included: the Seashore Revised Measure of Musical Talent - pitch, rhythm, and tonal memory sections; the American Council on Education Psychological Examination; and the Nelson Denny Reading Test. The original tests included: an extension of the Seashore test; tests for tonic memory, melodic dictation, melodic dictation without rhythm, and melodic modulation; five tests on intervals; a music literacy test; and two questionnaires, one from the faculty, and the other a student self-evaluation.

The sight-singing criterion for the study was an unidentified melody by Cherubini. The melody contained all but two of the traditional intervals within the octave, from which Ottman concluded it would be a suitable instrument. The sight-singing criterion was administered as a pretest and again as a posttest. The reliability of the pretest, computed using the Spearman-Brown method was found to be r = .861, while that of the posttest was r = .904.

The results of the standardized and original tests were correlated with the sight-singing criterion and it was found that only two of the standardized tests achieved a coefficient significant at the .05 level of confidence. These two were the tonal memory segment of the Seashore tests and the Nelson Denny Reading Test. Conversely, all but two of the original tests correlated significantly with the criterion at the .01 level of confidence. The results showed that the original tests written by Ottman correlated hierarchically in the following order.

1) music literacy - r = .733.

2) the six interval tests had a group correlation of .678.

3) the tests of Melodic Dictation, Melodic Modulation, and Tonic Memory.

4) the two questionnaires.

A later study, strikingly similar to Ottman's was conducted at North Texas State in 1968 by Read.[5] This study also involved the investigation of the relationship of selected variables to sight-singing ability. Twelve musical and seven extramusical variables were selected for consideration. The purpose of his study was 1) to determine the relationship of the seventeen variables to the sight-singing abilities of students who scored high or low on the sight-singing criterion; 2) to determine the strength of said relationship; and 3) to discover which combination of variables produced the maximum likelihood for success in sight-singing. One hundred and twenty-five students were selected from the choral ensembles at North Texas State and given the sight-singing criterion as well as a battery of tests similar to the Ottman study. The results indicated that the student who excels in sight-singing has greater natural musical ability, more classwork

[5] John William Read, "An Investigation of the Relationship of Selected Variables to Sight-Singing Ability," (unpublished Ed.D. dissertation, North Texas State University, 1968).

in music, is older than the student with less ability, gives greater attention to details, and appears to be more intelligent.

The importance of the first three studies is that those factors considered significant to sight-singing ability were directly related to pitch or melody rather than strictly to rhythm. In the work of Salisbury and Smith, the factors that correlated most significantly with the criterion were melodic dictation, the Seashore pitch and tonal memory tests. In the Dean study, again the pitch and tonal memory tests of the Seashore battery were most significant when compared with the sight-singing criterion. Finally, in the Ottman study, the results of the Seashore tonal memory test and the interval tests point to this conclusion. Even in the sight-singing criterion used by Ottman, the rhythmic continuity was virtually destroyed as a result of the administration technique. In the event of a melodic error, the singer was stopped, given the correct pitch, then instructed to continue from the point of the error. This technique was also used in several other studies to be reviewed later.

Group Sight-Singing Tests

One of the concerns of researchers and teachers using a sight-singing test is the usability of the instrument. As indicated earlier, the usability is generally stated in terms of ease of scoring and administration and length of time necessary to obtain the results. One approach to this concern is to assess the entire group at one sitting. Several researchers have proposed group sight-singing tests in an attempt to facilitate quick and easy measurement of large numbers of persons. The tests constructed for this purpose essentially follow the same pattern. They generally consist of items played on the keyboard or some other instrument while the student selects the item heard from the four or five available choices. The student's sight-singing ability is determined by the number of correct responses on the multiple-choice test. Music theorists would certainly identify these tests as melodic perception, pitch discrimination, error detection, tonal imagery tests, but definitely not sight-singing tests as no singing is necessary.

One of the first sight-singing tests of this nature was constructed by Raymond Mosher in 1925.[6] The purpose of his study was to determine the correlation of the group tests with a sight-singing criterion to identify the combination of group tests which would provide the best index of sight-singing achievement. The battery of group tests included:

1) knowledge of musical symbols, marks of expression and general music information;

2) recognition of scales, chords, and intervals;

[6] Raymond M. Mosher, A Study of the Group Method of Measurement of Sight-Singing, (New York: Teachers College, Columbia University, 1925).

3) knowledge of measure and note values;

4) ability to identify well-known melodies when read silently;

5) ability to write tonal figures or patterns from hearing them played on the piano;

6) ability to write rhythmic patterns from hearing them played on the piano;

7) ability to write melodies from dictation.

The results of these group tests were correlated with the individually administered sight-singing criterion, a test compiled by Mosher from songs in standard elementary music texts. Twelve songs were used ranging in difficulty, time signatures, keys, but not in mode. The method of scoring was to use the measure as the unit of error. If an error occurred in pitch within a particular measure, a minus (-) was indicated above the measure. If the same occurred in rhythm, a minus was indicated below the measure, creating a total of two points per measure.

The tests were administered to 723 children from grades five to eight in three communities across the country. Reliability of the group test was computed in two ways; Spearman-Brown (r - .909), and split halves (r = .952). The individual sight-singing criterion was found to correlate with the seven group tests above having the following coefficients.

1)	.3375
2)	.3708
3)	.3407
4)	.4730
5)	.6212
6)	.3425
7)	.4847

The results of the study indicated the segments of the group tests that correlated with sight-singing were the ability to recognize familiar melodies, the ability to take melodic dictation, and the ability to write melodies from dictation.

The overall correlation of the group test to the sight-singing criterion was found to be r = .65. In Mosher's words, "Assuming the latter instrument [his individual test] to be the criterion for achievement one concludes that the effectiveness of the group test is limited."[7] He further stated:

> It is obvious that the error in judging sight-singing will be fairly wide should a teacher estimate an individual pupil's achievement in terms of a group test score alone. Nevertheless, it seems safe to say that however rough and crude such an estimate may be, it is likely to be more accurate than the subjective judgment of the classroom teacher.[8]

About the same time, Herbert Hutchinson,[9] Director of Music in Columbus, Ohio, developed a test subtitled "Silent Reading and Recognition." This test consisted of the musical notation of twenty-four different familiar compositions from which a phrase or two was extracted from the melody and presented. The items were divided into six

[7] *Ibid*., p. 59.

[8] *Ibid*.

[9] Herbert E. Hutchinson, *The Hutchinson Music Tests*, (Bloomington, Illinois: Public School Publishing Co., n.d.).

groups of four selections each. Below each group the names of eight familiar selections appeared. The task for the student was to match the titles of the pieces with the appropriate excerpts. Kwalwasser wrote concerning the test; "This test is a significant one, for it measures a mental process that conditions success in music reading; namely, tonal imagery."[10]

Several years later, William Knuth[11] devised a test to measure a student's silent recognition and mental comprehension of notated musical ideas. He defined mental recognition and comprehension as sight-reading, rather than the actual physical performance of the notated symbols. He indicated that this test would determine whether difficulty in sight-singing was caused by incomplete comprehension or by actual physical problems.

The student was asked to listen to several short melodies while observing the musical score before him. The problem was to determine which of the three choices was the one played by the test administrator to complete the melody.

The reliability coefficient computed for the test was found to be r = .958. Knuth claimed that the test was valid for three reasons: 1) a detailed textbook analysis

[10] Jacob Kwalwasser, Tests and Measurements in Music, (Boston: C. C. Birchard and Company, 1927), p. 79.

[11] William E. Knuth, Knuth Achievement Tests in Music, (Philadelphia: Educational Testing Bureau, 1936).

which isolated the various problems of sight-singing as taught in the elementary school; 2) the pooled judgment of experts on each of the test items; and 3) an actual usage of the items discarding those whose percentage of correctness did not increase at higher grade levels.

In his 1963 study, Adolph Peter White[12] constructed and validated a group test in music reading for intermediate grades. The method of his test was similar to that of Knuth and Mosher. Again the students heard a series of melodic and rhythmic items and identified the one played from the possible choices. There were sixty items in his test; thirty-seven were pitch items and twenty-three were rhythmic items. The final form of the test was derived following a pilot study and was subsequently administered to 5,641 students from ten Minnesota school systems.

An individual sight-singing test consisting of ten pitch items and seven rhythmic items was administered to 398 of these students. The correlation between the group test and the sight-singing criterion for these students ranged from .83 to .89. The author concluded that "test results seem to indicate that music reading ability may be determined by measuring the ability to hear the musical

[12] Adolph Peter White, "The Construction and Validation of a Group Test in Music Reading for Intermediate Grades," (unpublished Ph.D. dissertation, The University of Minnesota, 1963).

score."[13] He also pointed out that there now exists a standardized group test for music reading whereas the previously cited studies were unstandardized. He identified several uses for the test and cited its value for testing large groups. He further indicated the norms to be used for comparison of groups of students.

In addition to the studies reported, several other tests have been constructed that utilize the same or similar techniques. The Kwalwasser-Ruch Test of Musical Accomplishment[14] has several sections that measure the recognition of familiar melodies from the notation and the detection of pitch and rhythm errors in familiar melodies. The Diagnostic Tests of Achievement in Music[15] by Kotick and Torgerson also includes a section devoted to recognizing familiar melodies and songs. The Farnum Music Notation Test[16] consists of forty four-measure melodies. In each melody one of the measures differs from the printed

[13] Ibid., p. 128

[14] Jacob Kwalwasser and G. M. Ruch, Kwalwasser-Ruch Test of Musical Accomplishment, (Iowa City, Iowa: Bureau of Research and Service, State University of Iowa, 1924).

[15] M. L. Kotick and T. L. Torgerson, Diagnostic Tests of Achievement in Music, (Los Angeles: Los Angeles Test Bureau, n. d.).

[16] Stephen E. Farnum, Farnum Music Notation Test, (New York: The Psychological Corporation, 1949).

score in either pitch or rhythm. The student must identify the measure in which the change has been made. This test was found to correlate favorably with musical aptitude tests and scores on instrumental performance scales. The Jones Music Recognition Test[17] is similar to the Hutchinson test previously reviewed. There are two parts to this test, one for elementary and junior high students consisting of eighty items, and the other for high school and college students containing one hundred items. Familiar musical excerpts in groups of ten are played on the keyboard and the student associates the appropriate title with the melody from the twelve possible titles provided for each group.

Each of the previously cited references is predicated on the same basic assumption; some form of music-reading is involved for the student to be able to recognize music from the printed score. This would seem to be a logical conclusions, although it is possible to recognize music from its familiar rhythm. Even if this were the case, there would still be some form of music reading occurring. The major issue concerning the above group tests is whether the ability being measured is sight-singing ability. At best, the activity being measured would be classified as melodic discrimination or tonal imagery. Perhaps it is a misnomer to refer to these tests

[17] A. N. Jones, Jones Music Recognition Test, (New York: Carl Fischer, Inc., n.d.).

as sight-singing tests since no actual vocal response is required.

The value of the above tests would be in the practicality of their use. As mentioned earlier, traditional standardized tests are used rather infrequently by teachers in the field due to the lack of usability. Certainly group tests could alleviate part of the problem of usability, but would create others concerning interpretation of the results and the relationship of the skills involved in mastering the group tests and sight-singing.

Relevant Studies Concerning Sight-Singing

Many studies have been conducted that relate specifically to some aspect of sight-singing. These studies contribute to the already large body of information that has been amassed regarding the skill of sight-singing. The studies to be reviewed are those that in some way are germaine to the present study.

In the early sixties, three studies were conducted by Ray,[18] Hammer,[19] and Barnes[20] involving the use of the

[18] Harry Burton Ray, " An Experimental Approach to the Reading of Pitch Notation," (unpublished Ph.D. dissertation, Indiana University, 1964).

[19] Harry Hammer, "An Experimental Study of the Use of the Tachistoscope in the Teaching of Melodic Sight-Singing," (unpublished Ed.D. dissertation, University of Colorado, 1961).

[20] James Woodrow Barnes, "An Experimental Study of Interval Drill As It Affects Sight Singing Skill," (unpublished Ph.D. dissertation, Indiana University, 1960).

the tachistoscope for improving the skill of sight-singing. This technique was used earlier by Bean[21] in 1938, and again in the mid-forties by Stokes[22] and Christ.[23] In each of these studies several techniques or findings relate to the present.

In the study conducted by Harry Burton Ray, a comparison was made between two types of drill with pitch patterns. Subjects from the first year music theory class at Western Michigan University were divided into experimental and control groups. An individual sight-singing test was administered to all subjects prior to the experiment and again following the experiment. Each group was given the same amount of instruction totalling fourteen hours over a ten-week period. There were two distinct differences in the type of instruction given the experimental group: 1) the hours of drill were using the tachistoscope whereas the control group was drilled from mimeographed sheets; and 2) the pitch patterns of varying length and complexity were not identified with a particular key,

[21] Kenneth L. Bean, "An Experimental Approach to the Reading of Music," Psychological Monographs, L (1938).

[22] Charles F. Stokes, "An Experimental Study of Tachistoscopic Training in Reading Music," (unpublished Ph.D. dissertation, Teachers College, University of Cincinnati, 1944).

[23] William E. Christ, "The Reading of Rhythm Notation Approached Experimentally According to Techniques and Principles of Word Reading," (unpublished Ph.D. dissertation, Indiana University, 1953).

only the appropriate accidentals.

The results of the comparisons indicated that the sight-singing skill of the subjects increased substantially, although there was no significant differences between the two groups. It must be noted that the results may have been somewhat contaminated by the conflicting influence of the tachistoscope and the absense of key reference.

Two elements from this study are important to the present study: 1) the sight-singing test used as the pretest and posttest; and 2) the pitch patterns used in the study were rhythmless pitch sequences consisting of intervals of major and minor seconds and thirds.

Harry Hammer chose two beginning fourth grade music classes as the subjects for his study. He devised a method for teaching sight-singing to children. An appropriate sight-singing test was also written to assess the melodic sight-singing abilities of the children. The investigator conducted the instruction in both classes, the only difference being the period devoted to tonal pattern practice.

The control group was instructed using conventional teaching methods, while tachistoscopic techniques were used with the experimental group. The groups were instructed in this manner bi-weekly for thirty-eight weeks at which time the sight-singing criterion was again administered. Hammer then inverted the groups so that the control group became the experimental group and vice versa. An additional fourteen weeks of instruction, three times per week, were given in the same manner as before only using improved techniques. Following the second period of instruction, the sight-singing criterion was administered the final time. Hammer concluded the following contributions were made:

1) an effective technique for teaching melodic sight-singing ability had been developed;

2) an experimental design for further research using the tachistoscope in the area of music reading had been developed;

3) an achievement test of melodic sight-singing ability was now available which could be used to measure the effectiveness of various techniques for developing sight-singing skills.

James Woodrow Barnes stated the purpose of his study was to 1) investigate the effect of group drill in sight-singing specific intervals upon the individual student's ability to sight-sing the same intervals and 2) identify the correlation between the ability to sight-sing isolated intervals and the ability to sight-sing melodies comprised of the same intervals. Two secondary purposes were also identified: 1) to determine the order of difficulty for the intervals used in the study and 2) to compare sight-singing modal or modulating melodies with major, minor or nonmodulating melodies in relation to the singing of intervals at sight.

The subjects for his study consisted of forty-six second term music theory students from Indiana State Teachers College during the year 1958 - 1959. The students were separated into experimental and control groups by the scores achieved on several standardized aptitude and achievement tests. The two groups received the same classroom instruction with the exception of approximately ten hours of tachistoscopic drill in the sight-singing of intervals given to the experimental group over a ten-week period. It should be noted that the interval drills and tests were notated in half notes, or unirhythmically, and without key signature. Only the necessary accidentals for the altered notes were used.

Barnes developed and validated two sight-singing measures; a melodic sight-singing test and an isolated interval sight-singing test. The validity and reliability of these tests were determined by administering both tests to sixty high school and college music students. The correlation coefficient between the test and retest scores for the Interval test was r = .968 and for the Melodic test, r = .945. A rank-order correlation coefficient was also determined for the forty college students by comparing the scores on the Melodic test with the composite ratings of three theory instructors. This coefficient was found to be r = .891.

The results of the study indicated that the experimental group performed significantly better than the control group in both the singing of intervals (.01 level of confidence) and the singing of melodies (.05 level of confidence). A high correlation (r = .916) was found between intervallic and melodic sight-singing abilities. This resulted from comparing the individual student's scores on the melodic and intervallic sight-singing tests and was significant at the .01 level of confidence. A correlation coefficient was also established between the improvement of melodic sight-singing ability and intervallic sight-singing ability (r = .525). This figure was also significant at the .01 level of confidence, though substantially lower than the correlation between intervallic and melodic sight-singing abilities.

Barnes also derived an intervallic difficulty rating from the performance of each interval 2080 times by 208 singers. Each of the intervals was presented ten times, five times in each direction. The intervals considered were the major, minor and perfect intervals through the fifth. Table 2-1, from Barnes's dissertation, illustrates the difficulty ratings of the intervals studied.

Table 2-1

Interval Difficulty Rating from Barnes's Study

Interval	Direction	Number of times presented	Number of times missed	Rank order of difficulty
Major third	Down	1040	432	1
Perfect fifth	Up	1040	394	2
Perfect fifth	Down	1040	382	3
Perfect fourth	Down	1040	372	4
Perfect fourth	Up	1040	345	5
Major second	Down	1040	342	6
Minor third	Down	1040	315	7
Major third	Up	1040	287	8
Minor third	UUp	1040	284	9
Major second	Up	1040	257	10
Minor second	Up	1040	161	11
Minor second	Down	1040	107	12

In a study somewhat similar to Barnes's, James Marquis[24] investigated the variation in performance of intervals within a melody sung at sight compared to that of intervals in isolation sung at sight. It was his assumption that more is involved in sight-singing than the ability to sing isolated melodic intervals apart from a melody. The purposes of his study were to determine:

1) whether the ability to sight-sing specific intervals changed as the context surrounding those intervals changed;

2) whether students tend to sight-sing intervals within a melodic context the same as those in isolation;

3) whether the degree of skill of students to sight-sing intervals in isolation differs from that of singing the same intervals under different conditions of melodic context.

The subjects for his study consisted of fifty-twc first-year music theory students at the University of Iowa. The investigator individually administered the Sightsinging Criterion and the Isolated Intervals Test to each student and tape-recorded the responses. When a mistake was made, the student was stopped, corrected, then allowed to continue from the point of error. This administration technique was similar to one used by Ottman.[25] In Ottman's study, the rhythmic continuity of the Cherubini melody was

[24] James H. Marquis, "A Study of Interval Problems in Sight Singing Performance with Considerations of the Effect of Context," (unpublished Ph.D. dissertation, University of Iowa, 1963).

[25] Ottman, "A Statistical Investigation...," op. cit.

virtually destroyed when mistakes were made using this procedure. In the Marquis study, the investigator minimized the effects of rhythm through simple usages. Presumably the singer should have been able to sing the rhythms with relative ease. The implication was that mistakes were directly related to the intervallic context.

The Sightsinging Criterion used in this study was developed by the researcher over a period of years. Validity of the criterion was established by correlating the scores on the criterion with the student's grade point average in Sightsinging and Ear Training and with scores on the final semester Aural Skills dictation test. The coefficients of validity for the Sightsinging Criterion and the Isolated Intervals Criterion test were found to be r = .776 and r = .828 respectively. The Sightsinging Criterion was found to have a reliability of .979 (Spearman-Brown), significant at the .01 level of confidence. The Isolated Intervals Criterion was found to have a reliability of .830 (Spearman-Brown), also significant at the .01 level of confidence.

Like interval types were classified into three categories according to their contextual setting; simple, moderately complex, and complex. Comparisons were made of multiple sets of interval types, each set containing one interval drawn from the three contextual settings. A total of forty-five complete sets and fourteen partial sets were analyzed. The descending tritone, ascending minor sixth, both ascending and descending major sixths, and the descending major seventh were not analyzed because their occurrence was always within the same context. The results of the study revealed that:

> 1) first year college level music students are significantly affected by differences in the contextual setting of the intervals while sight-singing;
>
> 2) ability or lack of ability in the singing of an isolated interval does not directly affect the singing of that interval in a melody;
>
> 3) in melodic sight-singing, ability to perceive the basic quality of intervals is considerably less important than ability to perceive the scalar. harmonic, and tonal changes across or surrounding intervals.[26]

The findings of Marquis are somewhat in conflict with those of Barnes. The latter found a significant relationship (r = .916) between the abilities of intervallic and melodic sight-singing. This conflict of results is minimized when one realizes that Marquis was measuring the ability to sight-sing intervals in isolation consistently with those sung in context. Conversely, Barnes was only measuring the correlation between sight-singing intervals and sight-singing melody. He was interested in whether the ability to sight-sing intervals would be helpful in predicting melodic sight-singing ability. It should

[26] Marquis, op. cit., p. 173.

also be noted that Marquis used all twelve common intervals while Barnes used only the major, minor and perfect intervals through the fifth.

In a recent study conducted by Lewis Danfelt[27] [1970] at Florida State University, the investigator attempted to determine if the type of melodic material used in the rehearsal and performance of sight-singing exercises would affect the student's ability to sight-sing. The two types of melodic material used in the study were composed music and contrived music. Danfelt defined the two types as music written for the purpose of performance, and music written for the sole purpose of acquiring the skill of sight-singing.

The subjects for the study consisted of forty-one first-year theory students at the University of Kentucky. The experimental group used the contrived music while the control group used the composed music. Each group received thirty minutes of instruction twice weekly for fourteen weeks. At the end of this period both groups were retested with the individual sight-singing test. The pretest and posttest melodies contained only quarter and half notes. The author cited two reasons for this: 1) to maintain an elementary rhythm, and 2) to make an accurate stroboscopic reading less difficult to obtain. A correct

[27] Lewis Seymour Danfelt, "An Experimental Study of Sight Singing of Selected Groups of College Music Students," (unpublished Ph.D. dissertation, The Florida State University, 1970).

response was identified as ± 50 cents on the equal tempermant scale.

The author reported that although both groups made substantial gains in the ability to sight-sing, there was no significant difference between the groups. Danfelt concluded: "As a result of this study, it was determined that the type of material selected [contrived or composed] had no influence upon sight-singing performance."[28]

The comparison of two methods of measuring achievement in sight-singing was the subject of a study conducted by John Charles Nelson.[29] The purpose of his study was to determine which of the two methods, the traditional long item test or a short item test, a newer approach, would prove more reliable.

Three tests were used in this study. The first was an objective multiple-choice dictation test comprised of three sections; pitch, rhythm, and melody. The second test was the short-item test including four sections: intervals, pitch patterns, rhythmic phrases, and melodies. The final test was a traditional long-item test constructed by the investigator expressly for the study. Both the short-item and the long-item tests were found to have reliability coefficients in excess of .95.

[28] *Ibid*., p. 58.

[29] John C. Nelson, "A Comparison of Two Methods of Measuring Achievement in Sight Singing," (unpublished Ph.D. dissertation, The University of Iowa, 1970).

The results indicated that the short-item test proved more valid when compared to the dictation criterion. It also proved to be significantly more efficient as a measurement device. The administration and scoring time for the short-item test was approximately twenty minutes as opposed to an hour for the long-item test. The author identified three advantages of using the short-item test over the long-item test:

1) it contained a wider variety of performance problems, especially in rhythm;

2) it was more valid in terms of its higher correlation with the dictation test as the criterion;

3) it was much more efficient.

He concluded that within the limitations of his investigation, the short-item method of measuring sight-singing was significantly better than the traditional long-item method.

Thomas Ritchie[30] conducted a study concerning the effects of diatonic harmony upon the ability to hear melodic fragments. His test consisted of twenty four four-note melodic fragments, each with three harmonized settings and one unharmonized. The subjects were to identify and write the second, third and fourth notes of the sequence in whole-note values, after hearing them played from a

[30] Thomas V. Ritchie, "A Study of the Effects of Diatonic Harmony Upon the Aural Perception of Selected Melodic Fragments," (unpublished Ph.D. dissertation, Indiana University, 1960).

tape recording.

As a part of the results of the study, Ritchie concluded that aural perception of melodic fragments was not aided by the presence of surrounding diatonic harmony; in fact, such harmonizations often substantially reduced accurate perception. He found that the correlation coefficients between his test and grades in sight-singing and harmonic dictation were r = .73 and r = .67 respectively. Both these figures were significant at the .001 level of confidence. Ritchie concluded:

> This would indicate a high degree of association between ability in sight-singing and achievement in taking dictation of the type which made up this test, and might have some relevance to the problem of the value of drill in dictation as an aid to sight-singing.[31]

The above conclusion supports the earlier premise that sight-singing is a complimentary or "building block" skill that contributes to other musical outcomes. In addition, Ritchie's conclusions support the premise that there are several music competencies that correlate highly with the skill of sight-singing. The group test methods revealed a high correlation with sight-singing; the Barnes' study verified a high correlation between melodic sight-singing and the ability to sight-sing intervals; and Ritchie found that sight-singing correlated favorably with melodic dictation. These studies seem to indicate that any or all of

[31] *Ibid.*, p. 137.

the correlating skills identified here might be used as predictors of success in sight-singing.

The Construction and Validation of Sight-Singing or Sight-Reading Tests

For definitive measurement of sight-singing or sight-reading to occur, valid and reliable tests are necessary. Several studies have been conducted whose purpose was defined as the attempt to construct and validate such measures. In addition to these tests constructed for specific studies, several others have also been written claiming validity, reliability, and in some cases, standardization. These tests fall into two categories, each requiring some type of performance: 1) vocal sight-singing tests and 2) instrumental sight-reading tests.

Sight-Singing Tests

One of the first attempts, if not the first, was a study in 1924 by Earl Hillbrand.[32] The purpose of his study was "to derive a scale for the measurement of ability in sight-singing."[33] The test was constructed using the repertoire found in elementary music books as examples. The unit of error for the twelve exercises was the single note and the errors were defined as follows:

[32] Earl K. Hillbrand, Measuring Ability in Sight Singing, (Ann Arbor, Michigan: Edwards Brothers Pub., 1924).

[33] *Ibid.*, p. 1.

