As I understood my responsibility as a reader, it was to react to this document as a report of a research project. The following critique is concerned with the design and general quality of the project as it is revealed by this report. This means that any deficiencies in the study which are discussed here may arise from either inadequate procedures or inadequate reporting of procedures.

Before entering into this critique, I would hasten to add that these negative observations are made with reluctance. This reluctance arises from a high personal regard for the investigator, from a knowledge of the great need for this kind of endeavor, and from an understanding of the very formidable complexity of the task which the investigator has undertaken. If the following can be construed as a desire to help refine what is already basically valuable, it will be an appropriate interpretation.

Note it)

CRITIQUE

Format

As the report was read, it became increasingly apparent that the "test manual" referred to in the text should have been incorporated or appended to the document for reference purposes.

Concerning the style of reporting, a difficult time was experienced where several elements of the test were discussed in parallel rather than in packaged sequences - see pp. 11-12.

It could be that subheadings would clarify the outline of the report. Cues for the content of sections is frequently helpful in reading the dry, reportorial style, which this has to be.

The division of tables between the appendices and the text made the referrals to these confusing and time consuming.

Another difficulty was the lack of a clean statement of the hypothesis being tested and an overall statement of procedures. In short, the failure to follow a traditional design made this reader climb out of a rut and scramble about a bit.

Achievement Test Construction

The difficulties of test construction as an endeavor in general cause some major questions. These difficulties are found somewhat inadequately treated in the development of the achievement test. The validity of an item rests in the technique by which it was developed and refined. Because these techniques were not described, the validity of the modifications in the test must be questioned. To sum up the position of this reader, when an item in a test is changed in any way whatsoever, even its place in the order of all items, it must be treated as a new item and a completely new analysis of its potency and validity must be made.(See p. 30 and p. 44)

Too, the validity of conclusions will depend on a reliable and valid test instrument which remains constant throughout. The

three achievement tests may have been a continuum comprising one test instrument but there is no way of deriving this fact from the discussion of the tests.

Evaluative Methods

The first evaluative method included only one measure of aptitude. Since three measures were made, why did the investigator not analyze the correlation between aptitude and achievement scores obtained at the end of the study? Why were not comparisons made between achievement - first year aptitude scores and achievement-third year aptitude scores to show what changes occurred? Here was an apparent strength (in terms of this reader's cursory examination) which was not exploited.

The second method for assuring validity, "the provision of a relatively uniform and extensive program of training", is inadequately described. The schools provided materials for instruction but what were the materials and how valid are they as a means for the development of musical sensitivity? What is meant by uniform? What variables were accounted for in this uniformity? Were individual differences in teachers accounted for such that progress differentials were weighted? Were teaching techniques, time for teaching - learning, and teaching goals controlled or accounted for?

The reliability of two judges seemed to be quite high. The validity of their judgements as individuals of musical taste cannot be questioned. However, the validity of their judgments in terms of musical judgment as this may be normally distributed throughout the profession can be questioned. The use of two judges is the most serious deficiency in the entire design. Because this test of accomplishment is perhaps the most specific and potent validating device in the project, this deficiency has regrettable major implications. It may be that these two judges are an adequate sample but we do not know this.

Another difficulty is the number of times each etude was

rendered by the subjects. This part of the procedures was generally well done. The reliabilities might have been different in three renditions. The rebuttal to this criticism is of course, how far does one go using youngsters of this age and obtain a valid reading on them.

An inconsistency in the procedures for the evaluating teachers may have biased the values for or against particular students. In some situations, two teachers rated a student for his progress. In others, one teacher made the rating. If one accepts that two heads are better than one, then there are several degrees of validity in these judgments.

Evaluation Instruments

Concerning the tape-recorded student performances. There is no discussion of the validity of the musical etudes as musically valuable compositions. Because the judge's form required a value judgment concerning "expression", a prerequisite would be a model agreed upon by experts as the appropriate "expression" for each of these etudes. Too, there is no indication of criteria by which the judges made a value judgment concerning expression, which is an aesthetic problem of many aspects. This would not be a point to belabor if musical aptitude was not ultimately judged in terms of the ability to interpret or express music.

