AN INVESTIGATION OF THE OPTIMUM LENGTH OF MUSICAL APTITUDE PROFILE SUBTESTS

by

Merrill Edwin Brown

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Music in the Graduate College of the University of Iowa

June, 1967

Thesis supervisor: Associate Professor Edwin Gordon
Thesis co-supervisor: Professor Leonard Feldt

The major problem of the study was to determine the optimum length of each Musical Aptitude Profile subtest with regard to reliability. Ancillary problems, all relating to the Musical Aptitude Profile, were to: 1) estimate its test-retest reliability, 2) investigate the effects of practice in taking the tests, 3) compare score distributions of parochial school students with those established for public school students in the national standardization program, and 4) determine the appropriateness of the Spearman-Brown Prophecy Formula for use with the test battery.

During the 1966-67 school year eight hundred and ninety-six students in Catholic parochial schools in Kenosha, Wisconsin, were tested and then retested on the test battery. This total included 331 fifth grade students, 307 seventh grade students, and 258 ninth grade students. The subjects represented all of the fifth and seventh grade classes in five elementary-junior high schools and the entire ninth grade class at one high school.

The optimum length of each of the seven subtests of the battery was determined separately for each grade level by subdividing each subtest into "test portions" of five items each. The test-retest reliability coefficient for each sequentially lengthened "test portion" was computed.

Resultant data indicated that, with very few exceptions, each subtest yielded higher reliability coefficients as "test portions" were added at each grade level tested and that the reliability of each subtest was at its maximum when the entire subtest had been administered.

The effects of practice in taking the Musical Aptitude Profile on retaking the battery were determined by comparing the eleven mean standard scores derived from the initial test administration with the corresponding mean scores derived from the retest. It was found that all mean score differences were small and that none was significant at the five percent level of confidence.

A comparison of split-halves reliability coefficients, standard deviations, and standard errors of measurement, between the parochial school students and public school students revealed negligible differences. Percentile rank norm differences between the parochial and public school students were very small.

The accuracy with which the Spearman-Brown Prophecy Formula predicted the reliability of subtests of increased lengths was determined for each of the seven subtests in the battery for students in each of the three grade levels. Reliability coefficients predicted through the use of the formula were very similar to, but generally higher than, those derived empirically. The differences varied from test to test and from grade to grade but no definite pattern was

established. Overall, the test-retest reliability coefficents were found to be lower than corresponding split-halves reliability coefficients.

Based on the data derived from this research it is evident that optimum reliability of each Musical Aptitude Profile subtest is realized only when all "test portions" are cumulated. Therefore, if any of the tests were shortened, reliability would be sacrificed. Actually, the findings even suggested that perhaps the reliability of the subtests might continue to increase if more test items were added.

Finally, it can be concluded that the normative data reported in the test manual is appropriate for use with parochial and public school students and for students who have previously been administered the test.

Abstract approved: Edwin Gordon, thesis supervisor

Associate Professor
Music Education
title and department

May 15, 1967
date

TABLE OF CONTENTS

Chapter		Page
I	PURPOSE OF THE STUDY	1
	Introduction	1
	Purpose of the Study	2
	Problems of the Study	3
	Description of the Musical Aptitude Profile	3
II	REVIEW OF RELATED STUDIES	8
	Introduction	8
	The Lanier Study	8
	The Larson Study	11
	The Gordon Study	19
	Comparison of Related Studies With Present Study	22
III	DESIGN OF THE STUDY	26
	Subjects	26
	Testing Procedure	26
	Data	28
	Normative Data	28
	Test-Retest Reliability	28
	Determination of the Optimum Length of Each Subtest	29
	Evaluation of the Practice Effects of Taking the Tests	31
	Appropriateness of the Spearman-Brown Prophecy Formula	31
IV	PRESENTATION AND INTERPRETATION OF DATA	34
	Standard Score Means, Standard Deviations, Standard Errors of Measurement, and Mean Differences	34
	Split-Halves Reliability	38
	Percentile Norm Data	41
	Test-Retest Reliability	45
	Optimum Length of Subtests	47

TABLE OF CONTENTS (cont'd)

Chapter		Page
	Effect of Practice in Taking the *Musical Aptitude Profile*	59
	Appropriateness of the Spearman-Brown Prophecy Formula	63
V	SUMMARY AND CONCLUSIONS	88
	Problems of the Study	88
	Design	88
	Results	90
	Conclusions	93
	BIBLIOGRAPHY	95

Chapter I

PURPOSE OF THE STUDY

Introduction

It is common practice in schools in the United States to adapt instruction to meet the individual needs and abilities of students in subjects such as reading and mathematics. With the exception of selecting students for participation in special performance groups, little consideration has been given to providing for individual differences in the teaching of music.

The subtests of the Musical Aptitude Profile[1] are primarily intended to furnish objective information about various musical aptitudes of students in grades four through twelve so that teachers can efficiently provide for individual differences. In addition, test results can be used to encourage musically talented students to participate in music performance organizations, to help students formulate educational plans in music, and to provide parents with objective information about their child's musical talent.