1) complete errors - that is, errors in oral sight reading which indicate clearly that the intervals are quite new to the child, or at least, not under his control or beyond his ability;

2) partial errors - that is, transpositions into another key of a group of notes;

3) omission of notes;

4) insertion of notes not included in the music score;

5) repetitions;

6) hesitations;

7) time in seconds for each song.[34]

Hillbrand selected 1487 fourth, fifth, and sixth graders as the subjects for his study. The subjects were tested individually by the author and the scoring was done as the students sang. The basic statistical computation was the percentage of error for the individual students and for the various grades on all the songs. It must be remembered that this study was a pioneer work and the statistical procedures and recording devices available in 1923 were unsophisticated or nonexistent.

The Otterstein-Mosher Sight-Singing Test[35] was an individual test for the purpose of determining a student's achievement in reading music at sight. The test consisted of twenty-eight exercises progressing in difficulty both

[34] Ibid., p. 2.

[35] Adolph Otterstein and Raymond Mosher, Manual of Directions: O-M Sight-Singing Test, (Stanford, California: Stanford University Press, 1932).

melodically and rhythmically. The first exercises consisted basically of scalewise motion, no altered tones, and no minor melodies. The later exercises often began on some other note than the tonic, frequent leaps, more altered tones, and minor melodies. The test was designed to measure pitch and rhythm only. Tone quality, dynamic variations, and articulation changes were not a part of the test.

After background information was gathered, the examiner indicated the tonality and the starting pitch and the student responded by singing the exercise. The unit of error was the measure and there were two points per measure, one for pitch and the other for rhythm. A total of two-hundred twenty-four measures were included in the test making the total number of points four-hundred forty-eight. If an error in rhythm occurred within a measure, a minus (-) was indicated below the measure. If the same occurred in pitch, a minus was indicated above the measure.

The authors claimed validity since the testing conditions closely resembled the actual classroom experience and the exercises were those typical of the music class. The reliability of the sight-singing test was computed in two ways. First, a correlation was made between the odd and even exercises resulting in a coefficient of r = .935 for rhythm and r = .979 for pitch. Second, the odd and even measures were compared and a correlation coefficient of r = .982 was obtained for rhythm and r = .996 for pitch.

Another sight-singing test developed by Thostenson was called the CSS76 Criterion Sight Singing Test.[36] The four sections of this individually administered test included: 1) an interval test of the twelve basic intervals; 2) twenty-four pitch sequences without rhythm; 3) twenty rhythm sequences without pitch; and 4) twenty two-measure melodies with both pitch and rhythm. This test can be administered in approximately thirty minutes. The reliability coefficient was calculated to be .95 when used with college students.

Thostenson later developed a group measure called the PRM78 Dictation Test. This test consisted of thirty pitch items with unchanging rhythm; twenty-four rhythm items with unchanging pitch; and twenty-four items with both pitch and rhythm. As in the previously discussed group measures, the student listened to the item and chose the answer to the question that correctly identified what was heard. The reliability coefficient for the group test was .88 and the correlation coefficient between the group test and the sight-singing criterion was r = .85. Both the individual test and the group test were weakest in the measurement of rhythm.

[36] M. S. Thostenson, "The Study and Evaluation of Certain Problems in Ear-Training Related to Achievement in Sight-Singing and Melodic Dictation," Council for Research in Music Education, XI (1967), pp. 14 - 35.

A recent study was conducted by John J. Cooper[37] [1965] in which he constructed and validated a test to measure the degree of achievement in sight-singing attained by college students. Secondary questions to be answered were: 1) what are the factors that constitute difficulty in sight-singing; 2) what arrangement will result in a hierarchy of items from simple to difficult; and 3) what are the characteristics of a good test and to what extent does this test meet this criteria. The subjects for the study were one-hundred-and-two music majors from five colleges in the state of Tennessee. He hoped the test could be used as 1) a diagnostic instrument for determining sight-singing errors; 2) a placement or qualifying examination; 3) a determiner of growth in sight-singing; or 4) a procedure suitable for further research in the development of performance tests in music.

In constructing the test, Cooper gave attention to the factors that should be included in a sight-singing test. The melodies were written specifically for the test in what he indicated was the style of conventional Western music. All of the traditional intervals were used, with the exception of the diminished fifth (the tritone was used) and the inclusion of the augmented second. The rhythms included figures that would normally be confronted

[37] John J. Cooper, "The Development of a Sight-Singing Achievement Test," (unpublished Ph.D. dissertation, University of Colorado, 1965).

in the literature becoming progressively more difficult in the later items. The tempi of the items were those of a moderate nature and the meters included 4/4, 2/4, 3/4, 6/8, and 4/8. The keys included all those through five sharps and flats, some minor keys and one modal melody (phrygian). It was decided to use the treble clef only and to have ten items per form.

Cooper elected to use the note as the unit of error. He indicated that "only an occasional study used the measure; the majority favored the note as the unit for scoring."[38] It was further decided that the interval was to be the only melodic element to be used in scoring. Since pitch is relative, it was thought that a change of key within an item would result in one error in interval. The number of errors possible with regard to rhythm was the number of notes and rests. The test was tape-recorded with the first listening for the purpose of scoring rhythm and the second for scoring pitch. Practically, the actual time needed for scoring the test varied with each subject depending upon the number of errors made. The investigator identified the range for scoring time to be from forty-five minutes to two hours per subject.[39]

Cooper claimed content validity because of the

[38] Ibid., p. 35.

[39] Ibid., p. 68.

factors mentioned above concerning the procedure used in constructing the test. He stated that because elements of music were used that are found in conventional Western music, content validity was obtained.[40] He further claimed that because the test displayed a wide range of scores, it was valid. Cooper also elected to use experts in the field to independently evaluate the content of the test and judge the appropriateness of the test for the measurement of the ability to sight-sing. This option was chosen for establishing criterion-related validity over a correlation with an already established test because the author felt that no suitable criterion existed. He did note, however, that "conclusive evidence of validity, other than that of face, or possible content validity, was not available from the data produced in the study."[41]

The coefficient of equivalence was chosen as the means to obtain the reliability coefficient. This method was selected because of the parallel forms of the test. It was found that the two forms of the test showed no significant difference between the means at the .05 level of confidence. The Pearson r computed between the rhythmic scores yielded r = .41. The same coefficient computed between the melodic scores produced r = .51. The author

[40] *Ibid.*, p. 61

[41] *Ibid.*, p. 64.

noted that these were fairly low, however, the composite score comparison resulted in r = .88. Cooper reasoned that from this data neither rhythm nor melody could be considered as separate entities, but must be considered as a whole. The results of Otterstein and Mosher and Thostenson would not support this reasoning.

The usability of the test was questioned by Cooper because of the time element required to score the test. The administration time for the two forms of the test was fourteen minutes combined, well within the time that would be convenient for other individual sight-singing tests. The reader will recall that the Thostenson test took thirty minutes to administer. The scoring time for the Cooper test, however, ranged from forty-five minutes to two hours, as indicated earlier. This would all but preclude its use by the average teacher. Cooper concluded the following from the results of the study:

> Within the scope and limitations of this study, the primary conclusion was that the test developed in this study is a reliable instrument for the measurement of sight-singing achievement. The establishment of its validity, as determined through analysis of content and the judgment of experts, indicated a face or content validity. It meets most of the criteria of a good test, with the exception of usability. It appears that the scoring method, although accurate and objective, is too lengthy to be considered practical and utilitarian.[42]

[42] _Ibid._, p. 84.

He further concluded:

> 1) the rhythmic difficulties seem to present less of a problem to the singer than tonal difficulties;
>
> 2) that certain musical intervals, namely the augmented second, augmented fourth, and major seventh appear to be difficult in any context;
>
> 3) that difficulty, as a factor of sight-singing achievement, cannot be defined or explained as an absolute quality of rhythm or melody, but rather is dependent to a large degree upon context.[43]

Cooper made several recommendations for further study, among the most significant to the present study was the question, "is it possible to develop a scoring technique for use with a sight-singing performance test that would render the test more usable without sacrificing objectivity and scoring accuracy."[44]

Instrumental Sight-Reading Tests

There have been several studies involving the development of performance tests in instrumental music reading which can give credence to the method of construction and scoring. While instrumental performance tests have little practical relation to sight-singing tests, the method of construction, the philosophy, the scoring method, and the administration procedures have significant relation to sight-singing tests. Three studies have been conducted

[43] Ibid.

[44] Ibid., p. 85.

that are worthy of note: one by John Watkins[45] in 1942; Stephen Farnum's[46] later adaptation of Watkins's work for all band instruments in 1954; and the last by Kenneth Gutsch,[47] a measure designed to evaluate instrumental music achievement from sight-reading rhythms only.

John Watkins prefaced his study with the observation that few, if any, reliable performance scales existed in music reading and even fewer existed for instrumental performance. He noted the "most carefully developed instrumental performance test [was] an organ scale by Stelzer."[48] This test by Stelzer used an objective scoring method, had high reliability, and had computed item difficulties. Watkins further stated, "this represents one of the few performance tests carefully constructed and validated by modern psychometric methods."[49] The purposes of Watkins's study were identified as follows:

[45] John Goodrich Watkins, Objective Measurement of Instrumental Performance, (New York: Teachers College, Columbia University, 1942).

[46] John G. Watkins and Stephen E. Farnum, The Watkins-Farnum Performance Scale for Band Instruments: Manual of Instructions, (Winona, Minnesota: Leonard Music Co., Inc., 1954).

[47] Kenneth Urial Gutsch, "Evaluation in Instrumental Music: An Individual Approach," Council of Research in Music Education, Nos. 5 and 6 (1964), pp. 21 - 28.

[48] Watkins, op. cit., p.9.

[49] Ibid.

1) to determine the possibility of measuring objectively achievement on a musical instrument;

2) to find out in a group of performers on this instrument the relation existing between sight-reading ability and technical skill after various periods of study.

Because of the lack of objectivity in measuring the inherent elements of sight-reading and technical skill, Watkins surmised that the only measurable element was that of performance. As a result, he substituted the terms "sight performance" and "practiced performance" for sight-reading and technical skill respectively.

Watkins chose to develop the performance scale for the cornet. He conducted an exhaustive procedure for developing and validating the test. Perhaps this systematic approach to measurement and the construction of measurement devices was the most significant contribution made by Watkins.

From an analysis of the method books available at that time, a series of sixty-eight graded exercises were selected. These exercises were administered to one-hundred-and-five students and from the data gathered, two equivalent forms of the scale, each containing fourteen exercises, were obtained. Based upon the preliminary study, the two forms had a correlation of $r = .982$. The students were rank-ordered with respect to their sight-reading ability by their respective teachers and the correlations of validity ranged from .66 to .91.

The final forms of the test were administered to cornet students of varying abilities by a group of music instructors. The sight-performance score was obtained from the initial reading of the test. The practiced performance score (measure of technical skill) was obtained by administering the same test again after the students had spent one week practicing the items. Data was obtained from one-hundred-fifty-three students who took Form A and seventy-one students who took Form B. The reliability coefficient between the two forms for the sight performance was r = .953 and for the practiced performance, r = .947.

Significant to the present study is the method of scoring and the elements that were scored. Watkins summarized the previous literature concerning scoring techniques and methods by analyzing the scoring methods of the tests conducted prior to his study. He concluded, as does Kwalwasser,[50] that the measure should "unquestionably" be the unit of scoring. Watkins stated:

> If the note is used as the scoring unit, certain difficulties become manifest. First, there is a tendency for any mistake to so upset the student that for the next few notes immediately following the error there is a greater chance for other errors. This tendency would not be so likely to cause additional errors if the measure were used as the scoring unit. Second, the note can be used as the scoring unit only on the assumption that it represents a specific response. A whole note sustained on a high pitch is obviously very different as a response, both qualitatively and quantitatively, from a single note in a sixteenth note run and on a low pitch. Here, the unit of scoring is uneven; no two units are the same in any sense.[51]

[50] Kwalwasser, Tests and Measurements, op. cit. p. 106.

[51] Watkins, op. cit., p. 43.

Watkins concluded that the weight of evidence favored the use of the measure as the scoring unit. This, of course, conflicts with Cooper's observations that unquestionably the unit of scoring should be the note. Accordingly, in his test, Watkins used the measure and scored the measure as incorrect if it contained one or more errors. The following were identified as errors:

1) pitch errors - a tone added or omitted or a note incorrectly played;

2) time errors - a note not held its correct duration;

3) change of tempo - more than twelve beats per minute variation;

4) expression errors - failure to observe any expressions mark;

5) slur errors - any slur omitted or a tongued note slurred;

6) rests - ignoring a rest or failing to sustain it its designated length of time as defined by the rules for rhythm errors;

7) holds or pauses;

8) repeats - failure to perform indicated repeats.[52]

This method of scoring was essentially the same as was used by Farnum in his later adaptation of Watkins's work.

In this subsequent work, Farnum adapted the Cornet Performance Scale for use with all band instruments. Essentially this adaptation was unchanged from the original, except for several transpositions and some additional

[52] *Ibid.*, pp. 44 - 47.

refinements. Farnum did determine reliability coefficients for all band students grades seven through twelve. The coefficients for the various groups are as follows: 1) grade seven - r = .87; 2) grades ten through twelve - r = .94; and 3) grades seven through twelve - r = .94. The validity of the band scale was established by using rank-order correlations. Each student was ranked by his instructor and the scores on the performance scale were correlated with the rankings. These correlations ranged from r = .68 to r = .87, having a mean correlation of r = .806. Farnum stated;

> The correlations compare favorable with those of Watkins. In view of these high correlations the scale may be used with a high level of confidence in measuring achievement in instrumental music.[53]

In a recent study, Kenneth Gutsch sought to determine 1) if an evaluative tool could be developed which would measure an individual's instrumental music achievement while sight-reading rhythms, and 2) whether or not such a tool, if it could be developed, could differentiate degrees of attainment for individuals who represented different amounts of instrumental music experience and reflected a variety of age levels.[54]

Gutsch chose to use Joseph Schillinger's system of

[53] Watkins and Farnum, _op. cit_.

[54] Gutsch, _op. cit_., p. 21.

rhythmical construction as the basis for developing the rhythmic items for the test. Through a complicated system of graphing two generators and determining the resultant, a translation was made of the resultant into corresponding musical notation. A series of nineteen resultants were identified which eventuated into three-hundred rhythmical problems. These were divided into two equivalent forms, A and B, and administered in a pre-testing period. From the results of the pretest, each form was reduced to one-hundred patterns. The administration procedure and the population were determined and the necessary details were attended to.

The subjects participating in the study consisted of 771 instrumental students in the public schools of Mississippi. Each student was above the fifth grade level and all had received at least six weeks of instrumental training. Of the 771 subjects individually tested, 336 were given Form A first and 405 were given Form B first.[55] When Form A was administered first, the correlation between the two forms of the test was r = .968. When Form B was administered first, the correlation coefficient was r = .964. The combined correlation coefficient was found to be r = .9659.

The test administrators and scorers were teachers

[55] Ibid., p. 26. A discrepancy is noted here since 336 and 405 do not add up to 771. Perhaps a typographical error occurred in the original manuscript and the figure should be a total of 741. This, however, is speculation on the part of this writer.

in the field. In order to establish scorer reliability, Gutsch selected a panel of four experts to score seventy-two randomly selected test sessions following the initial scoring by the teachers. The correlation coefficients between the original scores by the test administrators and the scores of any of the experts never fell below r = .99.

Gutsch also computed intercorrelations in order to determine if musical experience, grade level, scholastic average, or I. Q. had any effect on the scores. It was found that musical experience was the only factor that contributed significantly to scores on the test. When levels of experience were defined in terms of years, an analysis of the data (t-test) revealed differences in mean performance scores between any level and the level immediately preceeding it, significant at the .05 level of confidence. The writer had previously determined that "if the items of the test differentiated between groups of students with varying degrees of instrumental music experience, construct validity would be demonstrated."[56]

Gutsch concluded that the results reflected both test reliability and scorer reliability and when experience was defined in years, there was a significant difference between adjacent levels of instrumental experience. He further stated that experience was the most influential factor governing performance and that age had relatively little effect upon the student's ability to sight-read rhythms.

[56] Ibid., p. 25.

Summary

During the recent years many studies have been conducted regarding the subject of sight-singing and sight-reading. For the purposes of this study, these were grouped into four categories: 1) studies that involved factors which influence sight-singing skill; 2) group sight-singing tests; 3) studies whose emphasis was on some specific aspect of sight-singing; and 4) studies involving the construction and validation of sight-singing or sight-reading tests.

Four studies were reviewed as to factors that influence sight-singing skill. The studies by Salisbury and Smith, Dean, and Ottman revealed that tonal and melodic abilities correlated most highly with sight-singing ability. The fourth study, by Read, was a duplication of the Ottman study.

Four studies were reviewed that contained or involved group sight-singing tests. In addition, several tests using similar techniques were also reviewed. Early studies by Mosher, Hutchinson, and Knuth purported that ability to sight-sing can be measured through group methods, although Mosher recognized that this was limited. A later study by White was more definitive in verifying these results due to the rigorous techniques used in his study. A fifth group test, by Thostenson, was analyzed in conjunction with an individual sight-singing test by the same author. Other tests using similar techniques included: 1) the Kwalwasser-Ruch Test of Musical Accomplishment; 2) the

Diagnostic Test of Achievement in Music; 3) the Farnum Music Notation Test; and 4) the Jones Music Recognition Test. Questions were raised as to the validity of measuring sight-singing when no actual vocal response was required. It was concluded that these devices are not sight-singing tests, but are more appropriately referred to as melodic discrimination or tonal imagery tests.

Studies concerning the use of tachistoscopic methods by Barnes, Hammer, and Ray were analyzed. Barnes found that the ability to sight-sing isolated intervals correlated significantly with the ability to sight-sing melodically. Marquis found that the ability to sight-sing intervals changed as the context in which those intervals occurred changed. Danfelt determined that the type of material used in sight-singing exercises, contrived or composed, was not relevant to the improvement of the student's ability to sight-sing. Nelson found that short items were more reliable, valid, and better predictors of sight-singing ability than the longer traditional items. Finally, Ritchie discovered that the context of an exercise confounded the student's ability to take melodic dictation.

Several tests have been constructed specifically for the measurement of sight-singing or sight-reading. Two types of scoring systems are commonly used in these tests. The first included those tests that used the note as the unit of error. The tests that reflected this system included those by Hillbrand, Thostenson, and Cooper. The second included those tests that identify the measure as the unit of error. The tests that reflected this system included those by Otterstein and Mosher, the Watkins Cornet Performance Scale, the Farnum adaptation of Watkins's work for all band instruments, and the Gutsch rhythm test.

CHAPTER THREE

METHODOLOGY

Introduction

This study was conceived to address the problem of the minimal usability of sight-singing tests. Fifty-four subjects volunteered to participate in the study. These subjects were tested using four data-gathering instruments: 1) a subjective rating by the researcher; 2) the Otterstein-Mosher Sight-Singing Test; 3) an original unirhythmic pitch reproduction test; and 4) a musical experience questionnaire. After the students completed the questionnaire, the tests were individually administered in the above order. The tests were scored during administration and were tape-recorded for scoring at a later date by the researcher and by several other music instructors for scorer reliability. The reliability and validity of the unirhythmic test was determined through analysis of the test data and the correlation between the unirhythmic and traditional sight-singing tests.

Several pilot studies provided the foundation for designing the present study. Prior to a complete description of the methodology of the present study, the pilot study methods and data will be described and discussed.

Pilot Study One

As an outgrowth of a research seminar in evaluation and measurement, a sight-singing test was constructed similar to the traditional tests reviewed in Chapter Two. The test consisted of fifteen items, three from each of the five major style periods in music history, and ranging in difficulty from easy tc difficult. The items were selected from previously composed pieces written by well-known composers of their respective era. The test was individually administered and audio tape-recorded for scoring at a later time. The subjects were given a maximum of forty-five seconds to peruse each item. The fifteen subjects that participated were then asked to sing each item at a speed of approximately one beat per second. Following the completion of the tests, the responses of the students were scored by two persons, the researcher and another choral music education doctoral candidate.

A Pearson product-moment correlation coefficient was calculated to determine scorer reliability and _r_ was found to be .971. Kuder-Richardson Formula No. 20 (KR 20) was used to compute the reliability of the test and was found to be .954. As a result of these findings, the test was determined to possess high reliability for the items and the test, and high scorer reliability. Content validity was inherent because of the construction based upon actual choral literature. Criterion-related validity could be claimed because of the close correlation between the skills necessary for completing this test and those necessary for sight-singing in the choral setting. The results of this small study closely resemble the work of Cooper.[1]

Four conclusions were drawn from the above study:

1) the administration procedure was too lengthy to prove usable. The average administration time ranged from twenty to thirty minutes;

2) the scoring system was too complicated to be useful, and too much time was required to score each subject. The scoring time ranged from 1.5 hours to 2.5 hours;

3) lengthy items were not necessary to determine success in sight-singing. This finding agrees with Nelson's work.[2]

4) the difficulty of the items was more attributable to pitch content than to rhythmic content.

It was determined that these aspects of the study would need modification if the test were to be useful in the future.

Pilot Study Two

Using the above findings as a basis for further test construction, a method of testing sight-singing was devised to incorporate these techniques. The resultant test 1) could be administered and scored in less than five minutes; 2) scored each note as correct or incorrect,

[1] Cooper, "The Development of a Sight-Singing...," op. cit.

[2] Nelson, "A Comparison of Two Methods...," op. cit.

rather than scoring the relative accuracy of the interval; 3) used between four and twelve notes per item; and 4) was entirely constructed of pitch sequences without rhythm. The first form of the test contained all thirteen intervals within the octave in equal frequency. This test was administered to twenty high school students and fifteen college students and found to be virtually unusable, particularly for high school students. The reason for this unusability was because the items selected were not representative of the average high school or college repertoire. In order to achieve equal frequency of the intervals, highly contemporary or atonal music had to be used. As a result, the voice leadings and leaps were not those normally found in traditional literature and hence substantially reduced the content validity and the accuracy of performance. Equal frequency of the intervals was subsequently abandoned.

This test was then revised using the following procedures to insure content validity. A list of standard repertoire selections was compiled from three sources: 1) the Michigan School Vocal Association state choral festival selections, 1974 - 1978; 2) a survey conducted among high school choral conductors in the state of Michigan; and 3) the choral library at Michigan State University. From these compositions, eighteen of the pieces found in all three sources were selected and the intervallic content was analyzed. A percentage frequency was calculated for each of the thirteen intervals by summing the number of times the interval occurred in the eighteen pieces and dividing by the total number of intervals. The actual frequency and the percentage, or proportional, frequency of the thirteen intervals are illustrated in Table 3-1. The actual number of each interval used in the final form of the test varied somewhat from the calculated frequencies for reasons to be explained later.

Once the relative frequencies of the intervals were ascertained, excerpts were selected that were representative of choral literature and reflected the tonal characteristics of traditional harmony. The total intervallic content of the test closely matched the expected frequencies with the exception of the unison. From the data in Table 3-1, the reader will notice that most choral pieces consist predominantly of unisons, minor and major seconds, and fewer minor and major thirds. Approximately 85% of all the intervals from these eighteen selections fall into one of these categories. It was determined that to construct a test that used all the intervals and provide definitive information concerning the singer's abilities to sing all of these intervals, it would be necessary to use each interval at least once, and more if possible. A quick perusal of Table 3-1 would show that a short test would virtually eliminate several of the intervals if the percentages were followed explicitly. With this in mind it was decided to use the percentages as a flexible

Table 3-1

Frequencies and Proportions for the Intervals from the Eighteen Choral Selections

Interval	Frequency	Proportion
Unison	3866	.248
Minor second	2452	.157
Major second	5057	.324
Minor third	1491	.096
Major third	789	.051
Perfect fourth	1041	.067
Tritone	58	.004
Perfect fifth	483	.031
Minor sixth	64	.004
Major sixth	118	.008
Minor seventh	29	.002
Major seventh	10	.001
Perfect eighth	137	.009
Totals	15595	1.000

foundation for the test. Several factors were considered which resulted in minor alterations from the expected frequencies. First, a test consisting of one-fourth unisons would not discriminate between singers, regardless of their ability, since most can perform them with ease. Therefore the excess unisons were distributed throughout the other intervals. Second, to remain consistently within the tonal tertial framework, additional thirds and fifths would be used. The result would be exercises that revolved around the basic tonic and dominant sonorities. Third, all intervals, with the exception of the major and minor seventh and the tritone, would be used at least three times. The above three intervals were not used the same number of times for two reasons: The percentages would not allow it, and the context in which these intervals are found does not vary. The result would be that each time the interval was presented the same skill would be measured.

Using the above information and decisions, twenty melodic tonal exercises were selected from the body of choral excerpts. These twenty exercises were selected so that the frequency of each of the intervals basically reflected the expected frequencies of Table 3-1, with minor alterations. The actual frequencies and proportions of the intervals in the final form of the test are illustrated in Table 3-2.

Table 3-2

Actual Frequencies and Proportions for the Intervals Used in the Unirhythmic Test

Interval	Frequency	Proportion
Unison	5	.052
Minor second	14	.144
Major second	24	.247
Minor third	12	.124
Major third	8	.082
Perfect fourth	7	.072
Tritone	2	.021
Perfect fifth	12	.124
Minor sixth	4	.041
Major sixth	3	.031
Minor seventh	2	.021
Major seventh	1	.010
Perfect eighth	3	.031
Totals	97	1.000

Once the test was constructed, a panel of experts was selected consisting of choral conductors at Michigan State University, high school choral conductors, and choral education doctoral candidates. Each was asked to determine the appropriateness of the exercises and the test and to rank the exercises according to difficulty. The mean ranking for each exercise was used to derive the final order for the test to be used in the pilot study.

The students that comprised the sample to be tested in the pilot study were those attending Michigan State University Summer Youth Music. A total of one-hundred-and ninety-one students were tested. Four additional music education doctoral candidates were solicited to assist in administering the test. The students were tested individually on the first day of both summer sessions; approximately ninety the first, and one-hundred the second. Prior to being tested, each student was asked to complete a musical experience questionnaire relating to their vocal, instrumental, keyboard, family, and Youth Music experiences. A copy of the questionnaire, the test, and the administration instructions can be found in appendices A, B, and C respectively. The following instructions were given to the students by the administrator following an explanation of the test.

