DOCUMENT RESUME

ED 129 839 TM 005 455

AUTHOR Oldefendt, Susan J.
TITLE Scoring Instrumental and Vocal Musical Performances.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
PUB DATE [Apr 76]
NOTE 12p.; Paper presented at the Annual Convention of the National Council on Measurement in Education (San Francisco, California, April 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Educational Assessment; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; Examiners; *Music; *National Surveys; *Performance; *Scoring; Skill Analysis; *Test Construction; Testing; Training Techniques; Vocal Music; Young Adults
IDENTIFIERS National Assessment of Educational Progress; *National Assessment of Music

ABSTRACT

The first National Assessment of Music, conducted in 1971-72, measured the knowledge, skills, and attitudes of 9 year olds, 13 year olds, 17 year olds, and young adults, resulting in estimates of proportions of people in the population who have certain attitudes toward music, knowledge about music terminology, notaion and history, and musical performance skills. For the assessment of performance skills, new types of exercises and administration procedures were designed, and scoring criteria for the variety of performance tasks were developed. Standard instructions were given at all levels, and responses were recorded so they could be evaluated and scored later by trained music educators. The scorers counted errors in completeness, pitch, and rhythm, and the summary of these was the score for overall quality. In each category, the error rate determined whether a performance was "markedly deficient" or not. This development of methodologies for constructing items and scoring criteria for measuring musical performance skills across a wide range of abilities in the population was a pioneer effect. (BW)

SCORING INSTRUMENTAL AND VOCAL MUSICAL PERFORMANCES

by

Susan J. Oldefendt

National Assessment of Educational Progress

Paper presented at annual convention of
the National Council on Measurement in Education,
San Francisco, April, 1976.

Scoring Instrumental and Vocal Musical Performances

Susan J. Oldefendt

National Assessment of Educational Progress

The first National Assessment of Music, conducted in 1971-1972, measured the knowledges, skills and attitudes of 9-year-olds, 13-year-olds, 17-year-olds and young adults (ages 26-35). Data were gathered from individuals in national probability samples, drawn without regard for whether the persons had received formal music instruction. The results of the assessment are estimates of proportions of people in the population who have certain attitudes toward music, knowledge about music terminology, notation and history, and musical performance skills.

The original panel of professionals, convened in 1965 to identify the objectives of music education, realized early in their deliberations that music is "... first of all a personal, aesthetic experience--in terms of composition, production or response. It is not easy to assess such an experience, and certainly not easy to set standards for it."[1] Nevertheless, a comprehensive music assessment was designed which included the measurement of performance skills since musical performance has traditionally occupied an important position in school music programs. This aspect of the assessment proved to be particularly challenging because new methods for measuring instrumental and vocal performance skills had to be developed. New types of exercises (test items) and

---

[1] Music Objectives, 1971-72 National Assessment of Music (Denver, CO: National Assessment of Educational Progress, 1970).

administration procedures were designed, and scoring criteria for the variety of performance tasks were developed.

Exercise Development and Administration.

Objectives for the assessment were defined by music educators, musicians, and lay people. The objectives for the performance tasks included: singing familiar songs, maintaining a harmonic line, inventing and improvising, sight-singing and repeating rhythmic and harmonic lines. Many exercises were developed to measure each of these areas. All of the exercises were designed to be administered to one respondent at a time in an interview setting. There were no attempts made to develop exercises to measure performance skills of groups, such as bands and choruses. Each item either included tape recorded voices or instruments as stimuli to sing along with or required a free response by the individual who could sing or play any song he chose.

After the exercises were field tested, the results were reviewed by a group of music educators who selected the exercises which provided the best measures of the objectives of the survey. The selected exercises were included in the regular assessment and were administered by a specially trained field staff. However, since the field staff were not musical experts and since standardization of field procedures and standardization and reliability of scoring procedures were essential to the integrity of the study, musical responses to the items were tape recorded in the field and the recordings sent to a highly trained central staff for scoring.

Development of Scoring Criteria.

Taped response data from the field trials were transcribed into musical notation in an attempt to quantify and classify the types of responses. Since the sample of responses included all levels of musical ability, the range of quality was extremely broad. It was decided early that the quality of voice per se and the quality of instrumental tones could not readily be quantified and therefore would not be included in the criteria. Instead, the accuracy of reproduction of rhythms and of pitch interval relationships would be the basis for quantifying responses.

After data had been collected from all ages during the regular assessment, two experts in music listened to hundreds of recordings and selected 30 sample responses for each item which typified the full range of responses. A conference was held at which eight music educators who had participated in the development of the objectives listened to the samples and developed the exact criteria for scoring. The approach used in developing the criteria for each exercise emphasized that the set of criteria must always relate to the purpose of that particular exercise. In order to focus the thoughts of the consultants on this purpose very specific questions were asked about which elements of the performance responses should and should not be considered in the efforts to quantify the response. The list of questions is presented below.