There is no indication on the judge's form or in the text as to what criteria the judges used for evaluating melody, rhythm, sight reading, etc..

The instructions for music teachers in their evaluation of students were very vague. What were the criteria? How were personal relations between student and teacher eliminated? Was the youngster who gained in technical, mechanical facility given a greater rating than one who could not perform with such facility but within his limited facility developed a greater expressiveness?

Predictability

There is a difficulty in accepting the idea of prediction when the study is based on a small population.

The investigator raises an old issue: can one predict the success of an individual in a music program. It is realized that line 15 on page 8 is phrased very carefully but its context gives this reader a very clear impression that the writer hints at the prediction of success. This should be clearly stated instead of backed into later on in the study. This claim and attempt of proof, of course, is regrettable. All the investigator can claim, assuming the test is valid, is that the test identifies musical ability. To claim any kind of predictive ability is to claim that identified ability properly nurtured, will mature. To predict success is to include in the prediction all the variables in native ability as well as environmental influences which will impinge on the organism.

Another major difficulty in the predictive aspects of the conclusions is the many revisions of test materials as the project proceeded. This created a problem of validity in the comparisons made between the years (see p. 55). Before comparisons between two versions of a test can be made, specific validating procedures must be used as in the development of the short form of a test. A prediction of success, it would seem to me, must be based on a single test administered at the beginning and at the end, holding aptitude constant.

Analysis of Data

The results of data analyses were not examined closely by this reader. This was considered a second level operation after questions concerning the procedures were resolved.

Conclusions of the Study

The conclusion of the study (see p. 58) may be warranted in

terms of the data reported but it is not warranted in terms of the procedures which yielded that data. It is possible that the relationship between achievement and this aptitude test is greater than found here, the problems cited on p. 58 not withstanding. The comparison of scores derived from different tests not validated as a continuum of difficulty may be the source of difficulty. The correlations reported on page 59 indicate some kind of common learnings but is this a validity or a reliability measure?

In addition, there are some interpretations of data which can be challenged. For example, to statistically account for a low aptitude in 23 dropouts and not account for the other variables influencing these dropouts leaves many questions unanswered.

Commentary

This reader has had considerable difficulty in expressing the preceeding reservations about the study. The candidness of the investigator's report has made possible these rather verbose reactions. Ironically, the aptitude test seems to be reliable and valid as it stands. This reader's difficulty has been in finding proof that it is in terms of the procedures used.

The major difficulty in developing a test of this kind is that not only must the test evolve but so must its validating procedures -- thesis and antithesis. The one is dependent on the other. It seems that the validating procedures (the achievement test, etc.) are developed for future use in this project. This alone is no small endeavor.

There are times when the behavioral sciences are more opinion than science. The investigator now has the further trouble of investigating the validity of this reader's opinions. They may not be so.

I apologize for having to do this under pressure and probably emphasizing the negative too much. There is much of value to Music Education endeavors in this study.

CRITIQUE

Format

As the report was read, it became increasingly apparent that the "test manual" referred to in the text should have been incorporated or appended to the document for reference purposes.

Concerning the style of reporting, a difficult time was experienced where several elements of the test were discussed in parallel rather than in packaged sequences - see pp. 11-12.

It could be that subheadings would clarify the outline of the report. Cues for the content of sections is frequently helpful in reading the dry, reportorial style, which this has to be.

The division of tables between the appendices and the text made the referrals to these confusing and time consuming.

Another difficulty was the lack of a clean statement of the hypothesis being tested and an overall statement of procedures. In short, the failure to follow a traditional design made this reader climb out of a rut and scramble about a bit.