1) sing each exercise on a neutral syllable (recommend "lah");

2) try not to sing the notes too fast, perhaps one note per second;

3) sing each excerpt straight through without stopping;

4) you are not being graded on your tone, so do not worry about how you think you sound. Try to do your best.

The test administrator was to establish the tonality and starting pitch or simply indicate the starting pitch with no tonality prior to each exercise. The administrators were instructed to alternate between tonality and pitch, and pitch only for every other student. The result was to have two randomly assigned groups, one receiving the starting pitch only, and the other receiving both tonality and starting pitch. Specific instructions were also given in the event of a breakdown on the part of the student.

The reader will recall that the scoring procedure that allowed immediate scoring used the measure as the unit of error. If any mistake was made in the measure, the measure was marked wrong. In the unirhythmic test, there were no measures. Conceptually, each note could be regarded as a separate entity. Since pitch was the only consideration, any mistake in pitch would imply that the note was incorrectly sung and hence marked wrong. With the student singing at a speed of one note per second, the scorer should have sufficient time to determine the correctness of each note. It was not necessary for the scorer to make a decision concerning the correctness of the intervals. An intervallic score would result using the decision rule devised for that purpose. A scoring sheet was constructed that would allow the scorer to record a correct or incorrect response for each note.

As the scorer listened to the tape-recording of each student's response, he was to indicate a plus (+) for each correct response and a minus (-) for each incorrect response. Since a new tonality was established every few notes, the scorer should be able to retain the tonality of an exercise without great difficulty. Not all students have the ability to remain in the same key throughout an exercise without accompaniment. This, unfortunately, is an inherent problem with sight-singing tests. There is no easy solution to this problem, although Danfelt[3] elected to use a stroboscopic reading of ± 50 cents on a equal temperament scale as a correct response. While this might remedy the problem of tuning, it also would preclude rapid scoring of any test. In actual practice, the situation only occurred several times where the student was continually transposing an exercise due to successive sharpness or flatness. When this occurred, the piano provided an excellent permanent pitch reference.

The time needed to become relatively comfortable with the scoring procedure and allow for immediate scoring was found to be approximately forty-five minutes to one

[3] Danfelt, "An Experimental Study...," op. cit.

and a half hours. It should be noted that in the pilot study the researcher did all of the scoring and rescoring for reliability. Initially, it was necessary to play back several exercises and relisten to ensure an accurate score. Following this initial period, it was only necessary to play back and relisten on rare occasions.

The expressed purposes of this pilot study were:

1) to establish the reliability of the test, the individual exercises, and the interval types as items;

2) to determine the correlation between the exercises and the entire test, the interval types and the entire test, and the previous experiences with success on the test;

3) to determine the practical order of difficulty for the exercises of the test.

To accomplish the above, the following statistical procedures were used:

1) Cronbach's alpha was used to establish the reliability of the test, the exercises and the intervals;

2) the Pearson product-moment correlation coefficient was used to determine the relationship between the exercises and the test, the intervals and the test, and the musical experiences and the test;

3) the central tendency data for the items and the intervals were determined using SPSS Condescriptive procedures.

Following the recording of the data on the score sheets, they were transferred to computer cards by the researcher. The data for this pilot study were processed and analyzed by the computer system located at Michigan

State University. The SPSS (Statistical Package for the Social Sciences) programs used to accomplish this task were listed above. While the scoring was done from note to note, the data were analyzed according to intervals. Table 3-3 illustrates the aforementioned decision rule used to determine the correctness of each interval.

Table 3-3

Decision Rule for Interval Accuracy

(IF) 1st note	(AND) 2nd note	(THEN) interval accuracy
* +	+	+
+	** -	-
-	+	-
-	-	-

* (+) denotes correct response.
** (-) denotes incorrect response.

The question immediately arises concerning the possibility of singing two consecutively incorrect notes and yet singing a correct relative interval. This possibility certainly exists and does occur. There were two reasons for the above decision. First, a correct interval at the wrong pitch level is not an acceptable response in any choral or vocal setting, except when an entire ensemble would go flat while singing. If this response is not acceptable and is discouraged in the actual classroom, yet allowed during the test, content validity is reduced. Second, and more practically, the ease of scoring would be virtually nullified if the scorer were required to concentrate on relative interval accuracy. Were the primary consideration the ability of the student to sing intervals in a tonal context, this would be a necessity. The intent, however, is to determine if this method is reliable for measuring sight-singing.

Analysis and Discussion of the Results

Although the stated purposes of the pilot study were rather broad, they were similar to those of the present study. To avoid duplicating the discussion of similar data on the same procedures, only reliability, exercise difficulty and the order of the exercises, and the t-test regarding presence of tonality versus starting pitch only will be discussed. The following remaining tables will be found in Appendix F: 1) the correlation between the exercises and the test; 2) the correlations between the experience variables and the test; 3) the interval difficulty and rankings; 4) the correlations between the interval types and the test; and 5) the reliability of the interval types as items.

The initial purpose of this pilot study was to determine the reliability of the exercises and of the entire test. The reliability coefficients, as calculated using Cronbach's alpha, are illustrated in Table 3-4.

Table 3-4

Reliability Coefficients for the Exercises and the Test

Exercise	Alpha
Entire test	.969
1	.919
2	.804
3	.784
4	.531
5	.877
6	.832
7	.743
8	.803
9	.771
10	.817
11	.904
12	.855
13	.883
14	.782
15	.773
16	.645
17	.774
18	.821
19	.781
20	.844

The data of Table 3-4 indicated a high reliability for the entire test and relatively high reliability coefficients for most of the exercises. With the exception of exercises 4 and 16, all exhibited a coefficient in excess of .72. Many of the studies reviewed in Chapter Two reported reliability coefficients ranging from .62 to .95. It would appear that the coefficients identified in Table 3-4 fall well within that range. The reliability coefficient for the entire test was as high, if not higher, than most of the studies reviewed.

The reliability data reported here indicated that the test satisfactorily measures some aspect of sight-singing. The question that immediately arises concerns the nature of this aspect of sight-singing. The assumption of this writer is that the test accurately measures the pitch element of sight-singing. This assumption may even be incorrect depending upon the confounding and compounding influence of the addition of rhythm to a pitch exercise. For the time being, however, the assumption will be that the test accurately measures some aspect of sight-singing. No inferences can be made to the broad competency known as sight-singing.

Another aspect of the pilot study to be considered was the difficulty rating of the exercises and the resultant order of the exercises based upon the difficulty data. For the pilot study, the difficulty rating was based upon the percentage or proportion of correctness for all the students on each exercise. The reader will recall that the order of the twenty exercises was determined by the pooled judgment of the panel of experts. As there was not total agreement concerning the difficulty of the exercises, the mean rating was calculated. It follows that the actual data might differ somewhat from the opinions of the experts. Using the proportion of correctness on each exercise, a new order of difficulty was obtained. The difficulty rating, mean, standard deviation, and new order for the exercises are found in Table 3-5.

The proportional means for the 191 students seem rather low. These 191 students, however, were simply high school music students, not necessarily singers. A sizeable majority of the subjects had never sung in an organized choral ensemble. Ironically, some of the best sight-singers fell into this category. In this sample there were students who did not correctly sing any of the pitches, much less any intervals. These students were predominantly instrumentalists who had never sung prior to this experience. The statistics for the subset of the sample having one year of more of choral experience were substantially higher. This is not of great importance except to those who might consider the test too difficult and non-discriminating because it did not register high scores as well as low.

The final calculation was a t-test of the differences in the means of the two groups; one receiving just the

Table 3-5

Mean, Standard Deviation, and Difficulty Ranking for the Items of the Unirhythmic Test

Exercise	Mean	Standard Deviation	New Rank
Entire test	.278	.203	
1	.635	.407	1
2	.590	.370	2
3	.391	.316	3
4	.353	.325	6
5	.360	.407	5
6	.302	.372	9
7	.332	.377	7
8	.286	.270	10
9	.267	.309	11
10	.382	.311	4
11	.185	.327	15
12	.207	.331	13
13	.146	.279	19
14	.183	.272	16
15	.099	.214	20
16	.325	.291	8
17	.155	.248	18
18	.181	.251	17
19	.216	.263	12
20	.196	.277	14

starting pitch and the other both the tonality (tonic chord) and the starting pitch. The t-test was calculated for the total score and for the exercises between the two groups. Of the individual exercises, seven were found to differ significantly at the .05 level of confidence or less. The group receiving both tonic chord and starting pitch achieved substantially higher scores. For the remainder of the exercises, with the exception of three, the same trend was noted, although these were significant at a probability exceeding the .05 level of confidence. The three exceptions showed a reverse trend, although the differences were not significant at the .05 level.

The mean for the group receiving the starting pitch only on the total test was .243 (decimal or proportion representation) with a standard deviation of .175 and a standard error of .018. The mean for the other group was .315, a standard deviation of .224, and a standard error of .023. The calculated t-value was 2.47, having a two-tailed probability of .014. The results imply that hearing the tonality prior to the performance of an exercise substantially improved the performance of the exercise. One would naturally expect this to be true. Perhaps the most unique aspect was the three exercises whose performance was negatively affected by the presence of the tonality.

Conclusions from the Pilot Study

From the data gathered in this pilot study, the following conclusions were made:

1) the unirhythmic sight-singing test constructed for this study demonstrated high reliability for the test as a whole and for the individual exercises;

2) the individual items show a relatively high correlation with the total score indicating the exercises are reasonably accurate in predicting success on this test and that they discriminate between various levels of sight-singing ability;

3) there is a high correlation between the ability to perform specific intervals and success on this test;

4) there is no one experience that greatly attributes to success on this test;

5) the presence of a tonal framework substantially increases the probability of success on this test.

The above conclusions support the contention that the test is highly reliable, one of the prime considerations when devising a test to measure any facet of music aptitude, achievement, or performance. The second consideration of importance is of validity. The only type of validity that is conclusively present in this study is content validity. The method of constructing the test insured that it had content validity. It might be possible to construe criterion-related validity in the manner used in some of the studies reviewed in Chapter Two. It would seem, however, that the best method of determining criterion-related validity would be to compare the performance on the unirhythmic test with the performance on a standardized traditional sight-singing test.

The concept of construct validity is somewhat more difficult to assert and verify. There is no question of the inherent construct within the concept of traditional sight-singing. The question arises as to the possibility of isolating a construct that accurately encompasses the ability to sight-sing without rhythm. The process of such an endeavor would need to be exhaustive, were it at all possible. Construct validity is not claimed in this pilot study, nor is any attempt made to circumvent its discussion. For the present, the unirhythmic test is reliable and has content validity for the aforementioned sample.

The final consideration of importance is that of usability. As indicated in Chapter Two, many of the traditional sight-singing tests exhibit reliability and validity, yet are not usable due to the difficulty of administration and the length of time needed to obtain the results. It was found that both the researcher and the four additional administrators were able to administer the unirhythmic test in less than five minutes. Although the scoring was not done during the administration, the researcher was able, after several initial scorings of a few subjects, to score each student listening to the tape-recording without playing back any exercises. The implication would be that after becoming familiar with the test, the teacher should be able to score the subjects during administration.

The unirhythmic test used in this study was found to be reliable, possess content validity, and was quite usable.

The Present Study

The results of the pilot study indicated that the unirhythmic test was highly reliable and capable of measuring some aspect of sight-singing, presumably the pitch element. The present study was conceived to establish criterion-related validity for the unirhythmic test, applicable to the sample studied. To accomplish the task, it was determined that criterion-related validity would best be established by a comparison between a standardized traditional test and the unirhythmic test. In addition, a subjective rating by the researcher was assessed each student to compare with the results of the other two measurement devices. At the time of the testing, a musical experience questionnaire was completed by each student for correlation with test performance.

Samples Studied

A total of fifty-four students volunteered to participate in the study. The subjects were solicited from three distinct samples: 1) twenty high school singers attending Michigan State University 1979 Summer Youth Music; 2) twenty-four students attending either John Wesley College of Jordan College, two small private religious institutions in the state of Michigan; and 3) ten students attending Michigan State University during the summer of 1979. The scholastic experience of the subjects ranged from freshmen in high school to graduate university students.

The question could arise concerning the diversity of the samples; perhaps they were not drawn from the same population. In many studies this might pose several insurmountable problems. In the present study, however, this is not a significant consideration as each subject is his/her own control and the point of concern is the similarity of the student's response on all tests being administered. For the ANOVAS calculated to determine the differences between the three groups, the diversity of the samples is a significant consideration as individuals are not being compared with themselves, but groups with other groups. In this instance, however, the Central Limit Theorem would apply. This theorem is true because the population from which this sample was drawn has a finite variance and mean with regard to sight-singing performance on any given test, and the sample size was large enough to approach a normal distribution.

The assumption of independence, imperative when using analysis of variance, is also claimed in this study. While the subjects were not randomly selected, the observations were independent of each other; that is, the responses of any student had no relation to the responses of any other student. The third assumption for ANOVA, equality of variance, will be discussed with each appropriate ANOVA discussion.

Measurement Devices

As indicated, four measurement devices were used to gather data for the present study: 1) a musical experience questionnaire; 2) a subjective rating; 3) the O-M sight-singing test; and 4) the unirhythmic test. The musical experience questionnaire was completed by each of the students, identifying the amount of experience in the following areas: 1) scholastic level; 2) environmental activities - school, home, church, and private instruction; 3) choral and vocal experience; 4) instrumental experience; 5) keyboard studies; 6) additional studies; 7) family experiences - other family members who participate in musical activities; and 8) Youth Music experience - high school students only. Two forms of the questionnaire were used; one for high school and the other for college (See Appendix A).

The first of the actual tests administered was the subjective rating. Each choral director has his/her own individual approach to assessing the sight-singing abilities of the students who enter their choirs. There are several approaches known to the researcher: 1) the student may be asked to sing a series of contrived exercises designed to tax their ability in a short time; 2) the student may be asked to sing the alto, tenor or bass part to a song already familiar to the student; or 3) the student may be asked to sing their vocal part to a song the choir might be currently singing or one of the conductor's choice. There are certainly other techniques used by choral director in the field to assess sight-singing ability, however, these three are commonly used. The third method, having the student sing their vocal part, was selected for use in the present study. Several selections were chosen from the choral repertoire list found in Appendix D to be used as pieces from which one would be sung by each student. The selections used in the subjective rating are found in Table 3-6.

Table 3-6

Selections Used for the Subjective Rating

Selection	Composer
Ascendit Deus	Gallus
Best of Rooms	Thompson
Cry Aloud	Beck
Das Lamm	J. S. Bach
Exultate Deo	Scarlatti
How Beautiful Upon the Mountains	Berger
O Cast Me Not Away	Brahms
Sing and Be Joyful	Graun
Song of Exultation	Beck
Super Flumina Babylonis	Palestrina

The students were asked to peruse the selections and identify those unknown to them. After one was selected, the student was requested to sing their vocal part on a neutral syllable ("lah") at a relatively slow tempo, while the investigator played the accompaniment on the piano.

A suitable portion of each selection was sung, sufficient to determine the ability of the singer. At the end of this segment of the testing, a numerical score or rating was assessed, ranging from one to ten. An attempt was made to rate all students on the same scale, rather than one for each of the groups.

The second of the tests used in the study was the Otterstein-Mosher sight-singing test. After careful study of the available traditional sight-singing tests, this one was selected for the following reasons:

1) the test statistics indicated high reliability;

2) the scoring system was consistent, definitive, and relatively simple;

3) it contained parallel measures;

4) the test demonstrated as great a validity as any of the others reviewed by the researcher.

Statement four above is the most tenuous of the rationale given and requires the greatest defense. The test authors state:

> The validity of a test is of prime importance. If the test tests what it sets out to test, it is a useful measuring instrument. The aim here has been therefore, to construct and instrument which in every possible way resembles the function in question. Thus the examinees sing here a series of exercises typical of a music class; the exercises severally correspond to those which might be involved in a singing lesson, and the vocal utterances are graded in accordance with the plan described in another section. Since the testing conditions resemble so closely the ordinary procedure of a music class, the validity is quite apparent.

[4] Otterstein and Mosher, <u>O-M Test...</u>, <u>op. cit</u>.

While this method of validation is somewhat suspect, a close perusal of sight-singing texts and workbooks revealed a striking similarity to the items of the O-M test. Evidently, many authorities in the field must consider this type of exercise appropriate for the teaching of sight-singing.

The reader is referred to the complete discussion of the Otterstein-Mosher test found in Chapter Two. The twenty-eight exercises were found to be highly reliable, r = .935 for rhythm and r = .979 for pitch, using the odd versus even exercises. The scoring procedure used the measure as the unit of error, indicating a minus (-) above the measure if an error occurred in pitch, and below the measure if an error occurred in rhythm. The only considerations were pitch and rhythm and the total number of points was 448, 224 for rhythm and 224 for pitch.

The final test in the battery of tests was the unirhythmic test described earlier. The discussion of the construction and methodology of this test is found on pages 66 - 77 of this chapter under the heading "Pilot Study Two." Several modifications were made from the data in the pilot study, however, that should be enumerated. First, in the original form of the test, exercises which needed no accidentals were given no key signature, regardless of the tonal framework. For example, exercise ten was definitely in the key of D major, however, since none of the notes that would be altered by the key signature were used, no key signature was indicated. A similar situation occurred with other items, particularly those that were minor or modal. While this did not seem to have an effect on the majority of the students involved in the pilot study, several of the better sight-singers were puzzled and somewhat confused by their absence.

Second, the information gathered from the pilot study indicated that the order was not the correct order of difficulty. Therefore, it was altered to reflect the data of the pilot study. The reader may refer to Table 3-5 to recall the alterations.

Third, while several of the exercises were extracted from a piece in a particular mode or key, the tonal framework of the exercise more closely resembled a different mode or key. For this reason it was determined to give as the tonic triad the one which most closely described the tonal framework of the exercise.

Finally, the data from the pilot study indicated that students who were given both the tonic chord and the starting pitch performed significantly better than their counterparts. Therefore, it was determined to give both to all subjects in the present study.

Test Administration

The students were individually tested by the researcher at one of the three locations identified earlier. At Michigan State University, an office with keyboard was provided and all high school and university students were tested at that location. At Jordan College, twelve students were tested in the college chapel. Because construction work was being done at Jordan, some distractions were apparent. These distractions, however, did not seem to adversely affect the performances of the students. At John Wesley College, all testing of the balance of the college students was conducted in an office provided for the researcher.

On the days when the testing was conducted, the time was partitioned into half-hour time periods. This was sufficient time for the students to complete the questionnaire and the three tests. The students had already signed up for the time period most convenient for them and the majority arrived for their scheduled appointment. For those that forgot or were late, additional time was provided or another time scheduled. Upon arrival at the test site, each student was asked to complete the questionnaire.

After the questionnaire was completed, the student was asked to peruse the selections used for the subjective rating. When one was selected, previously unknown to the student, the procedure was explained and the tonality and starting pitch were given. In most cases, the students performed a sizeable majority of the selection. In several instances, however, this was not sufficient to obtain a satisfactory rating. In such instances, the students performed the entire selection. The reader is referred to Table 3-6 to review the compositions selected for the subjective rating.

Following this phase of the testing, the O-M test was placed before the student and explained. The following directions were given to each student:

> 1) The purpose of this test is to determine your ability to sing at sight. You will be graded on two points-pitch and rhythm. Tone quality does not matter; so if the note seems a little high when taking the test, and the quality of the tone is not pleasing, remember that does not affect your score. Do the best you can.
>
> 2) Set your own tempo. Perhaps a tempo of one note per second would be appropriate. For the items containing whole notes and half notes, perhaps a faster tempo would be appropriate, however, when the note values become faster you may want to slow the tempo down.
>
> 3) In each exercise the key will be established for you. I will play the tonic chord like this, (play arpeggiated major triad) and follow that with the note you begin on. (demonstrate using sample item)
>
> 4) You may sing using any method you wish. If you are not familiar with the numbers or the syllables you might like to sing on "lah."
>
> 5) Remember that all rests and final notes have a specific duration.

After the directions were given, the student was asked if there were any questions, after which the testing began. The test was sung in its entirety for all subjects except one. This subject was stopped prior to the last page because he scored only one correct measure, a whole note, out of the ninety-six possible measures. It was not deemed necessary for him to continue.

Several problems were encountered because of the incomplete directions for the O-M test. As a result, the following decisions were made to rectify the problems:

1) the singer might pause for an instant, then continue from the point of the pause. If this occurred, an error was recorded for the rhythm in the measure where the pause occurred;

2) the singer might stop, ask if they could start again, no pitch or rhythm error having occurred. In this case, they were permitted to do so, again recording an error in rhythm for the measure in which the stop occurred. In the event that pitch or rhythm errors occurred on the repeat where before there had been none, no error was recorded since none had occurred on the initial reading;

3) the singer might make several pitch or rhythm errors in the measures, realize their error, stop, and ask to start over. In this event, they were again permitted to do so, again recording an error in rhythm for the stoppage. On the repeat, regardless if they sang the portion correct or incorrect, the item to the point of the stoppage stood as initially scored and sung.

The reader will recall from the survey of literature that the scoring procedure for the O-M test consisted of making judgments as to the correctness of the pitches and rhythms in each measure. If an error was made in pitch, a minus (-) was placed above the measure. If the same occurred in rhythm, a minus was placed below the measure. After completion of the test, each melodic exercise was given a pitch score and a rhythm score depending upon the number of measures correctly sung in the exercise. The pitch and rhythm scores were summed to obtain a total for each exercise, then again to obtain the grand total score.

The third measurement device was the unirhythmic test. The students were instructed according to the following directions:

1) These items are unirhythmic, that is, all the notes have the same duration. They are not whole notes, simply equal. A tempo of one note per second would probably be appropriate.

2) Sing each item straight through without stopping, keeping each note equal.

3) Before each exercise I will play the tonic chord, or key center, and the starting pitch. You should begin singing immediately after the starting pitch is given.

4) Remember tone is not important, so do not be concerned if you think your voice is not sounding good.

5) Try to do the best you can.

This test, like the others, was scored while listening to the student. For each note either a plus (+) or a minus (-) was indicated depending upon the correctness of the response. A note score or total was calculated as the summation of all the correct responses (+). An interval total was obtained by applying the aforementioned rule (Table 3-3) to each exercise and summing the interval scores for all twenty exercises. A complete description of the unirhythmic test can be found in the first section of this chapter under the subheading "Pilot Study Two."

Scorer Reliability

During each session the researcher scored the performances on all the tests. This was intended to determine the usability of the tests, particularly the unirhythmic test, with regard to the ease and accuracy of the scoring method. At a later date, approximately one week following the sessions, the tests were rescoıed by the researcher to determine the reliability of the scoring system and the consistency of the investigator using the system. This would verify whether the scores for the individual subjects remained stable across time, although the setting, environment and personal characteristics accompanying the second scoring were altered from the first. Every reseacher logically is subject to the same daily fluctuations that singers experience, but if the scores remain stable, or similar, using a Pearson correlation, then the scoring system and the investigator's scoring abilities would be reliable with respect to the specific sample.

In addition to the single scorer reliability coefficient calculated for the researcher, several other music teachers, previously unfamiliar with the study, the test, and the scoring procedure, were asked to score the responses of the students on the unirhythmic test. Four persons agreed to participate; two public school choral directors, one college choral conductor, and one college music professor. This was done to establish the reliability of the scoring system when used by persons other than the researcher. The coefficient obtained from the above procedure would represent the level of agreement or concordance between a cross-section of music educators on a group of students using the same evaluating procedure. A high correlation would indicate that the scoring system was reliable for all levels of teachers, high school and college, when used to evaluate this sample.

Statistical Procedures

As previously stated, the research questions to be answered in this study concerned the determination of the reliability and the validity of the unirhythmic test and the differences between the three groups in the sample. To accomplish the above, the following statistical computations were made:

1) the reliability coefficients for the exercises and the entire test;

2) the correlation between the unirhythmic test a and the standardized test;

3) the discrimination index for each exercise on the unirhythmic test and the correlation of each exercise with the entire test;

4) the reliability of the battery of measurement devices;

5) the coefficient of concordance between the multiple scorers of the unirhythmic test, and the correlation between the first and second scorings of the unirhythmic and O-M tests by the researcher for scorer reliability.

In addition, several other considerations were of secondary importance in determining the above:

1) the correct order of difficulty for the twenty exercises of the unirhythmic test;

2) an order of difficulty for the thirteen intervals as used in the context of this test;

3) the correlation between the total score on the unirhythmic test and each interval, identifying the intervals that best predict success on this test;

4) the correlation between the unirhythmic test and the variables of the music experience questionnaire, identifying the experiences that best predict success on the test;

5) an ANOVA computed between the scores of the three groups on the unirhythmic and the O-M tests, and between the musical experiences of the three groups.

The statistical calculations were computed at Michigan State University using SPSS (Statistical Package for the Social Sciences), version 7.0. The programs used to compute the above calculations included: 1) Cronbach's Alpha for all reliability coefficients; 2) Pearson product-moment correlation coefficient for all correlations; 3) one-way analysis of variance for group mean comparisons; 4) Condescriptives for central tendency statistics; 5) Kendall's Coefficient of Concordance (W) for reliability of multiple scorers and multiple measurement devices.

CHAPTER FOUR

ANALYSIS AND INTERPRETATION OF THE DATA

Introduction

The method for the pilot studies and the related data in those studies were reported in Chapter III. The method of the present study was also articulated and the various statistical procedures identified. The data were collected using four types of data-gathering instruments: 1) a musical experience questionnaire; 2) a subjective rating obtained from listening to the student's performance on a choral selection from standard repertoire; 3) a standardized traditional sight-singing test, the Otterstein-Mosher (O-M) test; and 4) an original unirhythmic test, consisting of tonal intervallic pitch patterns without varying rhythm. The data were gathered from three distinct samples, all subsets of the larger population of singers: 1) high school choral students attending Michigan State University Summer Youth Music; 2) students attending two small private religious colleges in the state of Michigan; and 3) choral students attending Michigan State University.