1. How much may be omitted before a response is scored as incomplete?
2. How much should a score of incomplete affect the overall score?

3. How much weight is to be given to (1) occasional out-of-tune notes, (2) generally poor intonation, and (3) intonation problems caused by the respondent's range limitations?
4. What percentage of notes may be missed before a lowering of the score?
5. How much weight is to be given to the location of errors; that is, will a good ending after a poor beginning get a better score than a good beginning with subsequent faltering?
6. Should minor errors in rhythm count as much as pitch errors?
7. How much should errors in words be considered?
8. Should a change of octave be considered in scoring?
9. If the respondent is interrupted in any way, how should this be considered in scoring?
10. Can there be a pitch or rhythm score as "marked deficiency" along with an overall score of "adequate"?
11. Should respondent's shortening of last note in vocal items be considered?
12. Can a standard percentage of correct pitches or durations be set to separate Very Good from Adequate, Adequate from Poor, Markedly Deficient from Not Markedly Deficient, etc.?

For each exercise the basic task was to determine the number of pitch and rhythm errors in a performance that would be permitted before that aspect of the performance would be classified as deficient.

Since each item varied in complexity as well as in length of response these determinations were made on an item-by-item basis, by a consensus of the group of music educators. Ultimately, judgments had to be made about the overall quality of each performance (i.e., whether it was, in general, poor, adequate or good). In order to use more information than simply the number of pitch and rhythm errors, the additional criteria regarding complete versus incomplete performances and regarding performances which were exceptionally "good" were applied to the responses. The above average performances were generally of higher quality and had fewer pitch and rhythm errors than average/adequate responses.

Because the objectives delineated performance skills that were to be measured at all ages in the assessment, most of the exercises were administered with identical instructions and procedures to individuals at all four assessment age levels. The scoring criteria were not designed to be more lenient at one age than another. Instead, the same criteria were applied to all responses to an exercise without regard for the age of the respondent. Because the criteria were constant across ages, the results for different ages for the same exercise can be directly compared to determine if, for example, more 17-year-olds than 13-year-olds can sight read a simple, short line of music notation. An example of an item and its scoring criteria are presented below.

(Turn on the RESPONSE tape recorder.)

(Before reading this exercise, give respondent the supplementary package opened to page 3.)

Here are the words to the song "America." You will hear it sung two times. You may join in singing at the beginning or when the announcer on the tape tells you to.

(Turn on the STIMULUS tape recorder. Do NOT comment on the quality of respondent's singing. If respondent stops singing, encourage him to continue. When the music stops, turn off the STIMULUS tape recorder and fill in the appropriate words below.)

	FIRST TIME	SECOND TIME
No response	○	○
Incomplete response	○	○
Complete response	○	○

My country, 'tis of thee.
Sweet land of liberty.
Of thee I sing.
Land where my fathers died.
Land of the pilgrims' pride.
From every mountainside
Let freedom ring.

EXERCISE R1A—SCORING GUIDELINES

Generally only the second performance of "America" was scored in this exercise. If, however, there was no response or an incomplete response the second time, the better of the two times was scored.

Completeness: *No Response, Complete, Incomplete*
Respondents who made no attempt to sing "America" received scores of *no response* for completeness and for pitch, rhythm and overall quality.

For a response to have been considered *complete*, it must have filled approximately 80% of the time span. For example, respondents could have delayed for two measures at the beginning of "America" and left off the last two notes of the song.

Pitch: *Not Markedly Deficient, Markedly Deficient*
To have been considered correct, a pitch must have been closer to the right pitch than to the next half-step. Changes in register were not considered errors. A response which maintained the correct pitch in all but three notes was considered to be *not markedly deficient* in pitch.

Rhythm: *Not Markedly Deficient, Markedly Deficient*
The rhythm was considered incorrect if the singing deviated from the stimulus rhythm to the extent that it could have been notated more accurately another way; slight tempo changes were acceptable. A response was considered *not markedly deficient* in rhythm if it contained less than four rhythmic errors.

Overall Quality: *Acceptable (Adequate or Good), Poor*
Completeness, pitch and rhythm all contributed to the overall quality score; other factors, such as the correctness of the words, were not included. An *adequate* response contained not more than three pitch and three rhythmic errors; that is, it was not markedly deficient in either pitch or rhythm. A *good* response was complete, maintained correct pitch in all but the first two notes and contained no more than one rhythmic error. Both good and adequate responses were considered *acceptable*.

Poor responses fell into three categories. Poor pitch was a response markedly deficient in pitch but not markedly deficient in rhythm. Poor rhythm was a response not markedly deficient in pitch but markedly deficient in rhythm. Poor pitch and rhythm was a response markedly deficient in both pitch and rhythm.