Because one of the primary purposes of MAP* is to

---

1. Edwin Gordon, Musical Aptitude Profile (Boston: Houghton Mifflin Company, 1965).

*For ease of presentation the Musical Aptitude Profile will be referred to as MAP.

evaluate each student's specific musical strengths and weaknesses, the battery is comprised of seven reliable subtests. As a result, MAP requires more time to administer than that demanded by a "talent" test on which a student may demonstrate only "overall" musicality. It has been suggested that the administration time of the battery be shortened by reducing the number of items in each subtest if, in fact, this could be accomplished without sacrificing or limiting the effective use of MAP results.

Purpose of the Study

The length of a test is known to effect its reliability. Generally, as the number of items in a test increases, its reliability also increases. However, a point may be reached at which its reliability may well decrease. Factors such as frustration, fatigue, or loss of interest in taking the test may cause an examinee to make random responses, which will, in turn, lower the test's reliability. For a test to be most efficient, it must offer maximum reliability in a minimal amount of administration time.

In an attempt to determine the optimum length of each subtest of MAP, Gordon administered a preliminary version of his battery to approximately 475 musically select junior high students in Wausau, Wisconsin. He found that test reliabilities increased as items were added.[2] However, a limiting

---

2. Edwin Gordon, Manual, Musical Aptitude Profile (Boston: Houghton Mifflin Company, 1965), 73-79.

factor of the study was that a homogeneous group of students was used in the investigation. Further, the Wausau study was conducted before MAP was in its final form. The purpose of the present research was to investigate -- under more favorable conditions than prevailed in the Wausau study -- the effect of variations in test length on the reliability of MAP. It was hoped that some conclusion might be drawn concerning the optimum length of the individual subtests.

Problems of the Study

The major problem of this study was to determine the optimum length of each subtest constituting MAP. Ancillary problems of the study were to: 1) estimate the test-retest reliability of MAP, 2) investigate the practice effects of taking MAP a second time, 3) compare the score distributions of parochial school students with those of students who participated in the national standardization program, and 4) determine the appropriateness of the Spearman-Brown Prophecy Formula for use with MAP.

Description of the Musical Aptitude Profile

Through the use of MAP, a student's musical aptitude is evaluated by both his objective and subjective responses to musical stimuli. This is an approach similar to that which music psychologists refer to as the Gestalt method of music test construction. This approach is different from the other main philosophy of musical aptitude testing, of which

Carl Seashore was the leading exponent, which follows the point of view that musical talent can be best evaluated by testing an individual's sensory acuity to the component parts of the sound wave: pitch, timbre, loudness, and time.

Over an eight year period, Gordon constructed test items which were administered to both public school students and professional musicians. From the results of item analysis data, items were revised and additional items were developed and evaluated. A total of five revisions of the original battery was made before the final test was published. Approximately 15,000 students participated in the pre-publication research reported in the test manual.[3]

During the 1964-65 school year, MAP was nationally standardized in a carefully planned program which included a representative sample of students in grades four through twelve. More than 12,000 students, representing twenty public school systems in eighteen states, were administered the complete test battery. The standardization program is fully described in the test manual.[4]

The basic musical factors which are measured by MAP are grouped into three main divisions: Tonal Imagery, Rhythm Imagery, and Musical Sensitivity. The former two divisions are non-preference tests and are subdivided into Melody and Harmony, and Tempo and Meter, respectively. The latter

---

3. Ibid., 12-23.

4. Ibid., 44-47.

division, Musical Sensitivity, is a preference test and it is subdivided into Phrasing, Balance, and Style.

There are forty items in each of the four non-preference subtests and thirty items in each of the preference subtests, making a grand total of two hundred and fifty items for the entire battery. Each item consists of a short musical selection with a musical answer; students are asked only to decide whether the selection and answer are alike or different, exactly the same or different, or to decide which of two renditions is indicative of a more musical performance. If a student is not sure of the answer, he is instructed to mark the question mark (?) column, indicating that he is "in doubt." Because MAP is an aptitude test, there are no questions concerning historical or technical facts about music.

Other important and unique aspects of the test are: 1) all test items are composed by the test author especially for use in the test battery, 2) string instruments are used as the performing media, and 3) professional artists* perform the selections. The entire test battery, which is recorded on high fidelity magnetic tape, requires one hour and fifty minutes of actual testing time. The battery is designed so that each of its three divisions may be administered within the time limits of a regular class period.

---

*The players are violinists Stuart Canin and Charles Treger, and the cellist is Paul Olefsky.

Eleven test scores are derived from the test battery: one for each subtest, a total score for each of the three main divisions, and a composite score for the complete battery. Electronic scoring is available for computing scores or the answer sheets may be scored manually.

Percentile norms for each of the eleven test scores are furnished for each school grade from four through twelve. In addition, percentile norms are provided for each of the eleven test scores for students participating in music performance organizations at the elementary school level (grades 4, 5, and 6), the junior high school level (grades 7, 8, and 9), and the high school level (grades 10, 11, and 12).