The Subjective Rating

The purpose of the subjective rating was to determine if one of the standard informal methods used by choral directors in the field would approximate the results of the traditional or unirhythmic sight-singing methods. After completing the music experience questionnaire, each student was asked to sing a portion of a selection from standard choral literature at sight. A numerical rating was given ranging from one to ten, based upon the performance of the student. In actuality, the maximum rating given was 9.0 and the minimum was 1.0. The mean rating for the fifty four subjects was 5.32 having a standard deviation of 2.24. The statistics for the groups are illustrated in Table 4-1 for the subjective rating.

Table 4-1

Mean and Standard Deviation for the Subjective Rating

Group	$\bar{X}$	SD	N (n)
High School	5.05	1.91	n = 20
Small College	4.63	2.23	n = 24
Large College	7.50	1.36	n = 10
Total Sample	5.32	2.24	N = 54

Assuming the above sample is normally distributed, the data revealed fairly large differences in the means of the three groups. Further discussion of the differences will be forthcoming following the presentation of the data for the O-M and unirhythmic tests.

The Otterstein-Mosher Sight-Singing Test

The O-M test was administered immediately following the subjective rating as the standardized traditional sight-singing test to be used as the criterion for comparison with the unirhythmic test. Three sets of scores were available from the O-M test: 1) the total score; 2) the pitch total score; and 3) the rhythmic total score. The test exhibited a wide range of scores from a maximum of 443 points out of a possible 448 to a minimum of 131 points. The pitch totals also demonstrated wide differences with scores ranging from 223 points to 45 points out of a possible 224. The rhythmic scores ranged from 86 to 222 of the possible 224 points. It should be noted that there was substantially greater variation in the pitch totals as opposed to the rhythm totals. Only five of the fifty-four subjects performed under fifty percent correct rhythmically, while nineteen performed under that level regarding pitch. This figure would seem to indicate that the pitch element of the O-M test discriminated more accurately than did the rhythm element.

The mean for the entire sample for the O-M test was 324.48, or as a proportion, .724 for the total score. The pitch total mean was 146.77 or .66, and the rhythm total mean was 177.71 or .79. The group statistics for the O-M test are exhibited in Table 4-2.

A study of the data reveals differences, or the lack of them, between the means of the three groups. A

Table 4-2

Mean, Proportion, and Standard Deviation for the O-M Test

Group	$\overline{X}$	Proportion	SD	N (n)
Grand Total Scores				
Entire Sample	324.48	.724	84.03	N = 54
High School	313.20	.699	85.57	n = 20
Small College	301.96	.674	80.28	n = 24
Large College	400.90	.895	38.96	n = 10
Rhythmic Total Scores				
Entire Sample	177.71	.790	37.82	N = 54
High School	178.00	.795	39.92	n = 20
Small College	166.52	.743	37.35	n = 24
Large College	204.00	.911	22.07	n = 10
Pitch Total Scores				
Entire Sample	146.77	.66	50.30	N = 54
High School	135.15	.603	50.51	n = 20
Small College	138.56	.619	47.18	n = 24
Large College	196.90	.879	19.60	n = 10

one-way analysis of variance (ANOVA) was calculated to determine the statistical significance of these differences. The summary for the ANOVA calculations is found in Table 4-3.

Table 4-3

Analysis of Variance Summary Between Groups
for the O-M Test

Sources of Variance	SS	df	MS	F
Between Groups	74034.33	2	37013.17	*6.276
Within Groups	300833.98	51	5898.71	
Totals	374868.31	53		

*Significant at the .01 level of confidence

The reader will recall in Chapter Three the sample was described and two of the assumptions of the one-way ANOVA were discussed; normality of the sample and independence of the observations. The third, equality of variance, was calculated for the above computations. Bartlett's test for homogeneity of variance was calculated and the probability for the rejection of the null hypothesis was found to be .059. Hence, the null hypothesis of no difference between the variances of the three groups was accepted. While one would not reject the null at the .05 level, its close proximity to that level might cause one to consider the findings with apprehension. As a result of the data illustrated in Table 4-3, the groups were found to vary significantly at the .01 level of confidence.

Three Scheffe comparisons were made to determine the nature of the differences. One complex and two pair comparisons were made: 1) between high school and combined college; 2) between high school and small college; and 5) between small college and large college. These calculations can be found in Table 4-4.

Table 4-4

Scheffe F' Statistics and Confidence Intervals for the Otterstein-Mosher Test

Comparison	*F'	**Difference of means	Confidence Interval
HS - College	.23	11.24	$-47.44 \le \psi_1 \le 69.92$
HS - SC	.68	17.86	$-36.76 \le \psi_2 \le 72.48$
SC - LC	***11.70	98.94	$25.99 \le \psi_3 \le 171.89$

*The F statistic for comparison at the .05 level is 6.36, and at the .01 level, 10.12.

**The difference between the means of the groups being compared.

***Significant at the .01 level of confidence.

The Scheffe comparisons revealed the majority of the differences between the groups existed in the Small College-Large College comparison.

The confidence intervals identified above indicate the range of differences in means that would still be significant at the above level. Intervals that include zero are non-significant. As can be seen, the only comparison that did not include zero was between small and large college. The differences identified above are in terms of points on the O-M test. The differences noted here were similar to those of the subjective rating. In the results of that test, the means reflected a similar trend, although an ANOVA was not calculated. This would seem to indicate that the results of both tests were reasonably consistent.

The Unirhythmic Test

The final test in the battery of measurement devices was the original unirhythmic test. The purpose of this test was to determine if this method of measuring sight-singing performance was consistent with and measured the same level of performance as the standardized test. A total of 117 notes were used in the twenty exercises, ranging from four to eight notes per exercise. The summation of the correctly sung notes comprised the note score of the unirhythmic test. The interval score was calculated in the same manner using the decision rule identified earlier (Table 3-3). A total of 97 intervals was possible. These two totals, note and interval, were identified for each subject and were used to compute the group statistics for the sample. It was found that the two totals demonstrated similar trends for all statistical procedures; reliability coefficients, Pearson r's, means and standard deviations, and ANOVAs. Despite these similar trends, the unirhythmic test was much more reliable, discriminating, and valid when scored by intervals. As a result, all further discussion will relate directly to the interval totals and scores. Table 4-5 illustrates the group statistics for the unirhythmic test.

Table 4-5

Mean, Proportion, and Standard Deviation for the Unirhythmic Test

Group	*$\bar{X}$	**Proportion	SD	N (n)
Entire Sample	55.33	.57	23.18	N = 54
High School	49.60	.51	21.82	n = 20
Small College	50.88	.53	2.23	n = 24
Large College	77.50	.80	13.18	n = 10

*These statistics were calculated based upon the initial scoring of the subjects. This was chosen in lieu of the second scoring because it was felt that group norms should be based upon the one most likely to occur in the scholastic setting when used by choral conductors.

**The means are illustrated in proportions to facilitate comparisons.

Substantial differences were again noted between the means of the groups identified in Table 4-5. A one-way ANOVA was calculated to determine the statistical significance of these differences. The ANOVA summary for the unirhythmic test is illustrated in Table 4-6.

Table 4-6

ANOVA Summary for the Unirhythmic Test

Sources of Variance	SS	df	MS	F
Between Groups	6072.05	2	3036.03	*6.912
Within Groups	22402.78	51	439.27	
Totals	28474.83	53		

*Significant at the .01 level of confidence.

Bartlett's test for homogeneity of variance was again calculated for the unirhythmic test and the probability for rejection of the null hypothesis of no difference was found to be .206. This finding was substantially stronger than that of the O-M test and would seem to imply that the three groups were from the same population. The data of Table 4-6 verifies the significant differences between the three groups allowing rejection of the null hypothesis of no difference at the .01 level of confidence. The Scheffe technique was again used to identify the sources of the differences.

The same three comparisons were made between the groups using the Sheffe technique. The F statistics and confidence intervals are illustrated in Table 4-7. The data revealed a similar tendency for the unirhythmic test as with the O-M test and the subjective rating. Again the major portion of the differences identified in the ANOVA

Table 4-7

Scheffe F' Statistics and Confidence Intervals for the Unirhythmic Test

Comparison	*F	Difference of means	**Confidence Interval
HS - College	2.37	9.11	$-6.49 \leq \psi_1 \leq 25.01$
HS - SC	.04	1.27	$-14.88 \leq \psi_2 \leq 17.42$
SC - LC	***11.35	26.63	$6.55 \leq \psi_3 \leq 46.71$

*The figure for comparison at the .05 level is 6.36; at the .01 level, 10.12.

**The differences of the means are expressed in terms of points on the unirhythmic test.

***Significant at the .01 level of confidence.

calculations existed in the comparison between the small and large college groups. The results of all three tests indicated that significant differences existed between the small and large college groups.

Discussion of the results

The preceeding data for both the O-M and the unirhythmic tests, as well as the informal data of the subjective rating, seem to indicate that significant differences existed between the small and large college groups, whereas no differences existed between either the high school and combined college groups or the high school and small college groups. Upon first consideration, these findings would seem to contradict what one would normally expect. The expectation would naturally be that no differences would exist in the sight-singing abilities of college students regardless of the institution attended. However, when one considers the greater selectivity available at the larger institution as opposed to the smaller, these findings seem more logical. At the larger institution, the majority of singers participating in college ensembles are students who have made singing a prime objective in their scholastic experience. In addition, they would generally be from the upper portion of the Gaussian curve with regard to ability. This would not necessarily be true at the smaller institution. One would assume a lesser number of students in the upper standard deviations as a result of the fewer number of students attending the smaller institutions. In order to maintain the same standards, the size of the ensembles would need to be substantially reduced. This does not generally occur. As a result, a broader range of abilities would be represented in the choral ensembles of smaller institutions.

An analysis of the experience of the subjects in the respective groups revealed no significant differences between the groups with regard to instrumental experience, significant differences between high school and both college groups concerning choral experience, and significant differences between all three comparisons with regard to private vocal instructions. The ANOVA and Scheffe tables verifying the above conclusions are found in Appendix G. The implication of these findings was that the college students had similar experiences, with the exception of private vocal instruction, yet their performance on all three tests was significantly different. The differences in the private vocal instruction would also seem to verify the premise that students participating in the choral ensembles of larger institutions are somewhat more intense concerning vocal study and singing, perhaps resulting in greater achievement.

The preceeding discussion was based on the three tests, identifying the group statistics, the group differences and the individual comparisons for each of the tests. The ensuing presentation of data and discussion centers around the unirhythmic test and the analysis of data specifically concerning reliability and validity.

Analysis of the Unirhythmic Test Data

One of the considerations for tests that purport progressive difficulty is the order of the individual exercises. As previously stated, the original order of the exercises was altered as a result of the data from the pilot study. The twenty exercises were rearranged according to that data and the resultant form was used in the present study. It would follow that a new order of the exercises could result from the data gathered in the present study, perhaps more appropriate as to difficulty level. Table 4-8 illustrates the means, standard deviations, and

Table 4-8

Mean, Standard Deviation and Difficulty Ranking for the Exercises

Exercise	$\overline{X}$	SD	*New Rank
1	.915	.232	1
2	.847	.272	2
3	.713	.361	3
4	.608	.330	10
5	.644	.447	8
6	.679	.414	7
7	.698	.411	6
8	.506	.359	14
9	.699	.377	5
10	.474	.272	16
11	.626	.337	9
12	.509	.279	13
13	.532	.360	12
14	.491	.373	15
15	.700	.398	4
16	.578	.360	11
17	.444	.402	17
18	.400	.390	18
19	.340	.431	20
20	.359	.365	19

*Based upon administration to 54 subjects.

the new difficulty ranks for the twenty exercises based upon the data gathered in the present study. The means are represented in proportions because the exercises did not contain an equal number of intervals. The difficulty ranking here was derived based upon the proportion of correctness for each exercise. The higher the proportion, or the easier the item; the smaller the rank (e.g. #1 was the easiest; #20 the hardest).

A review of Table 3-5 will reveal substantially lower means than in the present study. The reader will recall that in the pilot study the students were simply music students, not choral students. Some of the subjects in the pilot study had never sung before. Many were instrumentalists having never sung in a choral ensemble. As a result, the proportion of correctness for the exercises was substantially lower. Conversely, in the present study, all participants were not only singers, but many were college students; some graduates. The result would naturally be that the proportion of correctness would be higher.

The item difficulty and item discrimination indices were also calculated using the Net D formulas.[1] These were found by summing the scores of the upper and lower 27% - tails of the sample. The upper and lower tails from this

[1] John C. Marshall and Loyde W. Hales, Classroom Test Construction, (Reading, Massachusetts: Addison-Wesley Publishing Company, 1971), pp. 230 - 233.

sample consisted of 14.59 or 15 subjects. One minor alteration was necessary to use these indices. Since each exercise consisted of several intervals, each dichotomously scored, the average of the tails for each exercise was divided by the number of intervals per exercise. The resultant indices of item difficulty and item discrimination are illustrated in Table 4-9.

The results of the calculations revealed a mean difficulty of .5995 and a mean discrimination of .5675. Generally the item difficulty should center around .50 or halfway between a chance score and a perfect score. With a test of this nature, it would be difficult to determine a chance score, although it would certainly be above zero. The result would be a mean difficulty somewhat above .50. It would seem that a range of difficulties would be advisable, the easy items needed to discriminate among the poorer students and the difficult items to discriminate among the better ones. In addition, a motivational factor could be present providing the possibility of success for students of lesser ability or challenge for students of higher ability. Concerning item discrimination, Marshall and Hales state:

> The interpretation of net D is somewhat comparable to that of a coefficient of correlation. Whenever the value is negative, the item exhibits negative discrimination; thus, it reduces the discrimination of the test. When the value is less than .20, the discriminatory power is so small as to be considered negligible. Items with indices between .20 and .40 are of some

Table 4-9

Net D Item Difficulty and Discrimination

Exercise	*Difficulty	Discrimination
1	.87	.25
2	.79	.42
3	.67	.57
4	.63	.56
5	.62	.76
6	.66	.43
7	.65	.53
8	.55	.54
9	.61	.72
10	.46	.52
11	.66	.54
12	.55	.49
13	.55	.50
14	.59	.65
15	.64	.66
16	.63	.57
17	.51	.79
18	.50	.69
19	.41	.60
20	.44	.56

*Based on upper and lower tails of 15 subjects each.

> value in discriminating between examinees. Items with indices between ·40 and .60 are good discriminators. Those with indices above .60 are unusually good.[2]

With the exception of exercises 1, 2, 6, and 12, all items were found to have discrimination indices above .50. One would have to consider exercise 1 almost valueless, and exercises 2, 6, and 12 only marginally valuable, although these exercises may be those needed to discriminate between poorer students and provide the motivational factor.

Marshall and Hales indicate that the index of discrimination is the ability of an item to discriminate between the better and poorer students, as established by the total test score. Another method of establishing this ability is to determine the relation between the items of a test with the total score on the test across subjects. This was accomplished by computing a coefficient of correlation. A Pearson r was calculated to determine the exercises that best predicted success on the unirhythmic test, and to determine which exercises best discriminated between subjects using the total score as a criterion. It should be noted, when using correlation as an index of discrimination, that the ability of an exercise to discriminate increases as the difficulty of that exercises increases. This is true to a point where the ability of the exercise to discriminate again declines. As the exercise reaches the point where neither the poorer students nor

[2] Ibid., p. 232.

the better students perform particularly well. When this occurs, the relationship between success on the item and success on the test is not very high. Hence, the difficulty is greater and the discrimination index and correlation coefficient for the exercise are substantially lower. It should be noted that to a slight degree, the correlation between each exercise and the total score is a correlation of the exercise with itself, since each is a subset of the twenty exercises that comprise the entire test. This would result in a slight inflation of the coefficient, but not substantially. The correlations between the exercises and the total score are found in Table 4-10.

For the coefficients illustrated in Table 4-10, all but two were significant at the .001 level, and those at the .002 and .003 level. While these significance levels seem acceptable, they are somewhat misleading when one considers that this simply means significantly different from zero, or no correlation.

Reliability of the Unirhythmic Test

The reliability of the unirhythmic test was calculated using SPSS Subprogram - Reliability: Cronbach's Alpha. Several coefficients of reliability were computed using the data gathered in this study. A coefficient was calculated using each interval as an item resulting in 97 items for the total test. Another coefficient was computed using each exercise as an item resulting in twenty items. These coefficients were used to establish the reliability

Table 4-10

Pearson Correlation Between the Exercises and the Unirhythmic Total Score

Exercise	Correlation Coefficient
1	.400
2	.584
3	.524
4	.723
5	.704
6	.411
7	.612
8	.675
9	.716
10	.710
11	.700
12	.790
13	.570
14	.733
15	.720
16	.682
17	.817
18	.782
19	.727
20	.700

of the total test. In addition, reliability coefficients were computed for each exercise as a separate entity. Table 4-11 illustrates the various coefficients calculated for the unirhythmic test and exercises.

From the data exhibited in Table 4-11, it can be seen that the unirhythmic test is extremely reliable when computed using all 97 intervals as separate items. When using the twenty exercises, the coefficient for the entire test is somewhat lower, although still highly reliable. The reliability coefficients for the individual exercises are also generally high, particularly when one considers the number of intervals in each ranged from three to seven. One would conclude from the above data that the unirhythmic test and the exercises are highly reliable.

Scorer Reliability

An important factor in establishing the reliability and usability of a performance test is the consistency of results across time and scorers. This is generally described as scorer reliability. Scorer reliability was computed for this study in the following manner. First, the scores on any given day should correlate closely with replicative scorings at later dates. This would indicate that scores on the test were not subject, to any great extent, to daily fluctuations in the mood or temperment of the scorer, nor changes in the environment or the setting where the scoring took place. It would also insure the degree of consistency

Table 4-11

Reliability Coefficients for the Exercises

Exercise	# of items	Alpha	Rank
1	5	.888	5
2	4	.766	18
3	4	.861	8
4	6	.820	14
5	4	.947	1
6	3	.856	9
7	3	.879	7
8	3	.748	19
9	4	.842	11
10	7	.781	16
11	5	.792	15
12	6	.731	20
13	4	.829	13
14	6	.886	6
15	5	.924	3
16	5	.780	17
17	7	.929	2
18	5	.893	4
19	6	.847	10
20	5	.840	12
Total Interval Score	97	.974	
Total Exercises Score	20	.933	

in the scorer's ability to use the specific instrument. To accomplish the above, the investigator scored all subjects during the administration sessions, then subsequently one week later. A Pearson r was calculated to determine the consistency of the first and second scorings. The obtained coefficient was r = .999 indicating an extremely high reliability for the test when scored by the investigator.

Second, the test should provide for accurate ratings by different scorers on the same sample of subjects. A high reliability coefficient would imply that the test could be accurately scored by competent music teachers, producing similarly reliable results. To accomplish multiple scorer reliability, the following procedure was employed.

Four competent music instructors were solicited for participation in the study. All were educators in high schools or colleges in the state of Michigan; two public school choral teachers, one college choral director, and one college music professor. Each was asked to listen to the tape-recording of the performances and score them using the identical procedures as the investigator. Verbal instructions were given to the scorers as well as written instructions for decisions involving specific occurrences. Kendall's W, a measure of consistency or concordance, was employed to determine the level of consistency between all scorers. In addition, Cronbach's alpha was also used, inputting the scorers as the variables for which the reliability coefficient would be calculated. These coefficients of reliability are illustrated in Table 4-12.

Table 4-12

Multiple Scorer Reliability

Statistical Method	Coefficient
Kendall's W	.995
Cronbach's Alpha	.999

The coefficients identified above revealed extremely high multiple scorer reliability for the unirhythmic test. Two reasons could be postulated for the excessively high coefficients. First, there is very little guesswork involved in the scoring procedure. The pitches are either correct or incorrect. No judgment of relative intervallic accuracy is necessary. This results in eliminating much of the possibility of variance between scorers. The only aspect requiring judgment is in determining the amount of incorrectness for poorly tuned notes. The frequency of out-of-tuneness was not great for most singers, resulting in limiting the varying judgments to a large extent. In addition, when out-of-tuneness was prevalent, there were also frequent wrong notes. As a result of the decision rule, this rendered many of the intervals incorrect regardless if one or more of the scorers heard individual notes differently. For example, scorer 1 might score an exercise "++--+" while scorer 2 might score the same exercise "++-+-." In both instances the score would be one correct interval. While this did not occur often, it certainly would be possible and occasionally did occur.

Second, the reliability coefficients were calculated using the total scores of the students rather than the exercise scores. The result would be that fluctuations in the exercise scores might off-set each other, not significantly affecting the total scores.

A correlation coefficient was also calculated between each scorer and the investigator. These coefficients are illustrated in Table 4-13.

Table 4-13

Pearson r Between Each Scorer and the Investigator

Scorer	Correlation Coefficient
1	.993
2	.998
3	.995
4	.998

Scorer reliability was also established for the O-M sight-singing test in a similar manner. The subjects were scored as the test was administered and then again one week later. It was not considered necessary to establish multiple scorer reliability because the O-M test is an established standardized sight-singing test.

As identified earlier, three scores were generated by the O-M test: 1) a rhythmic score consisting of 224 possible points, one point per measure; 2) a pitch score consisting of the same; and 3) a total score consisting of the summation of the two previous scores. A Pearson r was calculated for the rhythmic, pitch, and grand total scores between the first and second scorings. The findings are illustrated in Table 4-14.

Table 4-14

Scorer Reliability for the Otterstein-Mosher Test

Totals	Correlation Coefficient
Rhythm total	.991
Pitch total	.999
Grand total	.998

The preceeding data seems to indicate that the Otterstein-Mosher sight-singing test is also extremely reliable regarding the scoring process and scoring consistency when scored by the investigator.

Validity of the Unirhythmic Test

The primary purpose of this study was to establish the reliability and validity of this method of measuring sight-singing for this sample of singers. The reliability was to be determined through statistical analysis of the unirhythmic test data and the consistency of scoring and the results identified by repeated scoring across time and multiple scoring of the subjects. The content validity was established through the method of construction of the test; drawing the exercises from standard choral literature. The criterion-related validity was to be established by comparing the results of the unirhythmic test with those of an established standardized sight-singing test, and with a more subjective approach to ascertaining sight-singing ability. The standardized test selected as the criterion was the Otterstein-Mosher Test, constructed in 1932. The validity coefficient was to be determined by correlating the scores on the three types of measures. A Pearson r was calculated between the total scores of the unirhythmic test and the rhythmic, pitch, and total scores of the O-M test. These correlations are found in Table 4-15.

The results indicated that the correlation between the rhythmic total score of the O-M test and the total score of the unirhythmic test was substantially lower than with either the pitch or grand total scores. Two reasons for this lower correlation exist. First, the unirhythmic

Table 4-15

Validity Correlations Between the Unirhythmic Test and the Otterstein-Mosher Test

Totals	*Correlation Coefficient
Rhythm total	.775
Pitch total	.965
Grand total	.926

*The coefficients are calculated here using the first scoring of the unirhythmic test, as it was the one most likely to occur in the classroom, and the second scoring of the O-M test, as it was the most accurate.

test and the pitch total scores measured pitch and rhythm separately, although theoretically the two are never mutually exclusive. Perhaps this coefficient would be a reasonably accurate representation of the interrelatedness of the two skills, pitch and rhythm. A second reason for the lower coefficient was that the rhythmic element of the O-M test did not discriminate as well as the pitch element. As indicated earlier, the range of the rhythmic scores was substantially narrower than the pitch scores. In general, subjects who scored poorly on the pitch element inevitably scored much higher on the rhythmic element; in some cases as much as forty and fifty percent higher. In most of the melodic exercises of the O-M test the rhythmic texture was not difficult for the average high school singer. As a result, the rhythmic component of the O-M test did not discriminate well between good and poor sight-singers and hence the lower correlation. It follows then that the pitch element of the O-M test was the most influential in discriminating between subjects.

The correlation coefficients between the unirhythmic and O-M test having been established, it remained to determine the relation between the subjective rating and the other tests. The correlation coefficients between the subjective rating and the unirhythmic and O-M tests are illustrated in Table 4-16.

Table 4-16

Pearson Correlation Between the Subjective Rating and the O-M and Unirhythmic Tests

Totals	Correlation Coefficient
Unirhythmic total	.922
O-M Grand total	.923

The above data reveals high correlations with both the unirhythmic test and the standardized test. The implication of these findings would be that this subjective method of assessing sight-singing was nearly as accurate as the unirhythmic method when used by the investigator on this sample of singers. While this method would not be totally accurate in measuring the precise differences among students, it would be reasonably accurate for obtaining a general index of sight-singing ability. In the test administration sessions, both the subjective rating and the unirhythmic test required approximately five minutes to administer, although the subjective rating was more difficult to score. The reason for this was that the researcher was responsible for maintaining a consistent scale for all subjects. This implied remembering the previous subjects and comparing each to the others. With the unirhythmic test, the pitches were either right or wrong.

The calculations in the previous tables were performed on the three tests in pair comparisons. A Kendall's W was computed to determine the consistency of the three ratings; the O-M test, the unirhythmic test, and the subjective rating. The findings are exhibited in Table 4-17.

Table 4-17

Kendall's W for All Tests

Group	W	Chi-Square	*Significance
Total Sample	.951	205.491	0
High School	.952	76.165	.0000
Small College	.952	95.167	.0000
Large College	.961	38.452	.0000

*the significance levels are exactly as identified by the computer. Theoretically, it is impossible to have a probability of zero, with or without decimals.

In addition to the Kendall's W, Cronbach's alpha was also computed to determine the reliability of the complete battery of measurement devices. The results of the calculations revealed a coefficient of .981. This figure indicates that the battery of tests was highly consistent in measuring the skill of sight-singing and that little of the variance was error variance. The implication of the coefficients, both Kendall's W and Cronbach's alpha, is that the three tests, to a high degree, were measuring the same skill.