After criteria were drafted the consultants listened to additional samples and scored each of these independently. The group discussed the scores and resolved differences of opinion by adding more specific criteria to written guidelines. Since the scoring of the actual assessment responses was to be carried out by a different set of individuals, it was imperative that the guidelines be specific, complete, unambiguous, and clearly communicated.

Development of Scoring Procedures.

The number of taped responses to be scored (approximately 112,000 total responses, 2,500 responses per item per age) necessitated that a staff of scorers be trained to complete the

adults; (3) performing ability in one or more areas of vocal or instrumental music. During a one week training period scorers studied the objectives and written scoring criteria. Several recorded sample responses from the criteria conference were heard and discussed. Then each scorer independently scored more samples, with reference to the scoring criteria, but without consultation with others or reference to the standard scores from the criteria conference. Further discussion on the guidelines took place whenever a scorer's decision disagreed with the standard score. This process continued until a high degree of understanding and consistency was achieved.

After completion of the training phase each scorer listened to responses through headphones for about six hours a day for five months. The chief scorer conducted a quality control in order to assure consistent adherence to the scoring guidelines by the scorers. He independently rescored the first twenty responses for each exercise for each scorer and discussed reasons for disagreements in scores with the individual scorer. Throughout the five months the chief scorer conducted an additional independent rescoring of five percent of the total number of responses. Periodically, all of the scorers relistened to the scoring conference calibration sample tapes to reinforce their familiarity with the scoring criteria.

Summary and Conclusions.

It must be remembered that each performance exercise was originally developed for the purpose of measuring the achievement

scoring task accurately and efficiently. Since each individual had responded to three or four (but not all) of the performance items there was a possibility that halo effects (either positive or negative) might bias the results if one scorer scored all of the responses for one individual.

Tryout data and draft scoring criteria were used in a small study in which two scorers evaluated taped responses under each of two scoring conditions. In one condition each scorer scored all of the performances for one respondent before starting to score performances of another respondent. The same sample of taped responses was used in the second condition in which each scorer listened to and scored four responses to one exercise before going on to score responses to a different exercise.

There were 224 pairs of score points for each scorer. Scorer A disagreed with himself on three score points; Scorer B disagreed with herself on eleven score points. The two scorers agreed with each other 95% of the time under Condition 1 and 94% of the time under Condition 2. These results lead to the decision that scoring all of the performances for a single respondent as a unit was an acceptable procedure. (One exercise had distinctly lower agreement between scorers than the other exercises. The problem of too much ambiguity in the scoring criteria was corrected when the criteria conference added specificity to the guidelines.)

Scorer Training and Scoring Quality Control.

The chief scorer, who had participated in the criteria conference, trained 10 scorers, whose qualifications were as follows: (1) bachelor's degree in music with evidence of successful graduate work in music; (2) experience in teaching music to children and

of a particular objective. Success in developing relevant scoring criteria was dependent on keeping the purpose of each exercise clearly in mind.

The scorers had to be carefully trained so that they understood the purpose of the exercise and the reasons why it was being scored for particular features in the responses. Obviously it would have been possible to score for features other than the ones selected by the criteria conference. Additional descriptive information about responses could have been gathered, but since the main purpose of the assessment was to gather data relevant to measuring achievement of the objectives, the criteria were limited to those most relevant to the overall purpose.

Scorer agreement during training was most easily obtained when an exercise was designed to elicit a definable range of response variations and when there was a relatively small number of scoring categories. There were two major problems encountered during the training and the regular scoring. First, the scorers sometimes had difficulties in counting the number of pitch and rhythm errors when they occurred simultaneously. Second, the criteria for the two exercises which allowed respondents to sing or play songs of their own choice were too broadly defined. The consultants were unable to develop specific criteria which could accommodate the wide range of performance skills evident in these free response exercises. Attempts were made to quantify the difficulty level of the passage being performed according to the Selective Music Lists prepared by the National Interscholastic Music Activities Commission of the Music Educators National Conference. However, since there

are both simple and complex arrangements of most musical pieces it was sometimes difficult for the scorers to determine the appropriate difficulty level classification. In addition, judgments about the quality of the free response performances were by far the most subjective decisions the scorers had to make. Consequently, the amount of descriptive information gained from scoring responses to these very unstructured items is limited.

For all of the performance exercises the use of sample recorded responses was essential for the purpose of clarifying the meaning of the written guidelines. The reporting of the results of the assessment has included the preparation of a cassette recording with example responses and their scores so that interested persons can better understand the assessment results in terms of the relationship between the written criteria and the actual responses.

The development of methodologies for constructing items and scoring criteria for measuring musical performance skills across a wide range of abilities in the population was a pioneering effort resulting in reliable and reportable data. Although the scoring system could certainly be improved and refined if a second assessment were conducted, the first scoring project was a success in terms of proving the feasibility of such an undertaking.