Achievement Test Construction

The difficulties of test construction as an endeavor in general cause some major questions. These difficulties are found somewhat inadequately treated in the development of the achievement test. The validity of an item rests in the technique by which it was developed and refined. Because these techniques were not described, the validity of the modifications in the test must be questioned. To sum up the position of this reader, when an item in a test is changed in any way whatsoever, even its place in the order of all items, it must be treated as a new item and a completely new analysis of its potency and validity must be made.(See p. 30 and p. 44)

Too, the validity of conclusions will depend on a reliable and valid test instrument which remains constant throughout. The

three achievement tests may have been a continuum comprising one test instrument but there is no way of deriving this fact from the discussion of the tests.

Evaluative Methods

The first evaluative method included only one measure of aptitude. Since three measures were made, why did the investigator not analyze the correlation between aptitude and achievement scores obtained at the end of the study? Why were not comparisons made between achievement - first year aptitude scores and achievement-third year aptitude scores to show what changes occurred? Here was an apparent strength (in terms of this reader's cursory examination) which was not exploited.

The second method for assuring validity, "the provision of a relatively uniform and extensive program of training", is inadequately described. The schools provided materials for instruction but what were the materials and how valid are they as a means for the development of musical sensitivity? What is meant by uniform? What variables were accounted for in this uniformity? Were individual differences in teachers accounted for such that progress differentials were weighted? Were teaching techniques, time for teaching - learning, and teaching goals controlled or accounted for?

The reliability of two judges seemed to be quite high. The validity of their judgements as individuals of musical taste cannot be questioned. However, the validity of their judgments in terms of musical judgment as this may be normally distributed throughout the profession can be questioned. The use of two judges is the most serious deficiency in the entire design. Because this test of accomplishment is perhaps the most specific and potent validating device in the project, this deficiency has regrettable major implications. It may be that these two judges are an adequate sample but we do not know this.

Another difficulty is the number of times each etude was

rendered by the subjects. This part of the procedures was generally well done. The reliabilities might have been different in three renditions. The rebuttal to this criticism is of course, how far does one go using youngsters of this age and obtain a valid reading on them.

An inconsistency in the procedures for the evaluating teachers may have biased the values for or against particular students. In some situations, two teachers rated a student for his progress. In others, one teacher made the rating. If one accepts that two heads are better than one, then there are several degrees of validity in these judgments.

Evaluation Instruments

Concerning the tape-recorded student performances. There is no discussion of the validity of the musical etudes as musically valuable compositions. Because the judge's form required a value judgment concerning "expression", a prerequisite would be a model agreed upon by experts as the appropriate "expression" for each of these etudes. Too, there is no indication of criteria by which the judges made a value judgment concerning expression, which is an aesthetic problem of many aspects. This would not be a point to belabor if musical aptitude was not ultimately judged in terms of the ability to interpret or express music.

There is no indication on the judge's form or in the text as to what criteria the judges used for evaluating melody, rhythm, sight reading, etc..

The instructions for music teachers in their evaluation of students were very vague. What were the criteria? How were personal relations between student and teacher eliminated? Was the youngster who gained in technical, mechanical facility given a greater rating than one who could not perform with such facility but within his limited facility developed a greater expressiveness?

Predictability

There is a difficulty in accepting the idea of prediction when the study is based on a small population.

The investigator raises an old issue: can one predict the success of an individual in a music program. It is realized that line 15 on page 8 is phrased very carefully but its context gives this reader a very clear impression that the writer hints at the prediction of success. This should be clearly stated instead of backed into later on in the study. This claim and attempt of proof, of course, is regrettable. All the investigator can claim, assuming the test is valid, is that the test identifies musical ability. To claim any kind of predictive ability is to claim that identified ability properly nurtured, will mature. To predict success is to include in the prediction all the variables in native ability as well as environmental influences which will impinge on the organism.