The preceeding data concerning reliability and validity reveals that the unirhythmic test is highly reliable, possesses content validity, and correlates highly with both the Otterstein-Mosher sight-singing test and the subjective method of evaluating sight-singing ability. In addition, the battery of tests has been shown to be highly reliable for measuring the skill of sight-singing.

Several secondary questions were identified concerning the relationship of performing the various intervals with the unirhythmic test, the analysis of the intervals as used in this study, and the relationship of prior experience with performance on the unirhythmic test. The following data analysis will address these secondary questions.

The Intervals

The method of construction for this test took into consideration the relative frequency of the thirteen intervals most commonly experienced in choral music. As one of the secondary purposes of this study, an analysis of the data in terms of intervals includes the order of difficulty, the correlation between the intervals and success on the total test, and the reliability of the intervals as items. Table 4-18 represents the mean scores, standard deviations, number of presentations, and difficulty ranking of the intervals as used in this study. It must be remembered that the intervals were not presented an equal number of times, and therefore these statistics must be considered with caution. As any discussion of interval difficulty must include the context in which they are sung, the difficulty ranking here is only applicable to this specific test using this sample of singers.

A review of Table F-2 in the Appendix F reveals a close correlation with the difficulty ranking from the interval data of the pilot study. A Spearman Rank Order Correlation Coefficient was calculated between the pilot study and the present study orders of difficulty. The coefficient was found to be .971. The implication of this figure is that the intervallic difficulty was consistent across the two studies. This would not indicate that the intervals universally reflect this order of difficulty, but within the context of this test the intervallic difficulty remained fairly stable. This is particularly interesting when the sample of the two studies are reviewed. The

Table 4-18

Group Data for the Intervals in the Unirhythmic Test

Interval	$\bar{X}$	SD	Rank	*n
Unison	.911	.163	1	5
Minor second	.545	.291	7	14
Major second	.631	.233	4	24
Minor third	.549	.276	6	12
Major third	.586	.279	5	8
Perfect fourth	.648	.257	3	7
Tritone	.333	.400	11	2
Perfect fifth	.480	.286	8	12
Minor sixth	.463	.341	9	4
Major sixth	.340	.402	10	3
Minor seventh	.213	.331	12	2
Major seventh	.167	.376	13	1
Perfect octave	.861	.265	2	3

*number of times each interval is used in the unirhythmic test.

sample for the pilot study consisted of 191 high school music students, many of which were non-singers. The sample for the present study consisted of choral students in high school and college. One would logically assume that choral students would have a much better grasp of the traditionally more difficult intervals. The only characteristic that seemed to be different, however, was the percentage of correctness had increased by approximately the same degree for all intervals.

Further analysis was done to establish the discrimination indices for the interval types and the correlation coefficients between the individual interval types and the total interval score. As indicated earlier, these are two methods of determining the discriminatory powers of the scales being identified. These two types of indices are similar and were used to determine the interval types that would best predict success on the overall test. The discrimination indices, as computed using Net D, and the correlation coefficients are presented in Table 4-19, with all the correlation coefficients significant at the .001 level of confidence.

The data reveals generally high correlations and discrimination indices for the interval types. Four of the correlation coefficients were substantially lower than the others, two of these being the unison and the perfect octave. The reader will notice that these were identified as the easiest in the context of this test. The other two

Table 4-19

Pearson Correlation and Net D Discrimination Indices for the Interval Types of the Unirhythmic Test

Interval	Discrimination index	Pearson r
Unison	.187	.496
Minor second	.638	.935
Major second	.564	.955
Minor third	.627	.923
Major third	.647	.883
Perfect fourth	.509	.843
Tritone	.766	.785
Perfect fifth	.670	.951
Minor sixth	.683	.817
Major sixth	.789	.837
Minor seventh	.483	.623
Major seventh	.400	.406
Perfect octave	.422	.517

were the major and minor sevenths, identified as the most difficult intervals. This pattern was also evidenced in the discrimination indices, the same four intervals showing the lowest. This pattern is reflective of the curvilinear relationship of difficulty to discrimination, and, if discrimination is held constant, to a lesser degree, correlation. The reader will recall that discrimination tends to increase as difficulty approaches the .500 range, then declines as the difficulty approaches either extreme. A similar, though not identical, trend can be shown with correlation. The correlation of any exercise with the entire test is an index of the relationship of success on the exercise with success on the test. When the difficulty approaches either extreme, the implication is that all the subjects are performing the exercise with a higher or lower degree of accuracy. Hence, the relationship between success on the exercise and success on the total test is substantially reduced. The data of Table 4-19 reflects an excellent example of this phenomenon.

In order to establish the reliability of the test in terms of the consistency of response for all thirteen interval types and using each interval type as an item, Cronbach's alpha was again computed at two levels: 1) consistency of the interval types; and 2) consistency across the intervals under each type. Cronbach states, "the coefficient [alpha] depends on the number of observations entering the person's score,"[3] [i.e. the number of times each interval was presented]. Paraphrasing Cronbach, with the addition of observations, the sample of performance becomes more adequate. By making more observations of the same general sort, a better estimate of ability is obtained.[4] In addition to the advisability of more observations, a minimum of three is necessary for using Cronbach's alpha. Because of the above, reliability coefficients for three of the interval types were not available using Cronbach's alpha without inappropriate data manipulation. The tritone, major seventh, and minor seventh could not be computed using this method. The tritone and minor seventh, however, each were observed twice in the test. Therefore, a Pearson r would yield as close a coefficient as is possible. The major seventh was observed only once and no reliability coefficient was available. The reliability data are exhibited in Table 4-20.

The reliability coefficients tend to support those of the Pearson r in that the interval types that correlated highly with the total test also achieved high reliability coefficients, and conversely, those that had low correlation coefficients also exhibited low reliability.

[3] Lee J. Cronbach, Essentials of Psychological Testing, (New York: Harper & Row, Publishers, 1970), pp. 165 - 166.

[4] Ibid., p. 167.

Table 4-20

Reliability Coefficients for the Intervals Used in the Unirhythmic Test

Interval	Alpha	*n
Unison	.620	5
Minor second	.887	14
Major second	.882	24
Minor third	.842	12
Major third	.759	8
Perfect fourth	.630	7
**Tritone	.418	2
Perfect fifth	.850	12
Minor sixth	.500	4
Major sixth	.785	3
**Minor seventh	.302	2
Major seventh	***	1
Perfect octave	.504	3
All intervals (unison - perfect octave)	.939	

* number of observations for each interval.

** Pearson r rather than Cronbach's alpha.

*** not computable.

A review of Tables 2, 3, and 4 in Appendix F reveals a striking similarity to the interval data presented here. The difficulty rankings, Pearson correlation coefficients, and reliability coefficients exhibit strong similarities. This observation supports the premise of reliability by demonstrating consistent results across studies. These findings also tend to support those of Barnes's[5] which established a high correlation between the ability to sight-sing intervals and the ability to sight-sing melodically.

Experience

As previously indicated, a questionnaire was completed by all subjects identifying their prior musical experiences. The purpose of this was to determine the relationship between success on this test and the various musical experiences available to high school and college students. Data were gathered concerning the following: 1) scholastic level; 2) environmental activities - school, home, church, and private instruction; 3) choral and vocal experience; 4) instrumental experience; 5) keyboard experience; 6) additional studies - theory, history, and all other formal music studies; 7) family experience - other family members who participate in musical activities; and 8) Youth Music experience - high school students only. The reader will recall from the review of literature the

[5] Barnes, "An Experimental Study of Interval Drill...," *op. cit*.

attempt by Read[6] to isolate the factors that produced the greatest likelihood for success in sight-singing, as well as the study by Ottman.[7] To determine the experience factors that significantly contributed to success on this test, a Pearson r was calculated between each of the experience variables and the total score on the test. These calculations are exhibited in Table 4-21.

The correlation coefficients for the experience variables verify the premise that no one experience contributes to success on this test. In addition, they seem to verify the fact that the best sight-singers possess a variety of experiences that contribute to their abilities. Perhaps it could be postulated that the ability to sight-sing is less the product of one's choral/vocal training and experience than it is of one's total musical experience.

[6] Read, "An Investigation of the Relationship...," op. cit.

[7] Ottman, "A Statistical Investigation of the Influence...," op. cit.

Table 4-21

Pearson Correlation Between Musical Experience and the Unirhythmic Test

Experience Variable	Coefficient
Scholastic level	.338
Environmental Experience	.506
Instrumental Experience	.585
Private Instrumental Lessons	.560
Keyboard Experience	.543
Choral Experience	.462
Private Vocal Instruction	.426
Additional Musical Studies	.376
Family Experience	.502
Youth Music Experience (high school only)	.343

CHAPTER FIVE

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

The primary purpose of this study was to develop a reliable and valid method for the measurement of the ability to sight-sing. To accomplish this task, a sight-singing test had to be constructed and validated. The test was constructed and its reliability determined in a pilot study involving 191 high school students attending Michigan State University Summer Youth Music. The test was found to be highly reliable when used with these students, and easily administered and scored.

The present study was conceived to validate the test and determine its reliability using high school and college choral students as the subjects. The two specific questions to be answered were 1) the reliability of the test and the individual exercises, and 2) the criterion-related validity of the test. Several secondary questions also were identified: 1) the correct order of difficulty of the test; 2) the reliability of the intervals used in this study; 3) the influence of musical experience upon success on this test; and 4) the comparison of the performances of the three groups on the tests used in the study through analysis of variance.

Review of Literature

A survey of the related literature revealed that the studies involving sight-singing fell into several categories: 1) factors that influence sight-singing; 2) group sight-singing tests; 3) studies specifically related to some aspect of sight-singing; and 4) the construction and validation of instrumental sight-reading or vocal sight-singing tests.

Several studies of the factors that influence sight-singing were reviewed. Studies by Salisbury and Smith, Dean, and Ottman revealed that tonal and melodic abilities correlated highly with the ability to sight-sing.

Studies whose emphasis was group sight-singing methods were discussed. In addition, several other tests using similar techniques in one or more sections were identified. Early studies by Mosher, Hutchinson, and Knuth postulated that sight-singing could be measured by determinine the melodic discrimination or tonal imagery abilities of the student. A later study by White verified these conclusions. The other tests using similar techniques included: 1) the Kwalwasser-Ruch Test of Musical Accomplishment; 2) the Diagnostic Test of Achievement in Music; 3) the Farnum Music Notation Test; and 4) the Jones Music Recognition Test. The researcher raised several questions concerning the validity of measuring sight-singing without any vocal response.

Studies involving some specific aspect of sight-singing were discussed. Three studies by Barnes, Hammer, and Ray using tachistoscopic methods were reviewed. Barnes found that the ability to sight-sing intervals correlated highly with melodic sight-singing ability. Marquis discovered that the ability to sight-sing intervals fluctuated as the context changed in which those intervals occurred. Danfelf determined that the type of material, contrived or composed melodies, used for sight-singing exercises was of little consequence to the ability to sight-sing. Nelson discovered that short items were more reliable, more valid, and better predictors of sight-singing success than the more traditional longer items. Finally, Ritchie found that the context of an exercise confounded the ability of the student to take melodic dictation.

The construction of reliable and valid sight-reading and sight-singing tests was the purpose of several studies. In general, these fell into two categories; those that used the note as the unit of error, and those that used the measure as the unit of error. The tests that reflected the first concept were constructed by Hillbrand, Thostenson, and Cooper. Those that concurred with the second philosophy included Otterstein and Mosher, the Watkins Cornet Performance Scale, the later adaptation of Watkins's work to all band instruments by Stephen Farnum, and the Gutsch rhythm test.

Method for the Study

The methodology for the study was predicated upon several pilot studies conducted by the researcher. One, in particular, involved the construction and administration of a unirhythmic test to 191 high school music students attending Michigan State University Youth Music. The test was constructed using melodic excerpts from standard high school and college choral literature. The results of the pilot study necessitated the reordering of the twenty items, but the entire test was found to be highly reliable (r = .969). Several improvements were made to the unirhythmic test as a result of the pilot study data.

Fifty-four students participated in the study which followed. This sample was drawn from three sources; high school, small college, and university. All the students were singers involved in choral ensembles in their respective environments. The twenty high school students were attending Michigan State University Summer Youth Music; the twenty-four students from small colleges were attending either John Wesley College or Jordan College, two small private religious colleges; and the ten university students attended Michigan State University. All the participants were volunteers, ranging in scholastic experience from freshmen in high school to Master's degree candidates.

A battery of four measurement devices was used in the study: 1) a musical experience questionnaire; 2) a subjective rating involving the singing of a standard choral selection; 3) a standardized traditional sight-singing test; and 4) the original unirhythmic test, designed to measure sight-singing ability without rhythm.

The standardized sight-singing test selected as the criterion for the study was the Otterstein-Mosher (O-M) Sight-Singing Test written in 1932. This test was selected for three reasons: 1) the available test statistics indicated high reliability; 2) the scoring system was consistent, definitive, and relatively simple; and 3) the test contained parallel measures. In addition, the O-M test was as reliable and valid as any of the others reviewed by the researcher, its other characteristics making it seemingly the best choice.

The battery of measurement devices was administered to the students individually by the researcher the same session. The responses were scored during administration by the researcher, and subsequently again one week later. Four additional high school and college music teachers were selected to score the responses of the students for scorer reliability.

The data were transferred from the score sheets to computer cards by the investigator. This data included the music experience of the student, the subjective rating, the O-M test scores, and the unirhythmic test scores. The data were analyzed and processed by the CDC 6500 computer system using SPSS (Statistical Package for the Social

Sciences), version 7.0.

Analysis of the Data

The chapter containing the analysis of the results of the study was divided into several sections. The first contained the group data for the three tests; the subjective rating, the O-M test, and the unirhythmic test. The data for the three groups were subjected to an analysis of variance and the Scheffe technique to determine the significance of the differences. The subjects from the university performed significantly better than those in either of the other two groups.

The second section was comprised of an analysis of the unirhythmic test data, particularly that of the individual exercises. The difficulty ranking was determined using simple percentage of correctness. The discrimination indices were calculated using Net D, and the correlation between each exercise and the total test score was established using a Pearson product-moment correlation.

The reliability of the unirhythmic test was the subject of the third section. This was found using Cronbach's alpha, a measure of internal consistency similar to Kuder-Richardson Formula No. 20. The reliability of each exercise was established and the reliability of the test, first by each melodic interval as a separate response resulting in 97 continuous responses, then using each exercise as a unit resulting in 20 units or items. The test was found to be highly reliable (.974) using the

97 melodic intervals.

The fourth section involved the establishing of scorer reliability. Two types of reliability were computed; single scorer reliability, and multiple scorer reliability. The scoring procedure and results were found to be extremely reliable, both when rescored by the researcher, and when scored by the outside music teachers.

The validity of the unirhythmic test was established in section five. The method for establishing criterion-related validity was to compare the results of the unirhythmic test with those of the O-M test.

Several secondary questions were addressed in the final section concerning the analysis of the data according to interval types, and the relationship of experience to success on the unirhythmic test.

A summary of the most important statistics derived from the data in the study is illustrated in Table 5-1.

CONCLUSIONS

The purpose of this study was to validate and ascertain the reliability of a method for measuring sight-singing ability. This method was somewhat less than traditional as it purported to measure sight-singing without measuring all the elements commonly considered inherent within the concept of sight-singing. It rather purports to measure sight-singing by measuring only one characteristic; namely, pitch. Prior to a discussion of the results that seem to verify the above position, however, an

Table 5-1

Summary of Significant Statistics

Statistical Procedure	Coefficient
Reliability for the Unirhythmic Test	.974
Reliability for all exercises	.933
Reliability for all interval types	.939
Single scorer reliability - Unirhythmic	.999
Single scorer reliability - O-M	.998
Multiple scorer reliability - Unirhythmic (Kendall's W)	.995
Validity - Pearson r: O-M and Unirhythmic	.926
Kendall's W for all tests	.951
Reliability for all tests	.981

examination of the problems of the study would be advisable.

Although it was felt that the subjective rating provided a reasonably accurate assessment of the ability of the singers, certain characteristics rendered this device less successful than might be desired. First, not all the students performed the same choral selection. With experience varying widely among the students, many had previously performed nearly all the selections used for this segment of the study. In order for each student to perform a selection unknown to them, some less frequently performed or more difficult selections were needed to insure that this occurred. An attempt was made to select pieces of equal difficulty. The results, however, may not have been totally successful because of the ill-defined nature of difficulty. The implication was that all students were not assessed using the same measure. This would not preclude the use of this technique since it frequently occurs when used for choral auditions. It does, however, increase the variability in judgment when the researcher must take into consideration varying levels of difficulty.

A second problem that arose was that many of the singers were previously known to the researcher. The reading abilities and musical abilities of many had been witnessed in classes either taught or attended by the investigator. This would not nullify the subjective rating as a legitimate measurement device, since greater knowledge should render a more accurate rating. The question arises whether the rating was a product of the performed selection or of some other consideration. Perhaps the rating would have been more representative of what normally occurs in the field had the above not been true. An accurate rating was achieved, as illustrated by the correlations between the subjective rating and the other two tests. Whether the rating was valid or representative is the question.

The Otterstein-Mosher test also posed several problems. While it had a distribution of a wide range of scores, discriminated between good and poor sight-singers, and subsequently proved to be a satisfactory criterion, the test seemed to lack discrimination with regard to rhythm. As mentioned earlier, the range of rhythmic scores was considerably narrower than that of pitch. The fundamental problem was that the rhythmic difficulty did not progress at the same rate as the melodic difficulty. It was not until the 21st of 28 exercises that the singer encountered any division of the beat. Prior to that time, all notes were either the beat or multiples of the beat. In addition, of the 224 measures, only 22 contained dotted notes that were not multiples of the beat. As a result, some very poor readers were scoring reasonably well on rhythm simply because they sang every note the same duration. The obvious implication here is that the difference between good and poor sight-singers would not be forthcoming when rhythmic performance alone was observed. This certainly would be considered a deficiency of the O-M test. It would seem more appropriate to increase the difficulty of the rhythms equally with the difficulty of the pitch. This does not occur. On behalf of the designers of the test, however, it must be realized that rhythmic difficulty is more clearly defined than melodic difficulty. The tonal and melodic difficulty is contingent upon so many variables that it is virtually impossible to accurately determine difficulty. Nonetheless, it would seem that greater care should have been taken to insure equality of difficulty for both rhythm and pitch.

Another problem with the Otterstein-Mosher test was the lack of availability of the directions for its use. The directions, obtained from the publisher, can be found in Appendix E, along with the O-M test. There are clearly situations that arise in any testing session that are not accounted for in these directions. For example, the directions indicate, "if an examinee stops, makes an error, and wants to sing an exercise over, permit him to do so; however, an error has been made and is to be so recorded." The question arises as to how to score the measure if no error has been made, yet the singer stops and wants to start over again. The directions are not explicit in this instance, nor do they address pauses during a measure, or changes in tempo. These questions and others demand decisions to be made by the researcher; decisions that should have been made by the designers of the test. While the researcher is confident that the decisions made were consistent with the concepts of traditional sight-singing measurement., perhaps another researcher conducting a replicative study would make different decisions resulting in somewhat different results. These problems could have been eliminated if clear directions were available.

One would be naive to believe that all tests are perfect, without any inconsistencies or deficiencies. The attempt is to eliminate as many as possible in the construction of the test. Such was the case with the unirhythmic test. Two problems presented themselves that needed resolution, although there may not be a solution for one of them without completely altering the test. The first concerns the subject who pauses to conceptualize the interval in question. Conceivably, one could pause between each note and silently sing a scale or other assisting device and accomplish the exercise without error. If one had the patience, this could continue for the entire test. Any student with some musical ability should be able to perform well if this should occur. A similar situation could arise when a student figures internally the tonal sequence prior to performing it. In essence, the student would no longer be sight-singing; this would now constitute a practiced performance, though not an audible practice. During the administration sessions for this study, the former actually occurred only once. The possibility of its occurrence was discouraged by the instruction to the singer to sing each exercise at a speed of about one note per second without stopping. Most students complied with this request. The latter occurred on several occasions. This problem was dispelled, for the most part, by the researcher immediately playing both the new tonality and starting pitch for the next exercise. Only on one occasion did the singer persist in humming the exercise silently before singing it. Other than this one incident, the problem was eliminated.

The second problem, which may not be amendable, occurred when an excellent sight-singer lost concentration for a moment or made a thoughtless mistake. In so doing, they greatly decreased their overall score on the test by five to eight percent, if it happened only once. This is due to the extreme shortness of the test. In traditional sight-singing tests, the singer has ample time to redeem themselves without a simple mistake having a significant effect on the total score. In this test, however, thoughtless errors tend to substantially reduce the total score. This problem is a by-product of the intent and purpose of this test. The assumption was made that for a test to be usable it should be easy to administer, quick to score, and definitive regarding the scoring results. It was discovered, however, that many of the thoughtless errors could be eliminated by simple admonishments to the singers, such as "perhaps you might want to sing just a little slower," or similar direction. These types of instructions have nothing to do with the singer's ability, nor with unusual instructions that could be classified as prompting or reinforcement, since they have nothing to do with the actual test material. They simply are foreseeing problems and curtailing them before they occur.

The final consideration that must be classified a deficiency in this study was the small number of subjects, particularly in the large college sample. Ten subjects simply was not enough for any type of generalization other than for the subjects tested. Conversely, perhaps the students at John Wesley or Jordan Colleges were not representative of the "average" choral student attending a small college. In addition, a larger number of females participated in the study than males. One would presume no difference in the sight-singing abilities between the sexes, although that possibility does exist. Two considerations are important with regard to sample. First, all the subjects were volunteers. It is virtually impossible to have a randomly selected volunteer sample. Second, the nature of the study was such that, with the exception of the ANOVAS, the only criterion necessary for the subjects was that they be singers. As each subject was his own control, and totally independent from the others, randomization or a more clearly defined sample was not important. For the ANOVAS, the reader will recall the discussion about the assumptions of analysis of variance. The only assumption that came close to being violated was that of equality of variance. The results of Bartlett's test for homogeneity of the variance was such that the null hypothesis could be accepted at the .05 level, though marginally. If the intent was to generalize to the universal population of singers, the results would need to be much stronger. The intent here, however, was to generalize only to this specific sample. Despite these considerations, a larger sample, particularly of university students, would have made the results more applicable to other samples.

Having considered the problems of the study, the accomplishments and implications merit consideration. Perhaps the most significant accomplishment of this study is that another sight-singing test has been validated and shown to be extremely reliable as used with this sample. The characteristic of this test that distinguished it from other available tests was its usability. This was primarily due to the ease of administration and the scoring procedure. Concerning the ease of administration, the test requires no elaborate equipment for adminstering the test. Only a one page copy of the test, some type of scoring sheet that will allow for right and wrong answers, a tape-recorder if desired, and a keyboard would be necessary. The test can be administered in less than five minutes and scored during the same time. With regard to scoring, the test can be scored during administration, if desired. The amount of time required by the additional scorers to score the subjects in this study ranged from five hours to eight hours, including the time necessary to familiarize themselves with the test and the scoring system. The researcher is able to score from twelve to fifteen subjects per hour. The reader will recall the correlation coefficients between the outside scorers and the researcher, and between the researcher and himself at a later date. Perhaps these considerations will encourage the use of the test by public school teachers or college choral conductors. Certainly other tests demonstrate high reliability and validity. This test, however, seems to be more usable than most. In short, a sight-singing test without rhythm is available that is 1) reliable, 2) valid, 3) easily administered, 4) can be scored in less than five minutes, and 5) has group norms available.

Another conclusion of this study is that not only was the test valid, but this method of measuring sight-singing was also valid. Despite the elimination of rhythm, a high correlation (.926) between the unirhythmic test and the traditional test with rhythm was found. The assumption would be that the singer who excels in pitch reproducing abilities has also achieved a similar degree of rhythmic competency. This is not to imply altering the instruction of sight-singing. If instruction in sight-singing were based upon the concepts and philosophy of this test, melodic and rhythmic sight-singing would soon be in a state of degeneracy. This is only a short cut approach for assessing sight-singing, not teaching it. Fortunately, this will not occur; teachers will continue to instruct in the manner currently practiced. This test is only valid as a sight-singing measure if the assumption can be made that rhythmic reading and tonal melodic reading are achieved simultaneously. If they were not, no inferences could be made from the achievement of pitch to the achievement of rhythm.

This study, as have others before it, raises serious questions as to the experience factors that contribute to achievement in sight-singing. A multitude of musical experiences were correlated with success on the unirhythmic test. Instrumental experience, private instrumental lessons, keyboard, environmental, and family experiences all exhibited higher correlations with sight-singing success than did choral experience or private vocal instruction. One would have to assume that a variety of experiences contribute to achievement of this skill, and all others. As of yet, however, the combination of experiences that best predicts success in sight-singing has not been identified.

Some interesting questions have arisen concerning the musical abilities of students from small colleges as compared to university students. Admittedly the samples were small, too small on which to base any far-reaching conclusions, but large enough to pose the questions for further research. One would wonder if the colleges used in this sample were not representative of the "average" small college; if the small college attracts a different type of student; whether it provides a different caliber of education; or if the results of this sample are primarily due to sampling error.

Despite the intriguing questions raised by the findings of this study, the most substantial remains the availability of a reliable, valid and usable sight-singing test. The results can only be generalized to this specific sample and further research would be needed to generalize any results universally.

RECOMMENDATIONS FOR FURTHER STUDY

The findings of this investigation suggest replication with a much larger sample, equal sample group sizes, and stricter controls on the testing procedures. In addition, the following recommendations are made:

1. To insure the usability of the unirhythmic test, a large sample of high school and college students should be tested by their respective teachers using this device.

2. A study, on a much larger basis, should be conducted to identify significant differences, if any, in students from large and small colleges with regard to experience, performance, and ability.

3. An investigation involving teachers in the field should be conducted using the subjective approaches to the measurement of sight-singing and comparing the results with a traditional standardized test. This should be accomplished using a large sample and careful controls.

4. A study which would be of great interest to this writer originated from observations made during the testing sessions. It was extremely humorous to witness a student, particularly a Master's degree student or upper level college student, totally obliterate an exercise, whereupon exclaim, "There, at least I got that one," or something similar. It would be interesting to conduct a study that allowed the students to rate their performances after each exercise. Singers seem to have an unusually inaccurate conception of their abilities.