Split-halves reliability coefficients, adjusted for length through the use of the Spearman-Brown Prophecy Formula, are provided for each grade from four through twelve for each of the eleven scores of the test battery. The reliabilities vary somewhat from grade to grade and from test to test, but are generally in the .70's and .80's for the subtests, in the .80's and lower .90's for the total tests, and in the range from .90 to .96 for the composite test.[5] Lee,[6] in an investigation of the use of MAP with college and university students, found that for older students the reliability coefficients are comparable to those found for students

---

5. Ibid., 16.

6. Robert E. Lee, "The Adaptation of the Musical Aptitude Profile for College and University Students" (Unpublished Ph. D. dissertation, University of Iowa, 1966), 41.

in grades four through twelve.

Many investigations, including diagnostic validity studies, have been conducted to determine various aspects of validity of MAP. A major study, a three-year longitudinal investigation pertaining to the predictive validity of the test battery, was completed in June, 1966. The validity of MAP as a predictor of judges' evaluation of instrumental tape-recorded performances, musical achievement test scores, and teacher ratings were determined. When all of the validity criteria were combined and then correlated with the Tonal Imagery, Rhythm Imagery, and Musical Sensitivity tests, the within-school coefficients were found to be .60, .60, and .62, respectively. The predictive validity of the MAP composite score, using the combined unweighted validity criteria, was found to be .75.[7]

---

7. Edwin Gordon, A Three-Year Longitudinal Predictive Validity Study of the Musical Aptitude Profile (Iowa City, Iowa: University of Iowa, In preparation).

Chapter II

REVIEW OF RELATED STUDIES

Introduction

Separate studies conducted by Lyle H. Lanier,[8] Ruth C. Larson,[9] and Edwin Gordon[10] are closely related to the present study in that they included an investigation of the optimum length of a musical aptitude test. The first two studies deal with the 1919 edition of the Seashore Measures of Musical Talent and the third study with a pre-publication version of MAP.

The Lanier Study

Lyle H. Lanier, while investigating the reliability of mental tests and tests of special abilities, conducted research involving the 1919 edition of the Seashore Measures of Musical Talent. Lanier tested 106 college students in October, 1924, on the Seashore Pitch, Intensity, and Time

---

8. Lyle H. Lanier, "Prediction of the Reliability of Mental Tests and Tests of Special Abilities," Journal of Experimental Psychology X/2 (April, 1927), 69-113.

9. Ruth C. Larson, "Studies on Seashore's Measures of Musical Talent," (University of Iowa Studies: Series on Aims and Progress of Research II/6. Iowa City, Iowa: University of Iowa Press, 1930), 16-33 and 77-79.

10. Edwin Gordon, Unpublished research report furnished the writer, 10-13.

tests and retested the students in March, 1925. Two administrations of the Consonance, Tonal Memory, and Rhythm tests were conducted with another group of 109 college students under similar conditions.

For the Pitch, Intensity, and Time tests, each student recorded one hundred responses pertaining to pairs of sounds produced on a record player. For the Consonance, Tonal Memory, and Rhythm tests, fifty responses were made for each test. Each of the six tests was then subdivided into ten parts, or "test portions." Thus, the former three tests provided ten items in each of the ten "test portions" and the latter three tests provided five items in each of the ten "test portions." Reliability coefficients, determined by correlating cumulative scores on "test portions" with corresponding scores on the "retest portions" derived four and one-half months later, are presented in Table 1.

Maximum reliability for the Pitch, Intensity, and Rhythm tests was reached or surpassed before all "test portions" were cumulated. For the Time, Consonance, and Tonal Memory tests, maximum reliability was not reached until all "test portions" were cumulated. Reliabilities for the Time and Consonance tests, however, decreased after the cumulation of fifty and sixty percent of the items, respectively, and then increased until all items were cumulated. The results indicated that the Pitch and Rhythm tests could be shortened by one-half and the Intensity test by

Table 1.

The Lanier Study: Seashore Measures of Musical Talent Cumulative Reliability Coefficients for College Students.[11]

Number of Items Cumulated	N=106 Pitch	N=106 Intensity	N=106 Time	Number of Items Cumulated	N=109 Consonance	N=109 Tonal Memory	N=109 Rhythm
10	.416	.260	.298	5	.179	.332	.210
20	.600	.238	.303	10	.158	.430	.268
30	.528	.495	.287	15	.353	.540	.385
40	.666	.306	.303	20	.415	.529	.443
50	.675	.438	.416	25	.466	.570	.426
60	.628	.513	.361	30	.510	.577	.369
70	.680	.506	.408	35	.500	.596	.373
80	.652	.603	.445	40	.470	.605	.340
90	.676	.573	.477	45	.461	.637	.350
100	.680	.597	.495	50	.543	.669	.425

11. Lanier, op. cit., 90-95.