Another major difficulty in the predictive aspects of the conclusions is the many revisions of test materials as the project proceeded. This created a problem of validity in the comparisons made between the years (see p. 55). Before comparisons between two versions of a test can be made, specific validating procedures must be used as in the development of the short form of a test. A prediction of success, it would seem to me, must be based on a single test administered at the beginning and at the end, holding aptitude constant.

Analysis of Data

The results of data analyses were not examined closely by this reader. This was considered a second level operation after questions concerning the procedures were resolved.

Conclusions of the Study

The conclusion of the study (see p. 58) may be warranted in

terms of the data reported but it is not warranted in terms of the procedures which yielded that data. It is possible that the relationship between achievement and this aptitude test is greater than found here, the problems cited on p. 58 not withstanding. The comparison of scores derived from different tests not validated as a continuum of difficulty may be the source of difficulty. The correlations reported on page 59 indicate some kind of common learnings but is this a validity or a reliability measure?

In addition, there are some interpretations of data which can be challenged. For example, to statistically account for a low aptitude in 23 dropouts and not account for the other variables influencing these dropouts leaves many questions unanswered.

Commentary

This reader has had considerable difficulty in expressing the preceeding reservations about the study. The candidness of the investigator's report has made possible these rather verbose reactions. Ironically, the aptitude test seems to be reliable and valid as it stands. This reader's difficulty has been in finding proof that it is in terms of the procedures used.

The major difficulty in developing a test of this kind is that not only must the test evolve but so must its validating procedures -- thesis and antithesis. The one is dependent on the other. It seems that the validating procedures (the achievement test, etc.) are developed for future use in this project. This alone is no small endeavor.

There are times when the behavioral sciences are more opinion than science. The investigator now has the further trouble of investigating the validity of this reader's opinions. They may not be so.

I apologize for having to do this under pressure and probably emphasizing the negative too much. There is much of value to Music Education endeavors in this study.

THE UNIVERSITY OF IOWA
IOWA CITY, IOWA 52240

University High School
Telephone: Area Code 319
Principal's Office: 353-4794
General Office: 353-5601

January 25, 1967

Mr. John E. Simmons
Director of Publications
University of Iowa

Dear Mr. Simmons:

I appreciate having had the opportunity of reading the critique of my report titled A Three-Year Longitudinal Predictive Validity Study of The Musical Aptitude Profile. I found the review to be both informative and confusing. Nevertheless, as you requested, I will attempt to specifically present my reactions to the critique. I shall accomplish this by referring to each statement which I have numbered on the report (beginning on page one) to correspond with the following numbers of this reply.

1. The Musical Aptitude Profile test manual, first referred to on page one of the study, is published by the Houghton Mifflin Company. It is 113 pages in length and the Technical Considerations begin on page 44. Therefore, it seems most impractical to me to append the manual to the study. Regarding incorporation, the footnotes on pages 3, 6, and 9, for example, are in keeping with the critic's suggestion.

2. I do not understand what is meant by "parallel rather than in packaged sequences".

3. Subheadings were kept to a minimum so that results could be reported in a less dry and reportorial style. Because the nature of the results to be reported were similar for each of the three years, strict organization could make the style quite rigid.

4. In the initial draft of the study, the tables were not divided between the appendices and the text. However, it was decided that the tables be divided because they could not really be absorbed when presented in such rapid order. For this reason, only tables for all schools combined appear in the text and those for individual schools appear in the appendices. I am not opposed to presenting all tables in the text if it is considered desirable.

5. In a predictive validity investigation of any test, it is obvious that the hypotheses being tested is that the predictive power of the test will be demonstrated. Only for more theoretical or abstract studies, in which there may be many "unknowns", could the statement of a hypothesis be reasonable. I suppose I could hypothesize the degree of predictive validity I expected to find but I am not certain what purpose this would serve.

Regarding the lack of an "overall statement of procedures" the purpose of the study can be found beginning on page 6, second paragraph and [redacted] at the bottom of page 8; the "procedures" are presented [redacted] on page 9 and end on the middle of page 11.