BIBLIOGRAPHY

BIBLIOGRAPHY

Books

Apel, Willi. Harvard Dictionary of Music. Cambridge: Harvard University Press, 1962.

Baggaley, Andrew R. Intermediate Correlational Methods. New York: John Wiley & Sons, Inc., 1964.

Bentley, Arnold. Aural Foundations of Music Reading. London: Novello and Company, Ltd., 1966.

Benward, Bruce. Ear Training. Dubuque: Wm. C. Brown Company Publishers, 1969.

____________. Sight Singing Complete. Dubuque: Wm. C. Brown Company Publishers, 1965.

Berkowitz, Sol, Frontrier, Gabriel, and Kraft, Leo. A New Approach to Sight Singing. New York: W. W. Norton and Co., Inc., 1960.

Colwell, Richard. The Evaluation of Music Teaching and Learning. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970.

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper & Row, Publishers, 1970.

Cronbach, Lee J. and Gleser, Goldine C. Psychological Tests and Personnel Decisions. Urbana, Ill.: University of Illinois Press, 1965.

Downie, N. M. and Heath, R. W. Basic Statistical Methods. New York: Harper and Row, Publishers, 1965.

Dykema, Peter W. and Cundiff, Hannah M. School Music Handbook. Boston: C. C. Birchard and Company, 1955.

Ebel, Robert L. Measuring Educational Achievement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1965.

Edlund, Lars. Modus Novus. Stockholm: Nordiska Musicforlaget, 1963.

__________. Modus Vetus: Sight-Singing and Ear-Training. in Major/Minor Tonality. Stockholm: Nordiska Musikforlaget, 1963.

Edwards, Allen L. Statistical Methods. New York: Holt, Rinehart and Winston, Inc., 1967.

Farnsworth, Paul R. The Social Psychology of Music. Ames Iowa: The Iowa State University Press, 1969.

Ferguson, George A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Company, 1966.

Fish, Arnold, and Lloyd, Norman. Fundamentals of Sight Singing and Ear Training. New York: Dodd, Mead and Co., 1969.

Gaston, E. Thayer. Music in Therapy. New York: The Macmillan Company, 1968.

Glass, Gene V. and Stanley, Julian C. Statistical Methods in Education and Psychology. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970.

Grove, Sir George. Dictionary of Music and Musicians. ed. H. C. Colles, 3rd ed. New York: The Macmillan Company, 1939.

Harder, Paul O. Fundamentals of Music Reading. St. Louis: Educational Publishers, Inc., 1950.

__________. Harmonic Materials in Tonal Music. Boston: Allyn and Bacon, Inc., 1968.

Hayes, William L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, Inc., 1963.

Hillbrand, Earl K. Measuring Ability in Sight-Singing. Ann Arbor, Michigan: Edwards Brothers Pub., 1924.

Hindemith, Paul. Elementary Training for Musicians. New-York: Associated Music Publishers, 1949.

Horacek, Leo, and Lefkoff, Gerald. Programmed Ear Training. New York: Harcourt, Brace and World, 1970.

Kwalwasser, Jacob. Tests and Measurements in Music. Boston: C. C. Birchard and Company, 1927.

Leeder, Joseph A., and Haynie, William S. Music Education in the High School. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1960.

Lehman, Paul R. Tests and Measurements. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.

Lieberman, Maurice. Ear Training and Sight Singing. New York: W. W. Norton and Company, Inc., 1959.

Lundin, Robert W. An Objective Psychology of Music. New York: The Ronald Press Company, 1953.

McHose, Allen Irvine, and Tibbs, Ruth N. Sight-Singing Manual. New York: F. S. Crofts and Co., 1944.

Mosher, Raymond M. A Study of the Group Method of Measurement of Sight Singing. New York: Teachers College, Columbia University, 1925.

Murphy, Howard A. Teaching Musicianship. New York: Coleman-Ross Company, Inc., 1950.

Mursell, James L. Education for Musical Growth. Boston: Ginn and Company, 1948.

______________. Principles of Musical Education. New York: The MacMillan Company, 1931.

Mursell, James L. and Glen, Mabelle. The Psychology of School Music Teaching. New York: Silver Burdett Company, 1938.

Nation Society for the Study of Education. Basic Concepts in Music Education. Fifty-seventh Yearbook, Part I. Chicago: The University of Chicago Press, 1958.

Nunnally, Jum C. Educational Measurement and Evaluation. New York: Mc Graw-Hill Book Company, 1972.

______________. Introduction to Psychological Measurement. New York: McGraw-Hill Book Company, 1970.

Ottman, Robert. Music for Sight-Singing. New York: Prentice-Hall, Inc., 1967.

Phelps, Roger P. A Guide to Research in Music Education. Dubuque, Iowa: Wm. C. Brown Company Publishers, 1969.

Popham, W. James. Educational Statistics. New York: Harper and Row, Publishers, 1967.

Roe, Paul F. Choral Music Education. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970.

Ruch, G. M. and Stoddard, George D. Tests and Measurements in High School Instruction. Yonkers, New York: World Book Company, 1927.

Schoen, Max. The Psychology of Music. New York: The Ronald Press Company, 1940.

Seashore, Carl E. Psychology of Musical Talent. New York: Silver Burdett and Company, 1919.

Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, 1956.

Solomon, Herbert. Studies in Item Analysis and Prediction. Stanford, California: Stanford University Press, 1961.

SPSS (Statistical Package for the Social Sciences) Manual and Supplement. New York: McGraw-Hill Book Company, 1970.

Squire, Russel N. Introduction to Music Education. New York: The Ronald Press Company, 1952.

Thomson, William. Introduction to Music Reading: Concepts and Applications. Bellmont, California: Wadsworth Publishing Company, Inc., 1966.

________________. Advanced Music Reading. Belmont, California: Wadsworth Publishing Company, Inc., 1969.

Trubitt, Allen R. and Hines, Robert S. Ear Training and Sight-Singing; An Integrated Approach. New York: Schirmer Books, 1979.

Watkins, John Goodrich. Objective Measurement of Instrumental Performance. New York: Bureau of Publications, Teachers College, Columbia University, 1942.

Wedge, George A. Advanced Ear Training and Sight-Singing. New York: G. Schirmer, Inc., 1922.

de Zeeuw, Anne Marie, and Foltz, Roger E. Sight Singing and Related Skills. Manchaca, Texas: Sterling Swift Publishing Company, 1973.

______________________________________. Sight Singing: Melodic Structures in Functional Tonality. Manchaca, Texas: Sterling Swift Publishing Company, 1978.

Periodicals

Bean, Kenneth L. "An Experimental Approach to the Reading of Music," Psychological Monographs, L (1938), pp. 1 - 79.

Buttram, Joe B. "The Influence of Selected Factors on Interval Identification," Journal of Research in Music Education, Fall, 1969, pp. 305 - 315.

Chadwick, J. E. "Predicting Success in Sight Singing," Journal of Applied Psychology, XVII (December, 1933), pp. 671 - 674.

Drake, Raleigh M. "Four New Tests of Musical Talent," Journal of Applied Psychology. XVII (April, 1933), pp. 136 - 147.

__________________. "The Validity and Reliability of Tests of Musical Talent," Journal of Applied Psychology. XVII, (August, 1933), pp. 447 - 458.

Dean, Charles D. "Predicting Sight Singing Ability in Teacher Education," Journal of Educational Psychology, XXVIII (November, 1937), pp. 601 - 608.

Gaw, Ester Allen. "Five Studies of Musical Tests." Psychological Monographs. XXXIX (1928), pp. 145 - 156.

Gutsch, Kenneth Urial. "Evaluation in Instrumental Music: An Individual Approach," Council for Research in Music Education. Nos. 5 and 6 (1964), pp. 21 - 28.

Hammer, Harry. "An Experimental Study of the Use of the Tachistoscope in the Teaching of Melodic Sight Singing," Journal of Research in Music Education, XI (1963), pp. 44 - 45.

Helwig, Herman. "Rhythmic Approach to Sight Reading." The Instrumentalist, IX (February, 1955), pp. 6 - 7.

Lundin, Robert. "The Development and Validation of a Set of Musical Ability Tests." Psychological Monographs, LXIII (1949), pp. 1 - 20.

Mc Naught, W. G. "The Psychology of Sightsinging." Proceedings of the Musical Association, XXVI (1900), pp. 35 - 55.

Ottman, Robert. "Music for SightSinging." Journal of Research in Music Education, IV (Fall, 1956), pp. 144 - 145.

Ruch, G. M. and Stoddard, G. D. "Comparative Reliabilities of Five Types of Objective Examinations. Journal of Educational Psychology, (January and February, 1925).

Salisbury, F. S. and Smith, H. B. "Prognosis of Sight Singing Ability of Normal School Students." Journal of Applied Psychology, XIII (October, 1929), pp. 425 - 439.

Seashore, Carl E. "The Measurement of Pitch Discrimination." Psychological Review Monograph, LIII (1910), pp. 21 - 60.

Thostenson, M. S. "The Study and Evaluation of Certain Problems in Ear-Training Related to Achievement in Sight-Singing and Music Dictation." Council for Research in Music Education, XI (1967), pp. 14 - 35.

Unpublished Material

Barnes, James Woodrow. "An Experimental Study of Interval Drill as It Affects Sight Singing Skill." Unpublished Ph.D. dissertation, Indiana University, 1960.

Buttram, Joseph. "The Influence of ertain Factors on Interval Identification." Unpublished Ph.D. dissertation, University of Kansas, 1967.

Christ, William E. "The Reading of Rhythm Notation Approached Experimentally According to Techniques and Principles of Word Reading." Unpublished Ph. D. dissertation, Indiana University, 1953.

Cooper, John J. "The Development of a Sight-Singing Achievement Test." Unpublished Ph.D. dissertation, University of Colorado, 1965.

Danfelt, Lewis Seymour. "An Experimental Study of Sight Singing of Selected Groups of College Music Students." Unpublished Ph.D. dissertation, The Florida State University, 1970.

Edmonson, Frank. "The Effect of Interval Direction on Pitch Acuity in Solo Vocal Performance." Unpublished Ph.D. dissertation, The Florida State University, 1967.

Gonzo, Carroll Lee. "An Analysis of Factors Related to Choral Teachers Ability to Detect Pitch Errors While Reading the Score." Unpublished Ph.D. dissertation, The University of Wisconsin, 1969.

Hammer, Harry. "An Experimental Study of the Use of the Tachistoscope in the Teaching of Melodic Sight Singing." Unpublished Ph.D. dissertation, University of Colorado, 1961.

Helbing, Devon Willis. "An Experimental Study of the Relative Effectiveness of 'Whole' and 'Part' Methods of Teaching." Unpublished Ph.D. dissertation, Indiana University, 1965.

Jones, Howell Thomas. "The Relationship of Selected Factors and Music Reading Achievement." Unpublished Ph.D. dissertation, Michigan State University, 1968.

Justus, Lane Dale. "Evaluation of an Innovative Instructional Design for Sight Singing." Unpublished Ph.D. dissertation, University of Arizona, 1970.

Karl, Harold Thomas. "The Effects of Melodic Dictation and Sight Singing on Music Reading Achievement." Unpublished Ph.D. dissertation, Michigan State University, 1971.

Kanable, Betty Mae. "An Experimental Study Comparing Programmed Instruction with Classroom Teaching of Sight Singing." Unpublished Ph.D. dissertation, Northwestern University, 1964.

Knuth, William Edward. "The Construction and Validation of Music Tests Designed to Measure Certain Aspects of Sightreading." Unpublished Ph.D. dissertation, University of California, 1932.

Madsen, Clifford K. "The Effects of Scale Direction on Pitch Acuity in Solo Vocal Performance." Unpublished Ph.D. dissertation, The Florida State University, 1963.

Marquis, James H. "A Study of Interval Problems in Sight Singing Performance with Considerations of the Effect of Context." Unpublished Ph.D. dissertation, University of Iowa, 1963.

Miles, Walter Richard. "Accuracy of the Voice in Simple Pitch Singing." Unpublished Ph.D. dissertation, University of Iowa, 1913.

Nelson, John Charles. "A Comparison of Two Methods of Measuring Achievement in Sight Singing." Unpublished Ph.D. dissertation, The University of Iowa, 1970.

Norwood, Earl. "The Design, Construction, and Validation of a Test of Melodic Pitch Discrimination Ability." Unpublished D.M.A. dissertation, University of Oregon, 1972.

Ottman, Robert. " A Statistical Investigation of the Influence of Selected Factors on the Skill if Sight-singing." Unpublished Ph.D. dissertation, North Texas State University, 1956.

Pagan, Keith Areatus. "An Experiment in the Measurement of Certain Aspects of Score Reading Ability." Unpublished Ed.D. dissertation, Indiana University, 1970.

Pottenger, Harold Paul. "An Analysis of Rhythm Reading Skill." Unpublished Ed.D. dissertation, Indiana University, 1969.

Powell, Ira Chesley. "A Study of the Relationship of Singing Accuracy to the Pitch Matching Abilities of Eight-One Subjects." Unpublished Ed.D. dissertation, The University of Oklahoma, 1969.

Ray, Harry Burton. "An Experimental Approach to the Reading of Pitch Notation." Unpublished Ph.D. dissertation, Indiana University, 1964.

Read, John William. "An Investigation of the Relationship of Selected Variables to Sight-Singing Ability." Unpublished Ed.D. dissertation, North Texas State University, 1968.

Rodeheaver, Reuben Ellis. "An Investigation of the Vocal Sight Reading Ability of College Freshmen Music Majors." Unpublished Ed.D. dissertation, The University of Oklahoma, 1972.

Sherburn, Merrell. Basic Sight Singing and Theory Skills. Experimental text used at Michigan State University, developed under the auspices of the Humanities Research Center and the Education Development Program. 1972.

Stelzer, Theodore G. W. "Construction, Interpretation and Use of a Sight Reading Scale in Organ Music with an Analysis of Organ Playing into Fundamental Abilities." Unpublished Ph.D. dissertation, The University of Nebraska, 1935.

Stokes, Charles F. "An Experimental Study of Tachistoscopic Training in Reading Music." Unpublished Ph.D. dissertation, Teachers College, University of Cincinnati, 1944.

Thiebe, Edward Henry. "Differential Effects of Interval Presentation Orders upon Developmental Sight Singing Behavior." Unpublished Ph.D. dissertation, The University of Connecticut, 1973.

Tucker, David Walter. "Factors Related to Musical Reading Ability of Senior High School Students Participating in Choral Groups." Unpublished Ed.D dissertation, University of California, Berkeley, 1969.

Tucker, Gerald L. "The Influence of Isolated Rhythmic Drill on Growth in Sight Singing." Unpublished Ed.D dissertation, The University of Oklahoma, 1969.

White, Adolph Peter. "The Construction and Validation of a Group Test in Music Reading for Intermediate Grades." Unpublished Ph.D. dissertation, University of Minnesota, 1963.

Wiley, Charles Albert. "An Experimental Study of Tachistoscopic Techniques in Teaching Rhythmic Sight Reading in Music." Unpublished Ed.D. dissertation, University of Colorado, 1962.

Wilcox, Eunice Ann. "The Effects on Sight Singing of Voice Class Instruction Utilizing Variants of Traditional Vocalises." Unpublished Ph.D. dissertation, Michigan State University, 1968.

Zimmerman, C. Robert. "Relationship of Musical Environment to Choral Sight Reading Ability." Unpublished Ed.D. dissertation, University of Oregon, 1962.

Published Tests and Manuals

Beach, Frank A. Beach Standardized Music Tests. Emporia, Kansas: Bureau of Education Measurements and Standards, Kansas State Normal School, 1920.

Farnum, Stephen E. Farnum Music Notation Test. New York: The Psychological Corporation, 1949.

Gildersleeve, Glenn. Gildersleeve Music Achievement Test. New York: Bureau of Publications, Teachers College, Columbia University, 1933.

Hutchinson, Herbert E. The Hutchinson Music Tests. Bloomington, Illinois: Public School Publishing Co., 1924.

Jones, A. N. Jones Music Recognition Test. New York: Carl Fischer, Inc., n. d.

Knuth, William E. Knuth Achievement Tests in Music. Philadelphia: Educational Testing Bureau, 1936.

Kotick, M. L. and Torgerson, T. L. Diagnostic Tests of Achievement in Music. Los Angeles: Los Angeles Test Bureau, n. d.

Kwalwasser, Jacob and Dykema, Peter W. K - D Music Tests. New York: Carl Fischer, Inc., 1930.

Kwalwasser, Jacob and Ruch, G. M. Kwalwasser - Ruch Test of Musical Accomplishment. Iowa City: Bureau of Research and Service, State University of Iowa, 1924.

Otterstein, Adolph and Mosher, Raymond. Manual of Directions: O-M Sight Singing Test. Stanford, California: Stanford University Press, 1932.

Seashore, Carl E. Manual of Instructions and Interpretations for the Seashore Measures of Musical Talents. Chicago: C. H. Stoelting Co., 1940.

Watkins, John G. and Farnum, Stephen E. The Watkins-Farnum Performance Scale for All Band Instruments. Winona, Minnesota: Leonard Music Co., Inc., 1954.

APPENDICES

APPENDIX A

I. High School Questionnaire

II. College Questionnaire

MUSICAL EXPERIENCE QUESTIONNAIRE

High School

NAME__________________

1. ____Grade in school (this coming year)

   a. 9th b. 10th c. 11th d. 12th

2. ____Musical activities (identify all types that you participate in)

   a. family activities (singing, concerts, etc.)
   b. church activities (solos, groups, church choir, etc.)
   c. school activities (band or choir)
   d. formal training (private lessons)

3. ____Instrumental experience (circle number of years)

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

4. ____Private Instrumental Lessons (circle number of years)

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

5. ____Piano experience (circle number of years)

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

6. ____Choral experience (circle number of years)

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

7. ____Private Voice lessons (circle number of years)

   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

8. ____Additional studies (regular classes in school only)

   a. music theory
   b. music history
   c. ear-training or sight-singing
   d. other (explain)______________________________

9. ____Years of MSU Youth Music (present year included)

   1 2 3 4 5

10. ____Family music experience (identify all correct statements)

    a. father plays instrument, keyboard, or sings.
    b. mother plays instrument, keyboard, or sings.
    c. brothers or sisters play instrument, keyboard, or sings.
    d. grandparents play instrument, keyboard, or sings.
    e. nobody does anything musical except me. (Boo Hoo)

THANK YOU FOR YOUR PARTICIPATION AND COOPERATION! ! !

MUSICAL EXPERIENCE QUESTIONNAIRE

College

NAME____________________

1. ____Scholastic level (next year's)

    a. freshman c. junior e. Master's
    b. sophomore d. senior f. Ph. D.

2. ____Musical activities (Identify all experiences)

    a. family activities
    b. church activities
    c. college or university ensembles
    d. formal training (private lessons)
    e. all of the above

3. ____Instrumental experience (circle number of years)

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

4. ____Private instrumental lessons (circle number of years)

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

5. ____Keyboard training (circle number of years)

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

6. ____Choral experience (circle number of years)

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

7. ____Private voice lessons (circle number of years)

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

8. ____Emphasis of musical activities at college or university

    a. choral/vocal
    b. instrumental
    c. keyboard
    d. music theory and/or composition
    e. music history and literature

9. ____Family experience (identify all correct statements)

    a. father plays instrument, keyboard, or sings.
    b. mother plays instrument, keyboard, or sings.
    c. brothers or sisters play instrument keyboard or sings.
    d. grandparents play instrument, keyboard, or sings.
    e. nobody does anything musical (except yourself)

THANK YOU FOR YOUR PARTICIPATION AND COOPERATION! ! !

APPENDIX B

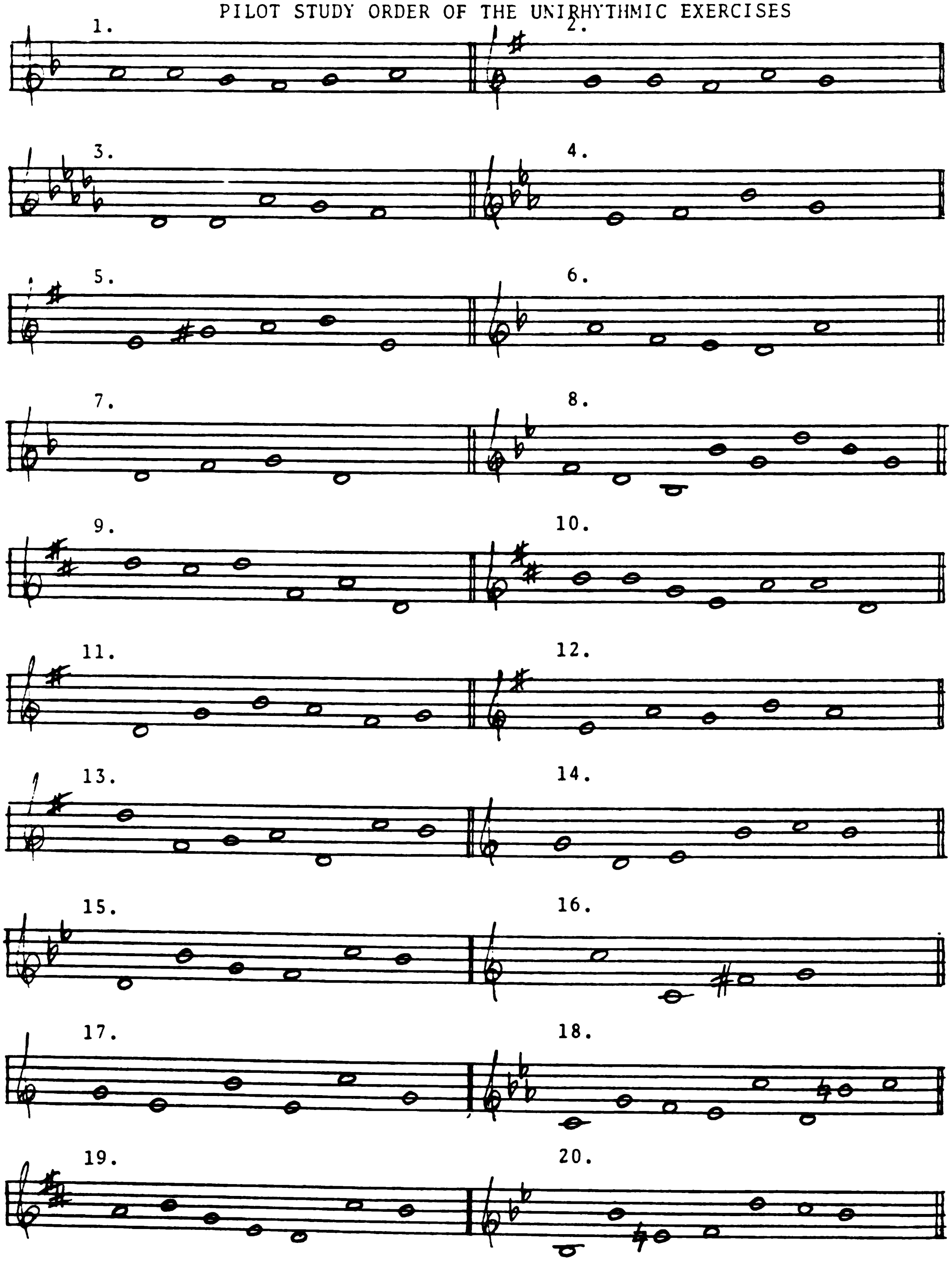
UNIRHYTHMIC TEST

I. Pilot Study Order

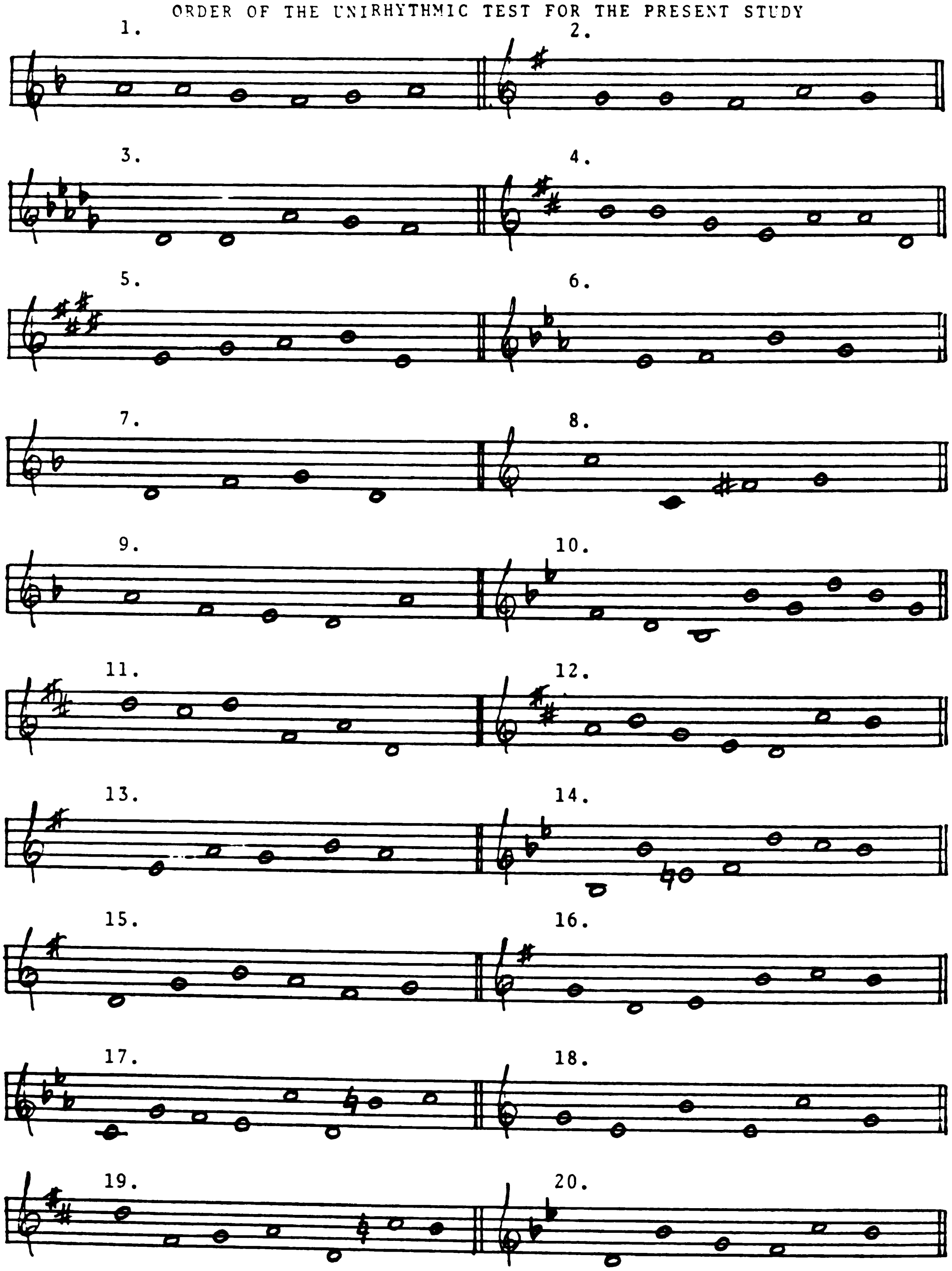
II. Present Study Order

III. New Order

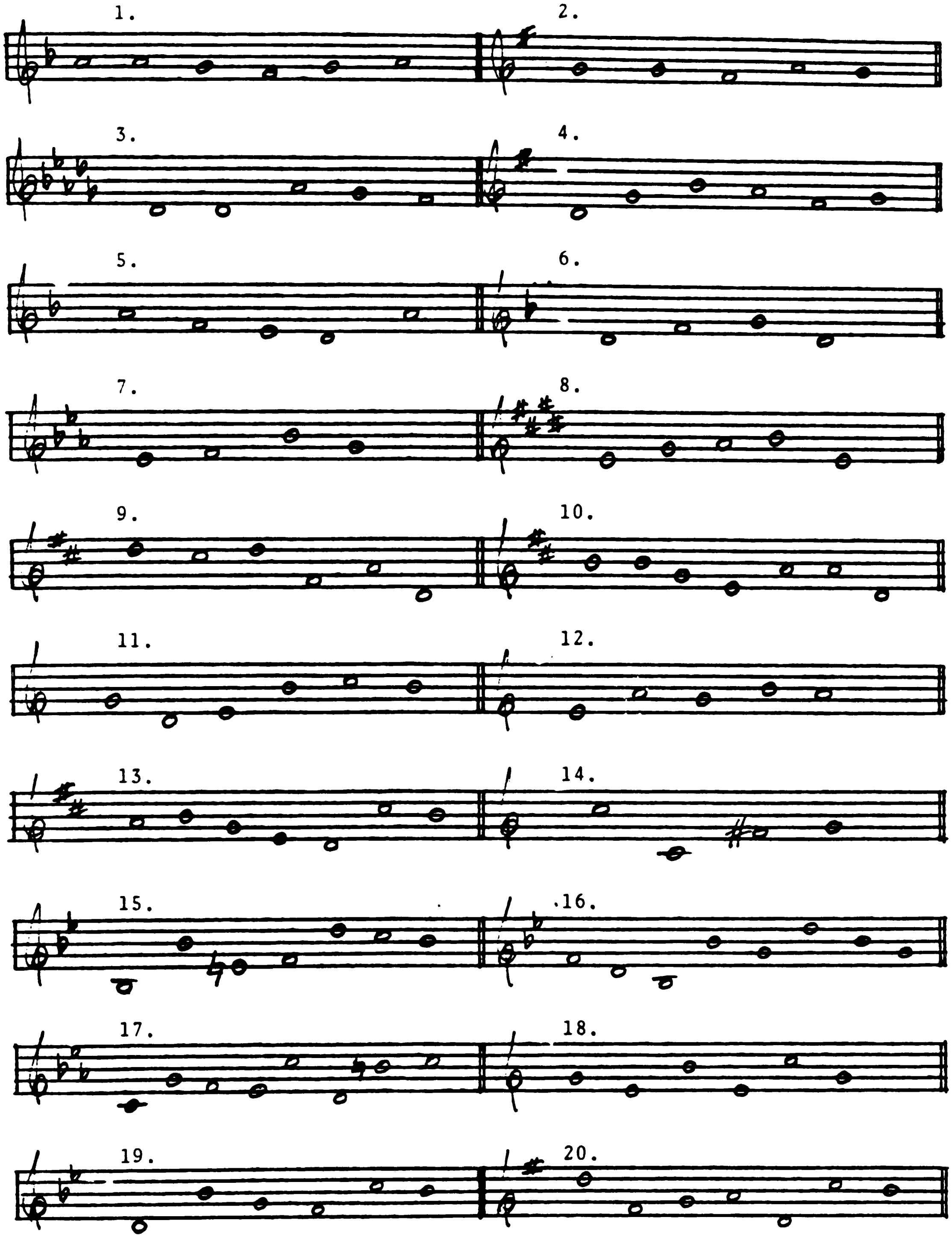
PILOT STUDY ORDER OF THE UNIRHYTHMIC EXERCISES



ORDER OF THE UNIRHYTHMIC TEST FOR THE PRESENT STUDY



NEW ORDER OF THE UNIRHYTHMIC TEST BASED UPON THE PRESENT DATA



APPENDIX C

Pilot Study Instructions for Test Administrators

INSTRUCTIONS FOR TEST ADMINISTRATORS

1. Before taking the sight-singing test, each subject should complete the musical experience questionnaire in total.

2. When taping the student responses, identify the student by name and number. The alphanumeric identifier will consist of a letter followed by the number of the subject. The letter represents the administrator and the number, the subject. An ID might appear A10, B16, etc.

3. The test should be recorded in its entirety, not stopping the tape between each excerpt.

4. The student should be asked to sing each excerpt on a neutral syllable of his choice (recommend "lah"), reminded not to sing too fast, (a tempo of about one note per second) and instructed to sing each through without stopping. In the event that the student should break down completely, give the student the pitch that he should be singing and continue to the end.

5. The administrator shall give the student either the starting pitch or both the starting pitch and tonic chord as indicated. Every other student shall be given the pitch while the alternate students are given both. Be sure to start the tape before giving the pitch or chord and pitch of the first exercise.

6. Allow the student only a minimum of time between each excerpt, no longer than ten (10) seconds. Ask if he is ready and then give the starting pitch or pitch and chord.

7. It may occur that a student will start, stop, and then ask for the starting pitch again. If he has not sung the second pitch, give him the pitch and indicate for him to begin. If he has sung the second pitch or more, merely give him the pitch that he should be singing and have him continue.

8. It may also occur that the student may, after looking at an exercise, indicate that he cannot sing it. If this occurs, encourage him to try and give the starting pitch or pitch and chord.

9. After the student has sung the entire test, thank him, and keep going.

THANK YOU FOR YOU HELP! !

APPENDIX D

Choral Selections Used as Basis for Construction of the Unirhythmic Test

CHORAL SELECTIONS

	All Breathing Life	J. S. Bach
	Alleluja	J. S. Bach
*	Crucifixus (B minor Mass)	J. S. Bach
**	Das Lamm, das erwurget ist	J. S. Bach
	Praise and Thanks	J. S. Bach
	Praise Him	J. S. Bach
	Sanctus (B minor Mass)	J. S. Bach
	Sanctus in D	J. S. Bach
	Komm Süesser Tod	F. M. Bach
	Let down the bars, O Death	Samuel Barber
	Three Hungarian Folk Songs	Bela Bartok
**	Canticle of Praise	John Ness Beck
**	Cry Aloud	John Ness Beck
** *	Song of Exultation	John Ness Beck
**	Hallelujah (from the Mount of Olives)	Ludwig von Beethoven
*	The Eyes of All Wait Upon Thee	Jean Berger
** *	How Beautiful Upon the Mountains	Jean Berger
	How Lovely are Thy Tabernacles (I, II, III)	Jean Berger
	From Isaiah	Jean Berger
** *	It is Good to be Merry	Jean Berger
	Gloria Tibi (from Mass)	Leonard Bernstein
**	The Bird	William Billings
**	Create in Me, O God	Johannes Brahms
** *	Grant Unto Me the Joy of Thy Salvation	Johannes Brahms
** *	How Lovely is Thy Dwelling Place	Johannes Brahms
	O Cast Me Not Away From Thy Countenance	Johannes Brahms
**	O Heiland, reiss die Himmel auf	Johannes Brahms
	Six Folk Songs	Johannes Brahms
**	I Hear a Voice A-Prayin	Houston Bright
**	Alleluia, Alleluia	Dietrich Buxtehude
	Praise Our Lord, All Ye Gentiles	William Byrd
	This is the Day	Paul Christiansen
**	Three Chorales from Tagore	Paul Creston
	Jesu, Priceless Treasure	Johann Crüger
**	Ain-a That Good News	William L. Dawson
	Hail Mary	William L. Dawson
	Soon-ah Will Be Done	William L. Dawson
**	Three Madrigals	David Diamond
	Praise Ye the Lord	Anthony Donato
	Ave Regina Coelorum	Guillaume Dufay
	Agnus Dei	Gabriel Fauré
	Sanctus	Gabriel Fauré
	Psalm 150	César Franck
	Magnificat	Andrea Gabrieli
	Benedictus	Giovanni Gabrieli
**	Jubilate Deo	Giovanni Gabrieli
**	Ascendit Deus	Jacobus Gallus

	Four American Patriots (Thomas Jefferson)	Earl George
**	Hosanna to the Son of David	Orlando Gibbons
*	Sing Praises	L. Stanley Glarum
** *	Sing and Be Joyful	Carl Heinrich Graun
	Surely He Hath Borne Our Griefs	Carl Heinrich Graun
	Antiphonal Hosanna	Christian Gregor
	Live A-Humble	Jester Hairston
	A Collect for Peace	Robert A. Harris
**	Glory to God	Robert A. Harris
	Kyrie	Robert A. Harris
**	The Heavens are Telling	Franz Joseph Haydn
*	Six Chansons	Paul Hindemith
*	Sixty-Seventh Psalm	Charles E. Ives
	Ain't Got Time To Die	Hall Johnson
	By the Rivers of Babylon	Joseph Kantor
	Counterpoint	Sven Lekberg
**	Crucifixus	Antonio Lotti
	Bitte	C. L. Madison
** *	Great Day	Warren Martin
	Two Sandburg Songs	Holon Matthews
**	He, watching over Israel	Felix Mendelssohn
	Magnificat	Daniel Moe
	Daniel, Daniel, Servant of the Lord	Undine S. Moore
	Laudate Dominum	Wolfgang Amadeus Mozart
	Laudate Pueri	Wolfgang Amadeus Mozart
**	Regina Coeli	Wolfgang Amadeus Mozart
	He's Gone Away	Edmund Najera
	The Gallows-Tree (Four Ballads)	Vaclav Nelhybel
** *	Psalm 150	Kent A. Newbury
**	Adoramus Te	G. P. daPalestrina
**	Super Flumina Babylonis	G. P. daPalestrina
	Jubilate Deo Omnis Terra	Flor Peeters
	Ascendit Deus	Peter Philips
**	Cantate Domino	Giuseppe Ottavio Pitoni
	The Face of God	Frank Pooler
** *	Exultate Deo	Francis Poulenc
	Salve Regina	Francis Poulenc
	Tenebrae factae sunt	Francis Poulenc
	Timor et tremor venerunt super me	Francis Poulenc
**	Ecce tu pulchra es	Josquin Des Prés
	Glory Be to God	Sergei Rachmaninoff
	A Psalm of Praise	H. Owen Reed
	Come, Come Ye Saints	Leroy J. Robertson
	All Glorius God	Ned Rorem
**	Canticles	Ned Rorem
**	Four Madrigals	Ned Rorem
	Ave Maria	Ronald Roxbury
** *	Exultate Deo	Alessandro Scarlatti
**	Friede auf Erden	Arnold Schönberg
*	The Unknown Region	William Schuman
	To All, To Each	William Schuman

**	Cantate Domino	Heinrich Schütz
	O All Ye Nations	Heinrich Schütz
	Three Hungarian Folksongs	Matyas Seiber
**	Morning Trumpet	Robert Shaw
	The King of love my Shepherd is	Harry Rowe Shelley
	Onward, Ye People	Jean Sibelius
	Anthem for Spring	Pietro Mascagni
	Ride the Chariot	William Henry Smith
	A Little Nonsense	Robert Starer
	To Every Thing There is a Season	Robert Starer
** *	Ave Maria	Igor Stravinsky
	Pater Noster	Igor Stravinsky
*	Alleluia	Randall Thompson
**	The Best of Rooms	Randall Thompson
	Choose Something Like a Star	Randall Thompson
	The Last Words of David	Randall Thompson
	Road Not Taken	Randall Thompson
	Say Ye to the Righteous	Randall Thompson
**	Festival Anthem	Rodger Vaughn
** *	Ave Maria	Tomás Luis de Victoria
	Where does the uttered music go	William Walton
**	Benedicamus Domino	Peter Warlock
	Battle Hymn of the Republic	Peter J. Wilhousky

EXTENDED WORKS

**	The Passion According to St. Matthew	J. S. Bach
	Christ lag in Todesbanden	J. S. Bach
	Fürchte dich nicht	J. S. Bach
	Ich hatte viel Bekümmernis	J. S. Bach
**	Jesu meine Freude	J. S. Bach
**	Komm, Jesu, komm	J. S. Bach
	Lobet den Herrn, all Heiden	J. S. Bach
**	Magnificat	J. S. Bach
	Singet dem Herrn ein neues Lied	J. S. Bach
**	Symphony No. 9	Ludwig von Beethoven
	Brazilian Psalm	Jean Berger
**	Hymn to St. Cecilia	Benjamin Britten
	Jesu, meine Freude	Dietrich Buxtehude
	In the Beginning	Aaron Copland
**	A Jubilant Song	Norman Dello Joio
**	The Mystic Trumpeter	Norman Dello Joio
	Missa Brevis	Antal Dorati
	Psalms	Lukas Foss
**	Canticle of the Martyrs	Vittorio Giannini
**	Messiah	George Fredrick Handel
**	The Creation	Franz Joseph Haydn
**	The Seasons	Franz Joseph Haydn
	The Last Days	Walter May

** Elijah	Felix Mendelssohn
Gloria	Francis Poulenc
** Messe en sol majeur	Francis Poulenc
** Magnificat	Johann Pachelbel
** Mass in G	Franz Schubert
Symphony of Psalms	Igor Stravinsky
** Gloria	Antonio Vivaldi
** Requiem	Giuseppe Verdi

* Selections analyzed to determine the intervalic content and the percentage frequencies.

** Selections from which melodic exercises were extracted, some of which were used in the various forms of the test.

APPENDIX E

Otterstein-Mosher Sight-Singing Test and Directions

TABLE II

Reliability Coefficients

	Method	r	Brown's Formula
Tone	O-E Exercises	0.959	0.979
	O-E Measures	0.992	0.996
Rhythm	O-E Exercises	0.878	0.935
	O-E Measures	0.966	0.982
Comparable Exercises ..	....	0.977	.....

Uses of the Test

The use to which the O-M Sight-Singing Test can be put is that of measuring the reading achievement of college students. This test will measure objectively an element of musical training which is highly essential to successful teaching alike in band and orchestra and vocal fields. For example, an entering student may be given the test to determine his or her placement in singing classes. Then again by employing at one stage the odd (even) exercises and at another the even (odd) the test may serve to measure the growth of the student through taking sight-singing courses. It also may be used as an entrance test for college freshmen planning to enter the music department to determine amount of sight-singing training.

Dangers in the Use of the Test

The outstanding difficulty with the test is its administration. The objectivity of scoring has been established (see Table I), but the data secured were in large measure dependent upon following explicitly the method described. It occurs to the authors that there may be inadvertent deviations which will affect the reliability of results. Prompting or other indication of approving favorable responses will elevate the scores. Likewise, showing irritation at wrong utterances may seriously affect the singer's performance and result in a general depression of the scores. The precaution of following directions strictly will obviate this apparent weakness. Experience with the test will tend to eliminate the personal equation of the examiner.

MANUAL OF DIRECTIONS

O-M SIGHT-SINGING TEST

By

ADOLPH W. OTTERSTEIN

Head of Music Department, San Jose State Teachers College, California

RAYMOND M. MOSHER

Professor of Psychology, San Jose State Teachers College

Stanford University Press, *Publishers*

THE O-M SIGHT-SINGING TEST is an individual test for the purpose of determining a student's achievement in reading music at sight. The items of the test were selected so as to be successively greater in difficulty, as to both rhythm and tone. The items advance scalewise, chord jumps being introduced in the ninth exercise, assuming (what is not proved) that the scale is the easiest to read and is the basis for singing the more difficult exercises. The test items progress through most of the major keys and introduce the minor mode. Some of the exercises in the latter part of the test begin on scale steps other than the tonic. The last two or three exercises in the test are taken from George Wedge: *Ear-Training and Sight-Singing*, Book II, by permission of the publishers, G. Schirmer, Inc., New York City.

Words are not used in the test. It is designed to measure pitch discrimination and rhythm only. Tone quality does not enter into its grading. There is a place for measure of this ability. Reading of pitch and rhythm are of primary technical importance. Without this knowledge, a student playing, for example, the violin, would not be able to play in tune. This is commonly expressed by the saying that music must first be heard as matters of pitch and rhythm, technically speaking, before it can be transposed on an instrument—piano, violin, voice, etc. This test is to measure the acquired ability of hearing a written score.

Administration of the Test

Thorough familiarity with the test is a prime requirement of the examiner. Almost any music teacher or supervisor can meet this essential by preliminary study of the test.

Before proceeding with the test proper, the examiner should hand a copy of it to the student to fill in the spaces for name, etc., or else the examiner quickly does that himself, asking for such information as he needs. This blank, with the student's name upon it, is then used by the

Copyright renewed 5/4/60 by [illegible]

examiner for recording the student's performance, the student reading from another copy.

The examinee should stand while taking the test. The examiner should be seated at the piano in order to play the chords (tonic) for establishing the pitch of the exercises in turn.

The directions are first read aloud by the examiner while the student reads silently. For reasons given below under Reliability, only the odd (or the even) exercises need be employed in a single test. If a retest is desired the even exercises may be used if the odd-numbered items are used the first time.

While the examinee sings, the examiner listens attentively and marks the exercises as follows:

1. Each measure being the unit for *tone* and *rhythm*,
   a) if a tonal error occurs in a measure, mark X or (—) just *above* the measure in which the error occurs;
   b) if a rhythmic error occurs in a measure, mark X or (—) just *below* the measure.
2. If an examinee stops, makes an error, and wants to sing an exercise over, permit him to do so; however, an error has been made and is to be so recorded.
3. Flatting or sharping on high or low tones is not considered a tonal error. The important thing is tonal relationship. Absolutely perfect vocal utterance in respect to pitch is not necessarily expected.

The examinee is permitted to continue until he has failed three complete exercises (both tonally and rhythmically).

Scoring the Test

The score is the sum of points earned on measures sung correctly. There are 28 exercises with 8 measures in each, making altogether 224 points for tone and 224 points for rhythm. Thus maximum score (tone plus rhythm) is 448.

Interpretation of Scores

For purposes of comparison, the following means for freshman college students have been computed:

No. Tested	*Mean*	*Standard Deviation*
42	Combined Score 305.52	83.4

Means are being prepared for music minors and general elementary students and will be available in the near future, also means for upperclass music majors.

Objectivity

The objectivity of the test was determined by having several examiners score the same singers. The test was scored by adding the number of wrong measures, both in pitch and rhythm.

TABLE I

DATA SHOWING OBJECTIVITY OF SCORING

Student	Scored by O.			By Ma.			By M.		
	P.	R.	Total	P.	R.	Total	P.	R.	Total
A	16	14	30	20	15	35	19	7	26
B	17	12	29	20	4	24	18	13	31
C				98	33	131	96	15	111
D	18	6	26	20	11	31	22	10	32
E	41	13	54	49	22	71	53	10	63
F	12	3	15	10	5	15	11	4	15
G	61	34	95	70	37	107	71	25	56
H	2	1	3	0	3	3	1	2	3
I	80	59	139	74	27	101			
J	95	35	130	99	31	130			
K	35	10	45	37	15	52	34	12	46

Validity

The validity of a test is of prime importance. If the test tests what it sets out to test, it is a useful measuring instrument. The aim here has been, therefore, to construct an instrument which in every possible way resembles the function in question. Thus the examinees sing here a series of exercises typical of a music class; the exercises severally correspond to those which might be involved in a singing lesson, and the vocal utterances are graded in accordance with the plan described in another section. Since the testing conditions resemble so closely the ordinary procedure of a music class, the validity is quite apparent.

Reliability

The reliability of the sight-singing test was established in the three following ways: First, by correlating the odd and the even exercises. Thus, Numbers 1, 3, 5, 7, 9, etc., were correlated with Numbers 2, 4, 6, 8, 10, etc., giving an *r* of 0.935 for rhythm and 0.979 for pitch. Second, by correlating the odd and even measures an *r* of 0.982 for rhythm and 0.996 for pitch was obtained, both correlations being stepped up by use of Brown's formula. And, third, by correlating comparable exercises, the table below gives the coefficients.

In reviewing these coefficients of reliability, one sees that the three attacks on the problem have all yielded very superior *r*'s. It would appear, therefore, that the measuring instrument as used offers a good tool to departments of school music in testing sight-singing achievement.

The correlations between sight-singing scores and other functions may be found in the table below. The data appear as inter-correlations which will be used in another study now in progress.

Published by
STANFORD
UNIVERSITY
PRESS
Stanford University
California

O-M SIGHT-SINGING TEST

By
ADOLPH OTTERSTEIN, B.S.
Head of Music Department, San Jose State Teachers College, California
RAYMOND M. MOSHER, Ph.D.
Professor of Psychology, San Jose State Teachers College

SCORE

Tone ....
Rhythm ....
Total ....
....

Name .... Date ....
(First name, initial, and last name)

Age at last birthday .... years. Birthday ....
(Month and day)

School .... City ....

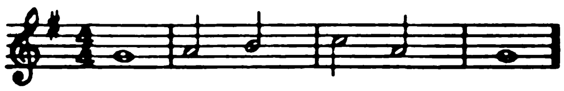
Do you play or sing? .... Name instrument or voice ....

How long have you studied music? ....

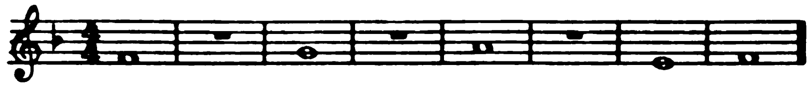
DIRECTIONS (*To be read aloud by Examiner while the student reads silently*):

1. The purpose of this test is to determine your ability to sing at sight. You will be graded on two points—pitch and rhythm. Tone quality does not matter; so if the note seems a little high when taking the test, and the quality of the tone is not pleasing, remember that does not affect your score. Do the best you can.
2. Set your own tempo. We advise setting a slow tempo so that you will have a better chance of singing accurately.
3. In each exercise the key will be established for you. For example, if the exercise is in G, I will play the tonic chord in the key of G (*examiner plays the broken chord to give the feeling of tonality*), during which you may hum "*do, mi, sol*" if you so desire.
4. If *sol. fa* syllables are a help to you, use them. If not, use neutral syllables. Remember you will be graded on pitch and rhythm.
5. To illustrate, look at the sample just below on this sheet. The exercise is in the key of G. I will sound the G tonic triad to establish the key and then you sing the exercise. (*The exercise is played or sung in strict rhythm.*)
6. Do you understand? All ready for the first exercise. This is the tonic chord. (*Examiner plays it.*) All ready—go.