6. I would be more than happy to include (1) a discussion of the development of the achievement test and/or (2) three twenty-two page achievement test booklets in the study. However, the test is adequately described beginning on page 11, second paragraph, through page 12; and means, standard deviations, and reliability coefficients for each version of the test are included in the results for each corresponding year of the study (See Tables 3, 4, 11, 12, 19, and 20). I think the description of overall test content, and not a report on individual item development techniques, is most important for evaluating the adequacy of test to serve as a validity criteria.

7. The statement that the three versions of the achievement test should be "a continuum comprising one test instrument" puzzles me. Surely, if I wanted to compare a student's academic achievement in fifth grade with his academic achievement in seventh grade, it would be most inefficient to administer the same achievement test to him in both grades. More practically, the student should be administered different and more appropriate tests of academic achievement in each grade. Further, in the study under review, achievement test means for different versions administered in different grades were not compared to estimate growth. Only the correlations between aptitude scores and achievement test scores were investigated.

8. These analyses are made. See page 41, last paragraph for second-year results and page 57 for third-year results. Further, more extensive analyses can also be found in the section titled The Effects of Practice and Training on Aptitude Test Scores which begins on page 66 and ends on page 71.

9. If a test, such as the Musical Aptitude Profile is to actually be used for predicting a student's success in instrumental music in a typical school under typical conditions, it must be validated on students who are studying instrumental music in a typical school under typical conditions. If all variables were controlled (even though they cannot be) in a predictive validity study, the results could only be generalized to students who use --- music, who have lessons at ----- A.M. or P.M., who are taught by a---------- teacher, who are enrolled in a school that has --------- goals, etc. The only important and realistic "control" in the study under review was that all students were given ample opportunity to learn to play an instrument.

10. As implied in the study, each of the two judges evaluated 1500 tape-recorded performances each year of the study. Undoubtedly, if time and money permitted, additional evaluations of students' tape-recorded performances would have been helpful. To the extent that a critic feels that under the circumstances two judges are not a sufficient number for the adjudication endeavor, he must necessarily consider this a limitation of the study and interpret the results accordingly. It is interesting to speculate (like how many hairs make a beard) on how many judges would be needed to satisfy every critic.

11. I am not sure whether this is a negative criticism or not. If so, the critic has answered his own question.

12. In some schools, more than one teacher was responsible for the students. Therefore, in those schools, more than one teacher evaluated the achievement of the students according to the directions on the rating form presented in Appendix A. Is there any other method that should have been employed? I guess I do not understand what is meant by "several degrees of validity".

13. The etudes are presented in Appendix C and therefore, the content validity of these etudes can be assessed by the reader. As should be obvious to any experienced music teacher, each etude progresses from easy to difficult and the etudes become more complex each year of the study. Content for the etudes is based on material found in appropriate graded methods books. I feel that a discussion would seem apologetic and would be perfunctory.

14. The criticism is helpful. I have added the necessary directionson on the judges' form presented in Appendix B.

15. Are personal relations ever "eliminated" between teacher and student? If in fact they could have been eliminated in this study, my comments in #9 explain why I would not have tried to "control" this factor.

16. What can I say? Is the critic saying that the Graduate Record Examination or the ACT tests do not and cannot predict success in college? Is he saying that the Musical Aptitude Profile did not predict success in instrumental music achievement? This is not a matter of opinion, it is a matter of fact. That is what the study is all about and it is explained ("and not backed into later on") on page 8, line 15.

17. Please see #7 above. Regarding the sentence circled, if someone can interpret it for me, I would be much appreciative. Frankly the sentence embarrasses me.

18. I do not understand the criticism. It is not clear but it seems contradictory.

19. The critic missed the whole point of the analysis and the interpretation. I can only request that you read the following pages: 59, last paragraph, through 61, first paragraph.

20. I guess I miss the point now.

Sincerely,

Edwin Gordon
Associate Professor
Music Education

EG:vc