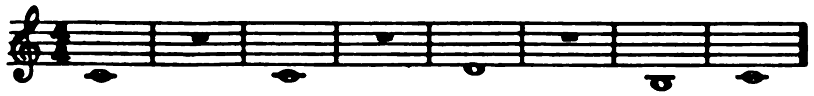
Sample:



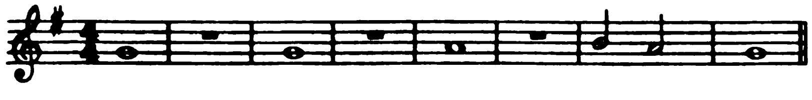
1

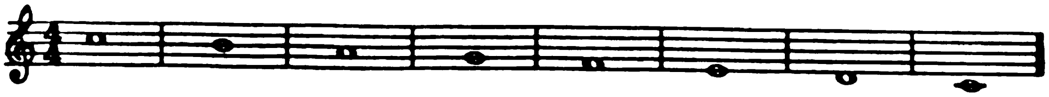


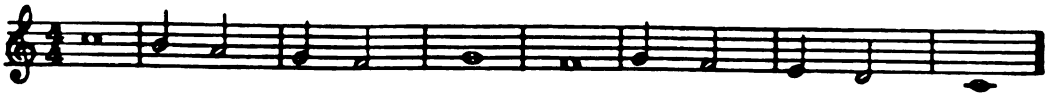
2

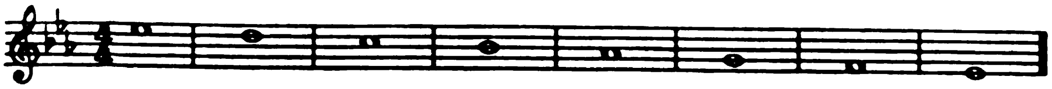


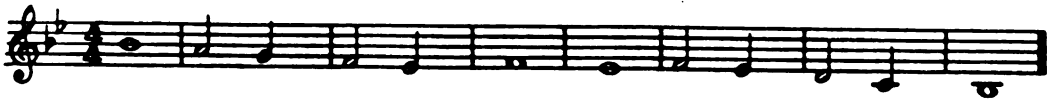
3

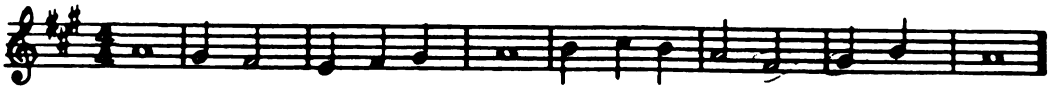


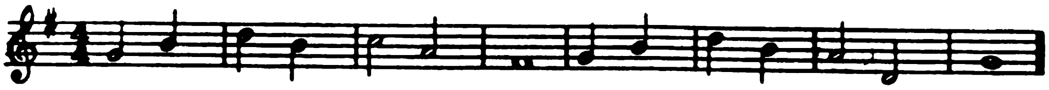
4 

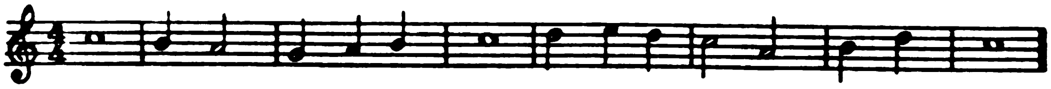
5 

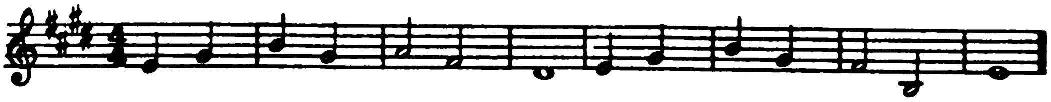
6 

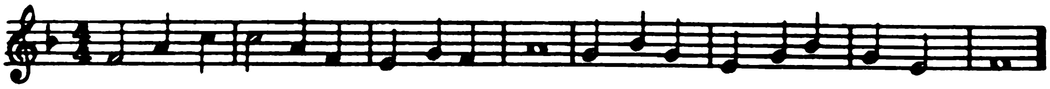
7 

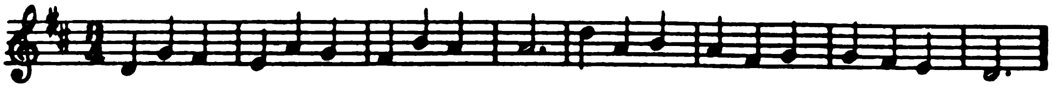
8 

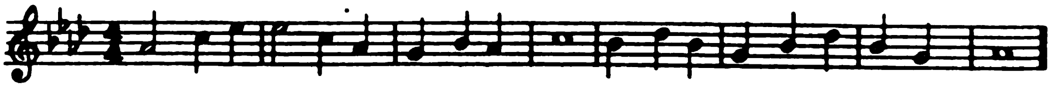
9 

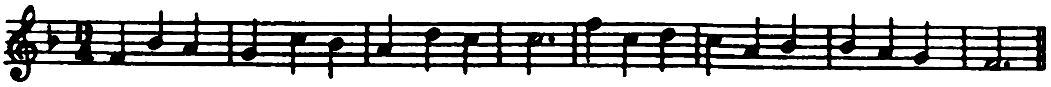
10 

11 

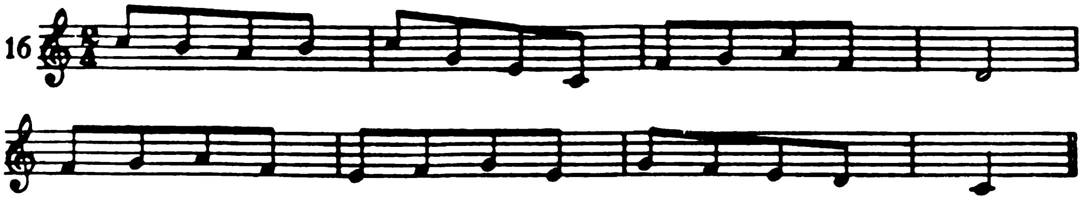
12 

13 

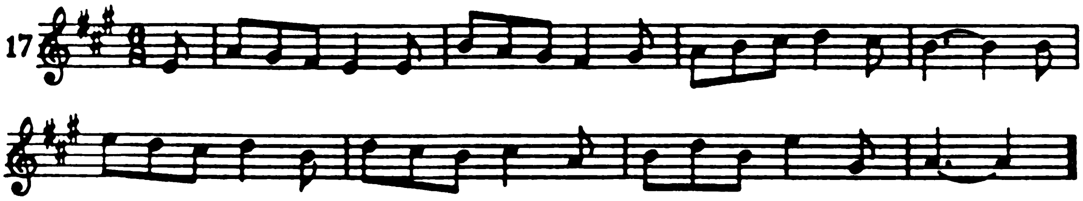
14 

15 

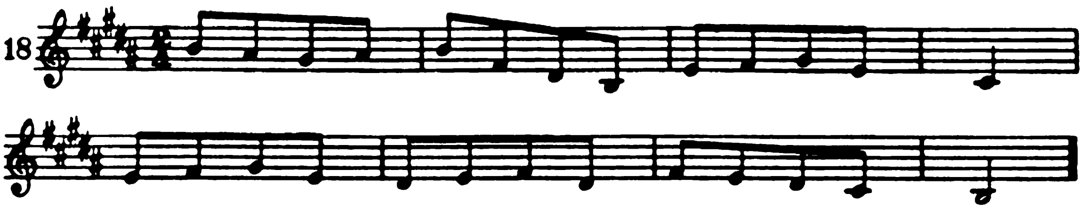
16



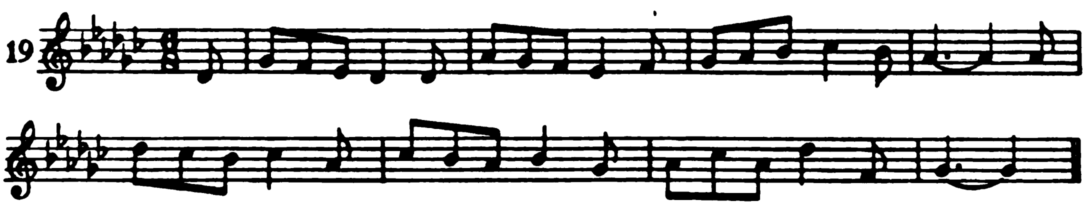
17



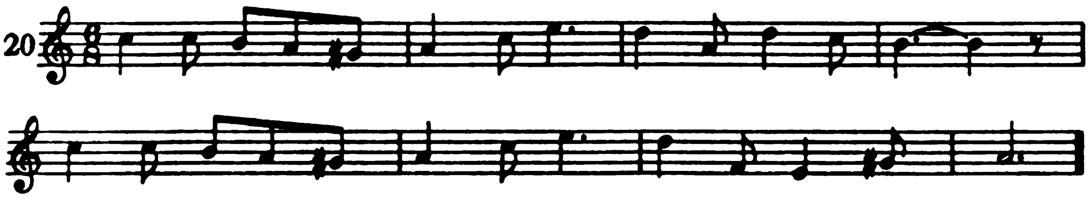
18



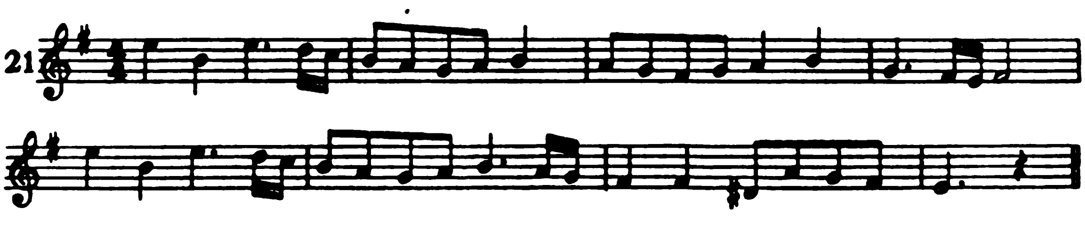
19

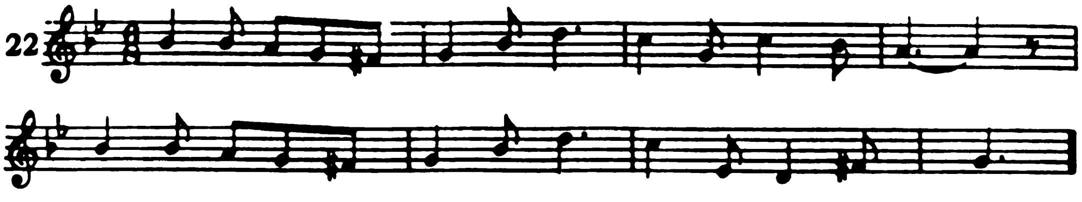


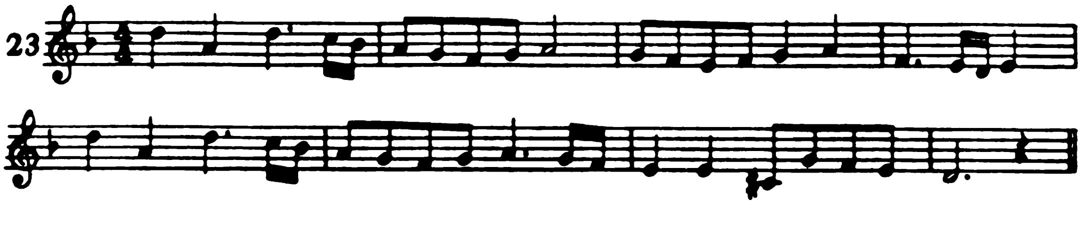
20

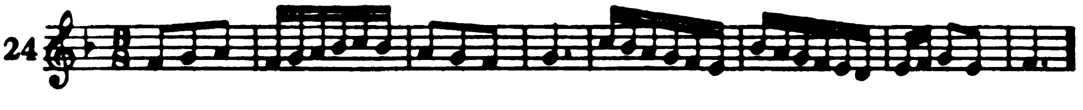


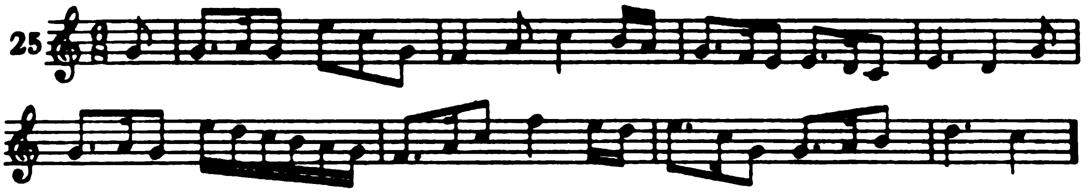
21

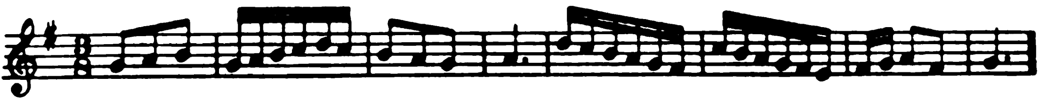


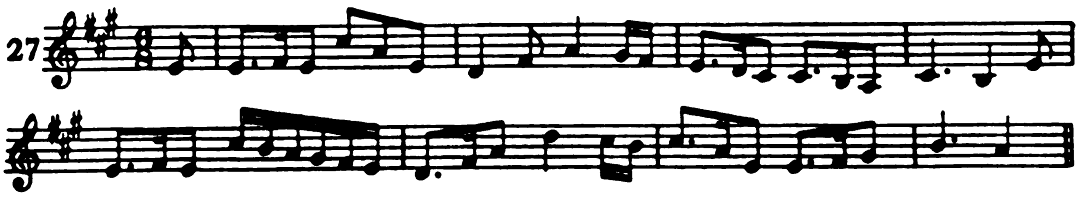
22


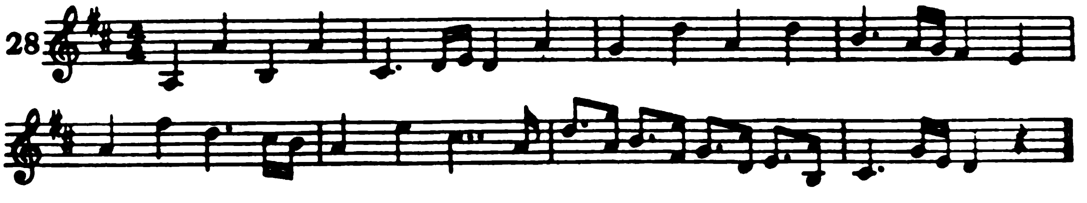
23


24


25


26


27


28


Numbers 21, 25, and 28, also Numbers 23 and 27 (transposed), from the George Wedge, *Ear-Training and Sight-Singing*, Book II, are here employed by special permission of the publishers, G. Schirmer, Inc., New York City

APPENDIX F

Tables from the Pilot Study

I. Pearson r Between the Unirhythmic Test and the Exercises

II. Interval Difficulty and Ranking

III. Pearson r Between Intervals and Unirhythmic Test Total

IV. Reliability Coefficients for the Interval Types

V. Pearson r Between Unirhythmic Test and Experience

Table F-1

Pearson r Between the Unirhythmic Test and the Exercises

Exercise	New Order	Coefficient
1	1	.556
2	2	.583
3	3	.586
4	6	.562
5	5	.661
6	9	.679
7	7	.668
8	10	.757
9	11	.645
10	4	.749
11	15	.688
12	13	.658
13	19	.636
14	16	.602
15	20	.581
16	8	.626
17	18	.712
18	17	.766
19	12	.698
20	14	.810

Table F-2

Interval Difficulty and Ranking

Interval	$\overline{X}$	Rank	*Frequency
Unison	.792	1	5
Minor second	.210	8	14
Major second	.302	4	24
Minor third	.253	5	12
Major third	.306	3	8
Perfect fourth	.246	6	7
Tritone	.126	10	2
Perfect fifth	.220	7	12
Minor sixth	.165	9	4
Major sixth	.079	12	3
Minor seventh	.094	11	2
Major seventh	.042	13	1
Perfect octave	.545	2	3

*number of times presented in the unirhythmic test.

Table F-3

Pearson r Between Intervals and Unirhythmic Test Total

Interval Type	Coefficient	*Rank
Unison	.568	12
Minor second	.901	4
Major second	.946	1
Minor third	.907	3
Major third	.841	6
Perfect fourth	.852	5
Tritone	.669	9
Perfect fifth	.914	2
Minor sixth	.679	8
Major sixth	.730	7
Minor seventh	.575	11
Major seventh	.547	13
Perfect octave	.662	10

*rank of the correlation coefficients; the highest coefficients are denoted by the smallest numbers.

Table F-4

Reliability Coefficients for the Interval Types

Interval	Coefficient Alpha
Unison	.730
Minor second	.829
Major second	.895
Minor third	.790
Major third	.735
Perfect fourth	.727
Tritone	.850
Perfect fifth	.788
Minor sixth	.503
Major sixth	.728
Minor seventh	.779
Major seventh	*
Perfect octave	.848
All intervals	.937

* the reliability coefficient for this interval was not computable without inappropriate data manipulation.

Table F-5

Pearson r Between Unirhythmic Test and Experience

Experience Variable	Coefficient	p level
Grade in school	.005	.473
Musical activities	.266	.001
Choral experience	.420	.001
Private vocal instruction	.053	.234
Instrumental experience	.059	.207
Private Instrumental Instruction	.302	.001
Keyboeard Training	.367	.001

APPENDIX G

Additional Tables from the Present Study

I. ANOVA Summary Between Groups for Choral Experience
Scheffe Comparisons for Choral Experience

II. ANOVA Summary Between Groups for Private Vocal Instruction
Scheffe Comparisons for Private Vocal Instruction

III. ANOVA Summary Between Groups for Instrumental Experience
Scheffe Comparisons for Instrumental Experience

Table G-1

ANOVA Summary Between Groups for Choral Experience

Sources of Variance	SS	df	MS	F
Between Groups	399.56	2	199.78	* 12.45
Within Groups	818.53	51	16.05	
Totals	1218.09	53		

* significant at the .01 level of confidence

Table G-1a

Scheffe' Comparisons Between Groups for Choral Experience

Comparison	** F	*** Confidence Interval
High School - College	* 16.19	5.45 ± 2.85
High School - Small College	* 23.30	4.88 ± 3.06
Small College - Large College	1.62	1.92 ± 3.80

* significant at the .01 level of confidence

** The tabled F statistic for 2 and 50 degrees of freedom is 3.18 at the .05 level. After mulitplying this figure by the number of groups minus one, the figure for comparison is 6.36. The figure for comparison at the .01 level is 10.12.

*** The confidence intervals have been calculated using the F statistic at the .05 level. The same are significant when calculated at the .01 level.

Table G-2

ANOVA Summary Between Groups for Private Vocal Instruction

Sources of Variance	SS	df	MS	F
Between Groups	93.88	2	46.94	* 16.47
Within Groups	145.45	51	2.85	
Totals	239.33	53		

* significant at the .01 level of confidence

Table G-2a

Scheffe' Comparisons Between Groups for Private Vocal Instruction

Comparison	** F	***Confidence Interval
High School - College	* 18.75	2.06 ± 1.20
High School - Small College	6.97	1.35 ± 1.29
Sm. College - Lg. College	* 14.27	3.90 ± 1.60

* significant at the .01 level of confidence

** The F statistic data is the same as that for Table 4-4 and 4-4a

*** The confidence intervals have been calculated at the .05 level. The first and third comparisons are significant. When calculated using the .01 level F statistic, only the third comparison is significant.

Table G-3

ANOVA Summary Between Groups for Instrumental Experience

Sources of Variance	SS	df	MS	F
Between Groups	246.62	2	123.31	* 4.69
Within Groups	1341.31	51	26.30	
Totals	1587.95	53		

* significant at the .025 level of confidence

Table G-3a

Scheffe' Comparisons Between Groups for Instrumental Experience

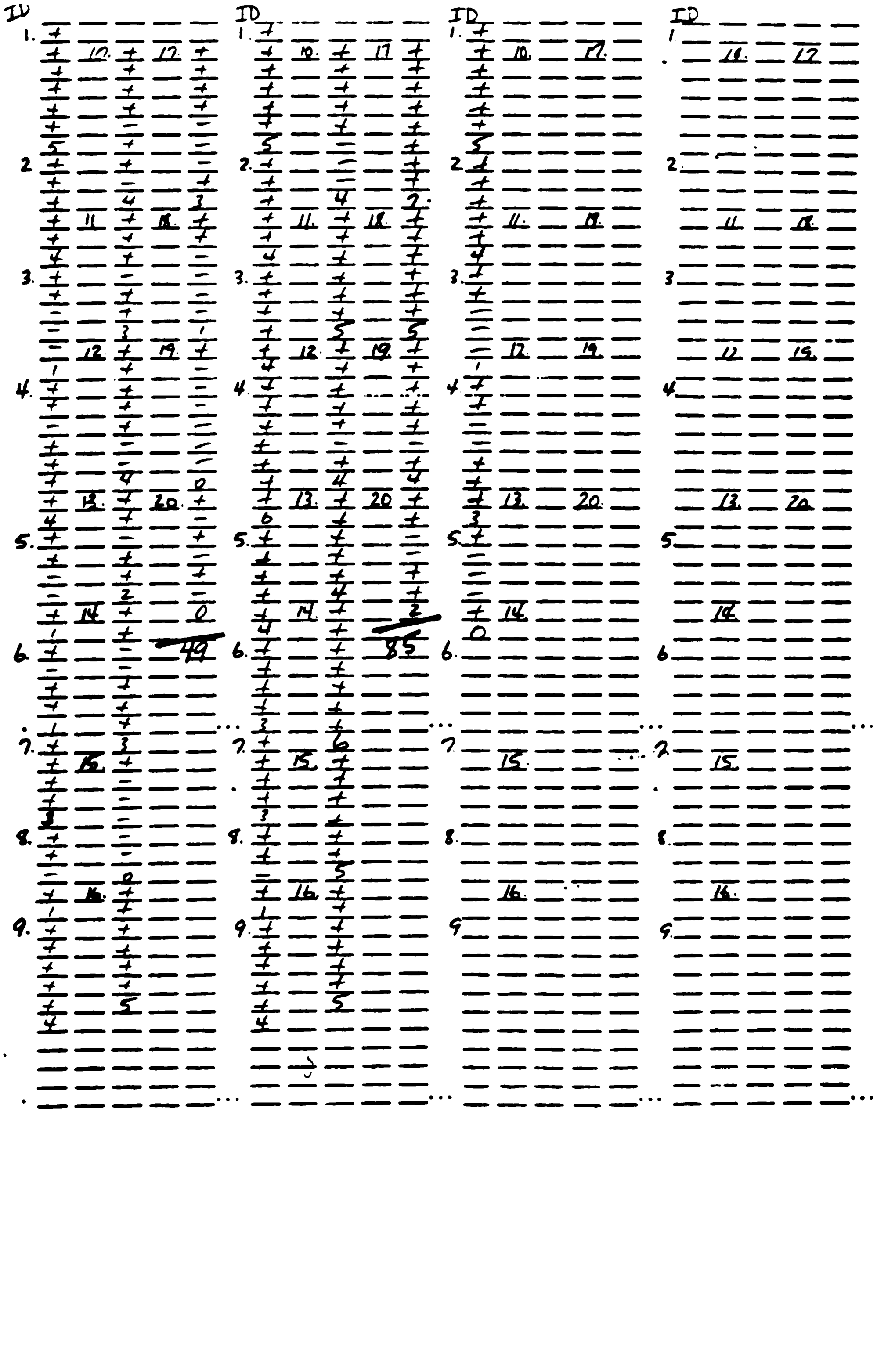
Comparison	* F	** Confidence Interval
High School - College	3.04	2.52 ± 3.65
High School - Sm. College	.49	1.09 ± 3.92
Sm. College - Lg. College	6.34	4.86 ± 4.87

* The F statistic for comparison at the .05 level for the Scheffe' technique is 6.36. None of the above comparisons are significant at that level.

** The confidence intervals are calculated at the .05 level of confidence. They verify the findings of the F statistics.

APPENDIX H

Scoring Sheets and Directions for Scoring Unirhythmic Test


IV
ID
ID
ID
1.
2
3.
4.
5.
6.
7.
8.
9.
10.
11
12.
13.
14
15.
16.
17.
18.
19.
20.
49
85

INSTRUCTIONS FOR SCORING UNIRHYTHMIC TEST

1. Prior to scoring any of the subjects, play and sing each item several times to become acquainted with the pattern.

2. When scoring the first few subjects, listen to an item, score it, then listen again to insure that it has been scored correctly. As you score each item, sing the student's response (where possible) and score as you sing. After scoring several subjects the need to relisten will become less frequent.

3. To score the test, a plus (+) is indicated for each correct note sung by the subject. A minus (-) is recorded for each incorrect note. The score sheets are numbered and provide one more space than the number needed for each item. This is for clarity and to insure that the correct number of notes have been scored for each item. A quick perusal of the score sheet for the blanks can insure that the correct number of notes has been scored.

4. You are asked to use your own judgment concerning the sharpness or flatness of a note. If the note sung sound (to your ears) closer to the note written than some other note, then mark a plus. If not, a minus. Please do not use any mechanical device to differentiate between a correct or incorrect response. Use your ears.

5. A student may start, sing several notes, realize that he is in error, and ask to start again. If this occurs, the notes that have already been sung should be scored correct or incorrect depending upon the initial performance, not the repeat performance. On the repeat, however, you may score the remaining notes as he sings them, since these would not have been repeated.

6. Do not mark an error for the first note unless the subject has been given the starting pitch a fourth time. Each subject has been given the correct pitch and failure to match the initial pitch does not constitute an error, unless it has been given four times.

7. Although this is a unirhythmic test, the students may not sing all notes unirhythmically. Do not count an error for this as rhythm is not to be scored.

8. As you score the subjects, please indicate their appropriate number as letter code at the top of each set of columns. This will insure that the correct responses are associated with the correct student.

9. Occasionally a student will repeat a note to retune or to catch a breath as it were. Where this is obvious, do not mark an incorrect response unless the initial note sung was incorrect.

10. Thank you very much for your cooperation and assistance.

APPENDIX I

Raw Scores for the Three Measurement Devices

Raw Scores of the Entire Sample for the Three Measurement Devices

Subject	OM1	OM2	U1	U2	SR
1	305	295	37	38	5
2	421	421	70	70	7
3	329	322	39	39	6
4	320	316	56	64	5
5	432	434	91	92	9
6	168	168	17	16	3
7	224	217	34	34	4
8	276	271	57	58	6
9	362	358	52	54	5
10	385	375	66	67	6
11	292	281	37	36	5
12	415	415	73	74	7
13	292	269	44	44	4
14	251	251	28	26	3
15	322	321	42	42	2
16	408	408	73	73	7
17	302	295	44	46	3
18	308	297	32	34	6
19	131	125	13	13	1
20	424	424	88	89	8
21	413	410	78	79	8
22	394	386	71	69	8
23	411	405	82	83	7
24	423	419	88	89	8
25	435	431	83	84	8
26	400	393	66	67	6
27	308	307	60	63	4
28	435	437	95	96	9
29	443	440	94	95	9
30	381	381	59	60	7
31	388	381	76	76	6
32	240	238	39	40	2
33	387	385	86	87	8
34	260	242	41	38	4
35	426	426	87	90	8
36	267	261	46	47	4
37	289	281	47	47	4
38	289	286	27	24	2
39	249	22	29	30	3

Raw Scores (Cont'd)

Subject	OM1	OM2	U1	U2	SR
40	173	159	18	18	1
41	302	300	52	52	4
42	257	251	24	23	3
43	177	173	28	28	2
44	277	269	50	51	3
45	268	265	33	31	3
46	206	204	34	34	3
47	287	296	35	35	5
48	324	316	45	45	5
49	401	387	62	63	6
50	429	430	88	89	9
51	428	428	88	89	8
52	405	405	79	80	8
53	347	347	68	69	6
54	287	285	40	39	4

OM1 and OM2 correspond to the first and second scorings of the Otterstein-Mosher test.

U1 and U2 correspond to the first and second scorings of the Unirhythmic test.

SR was the subjective rating.

Subjects 1 - 20 were the high school students.

Subjects 21 - 30 were the university students.

Subjects 31 - 54 were the students from the small colleges.