Number 15 Special Research Interest Group in Spring 1993

Measurement and Evaluation

In This Issue...

In New York State: Assessment and "A New Compact for Learning" page 1

The Assessment of Music Performance: Techniques for Classroom and Rehearsal page7

Content Area Performance Assessment in Music: California's New Test for Prospective Music Teachers page 11

Editor's Desk page 15

Regional Evaluation Representatives

East
Richard Sang
Queens College/CUNY
Aaron Copland School of Music
Flushing, NY 11367
(718) 997-3850

West
Carol Harrison
California State University
Department of Music
Fullerton, CA 92634
(714) 449-5342

South
Patricia Sink
University of North Carolina
School of Music
Greensboro, NC 27412
(919) 384-5969

Southwest
Charles Chapman
Southwest Oklahoma State University
Department of Music
Weatherford, OK 73096
(405) 772-6611

Northwest
Ed Asmus
University of Utah
Department of Music
Salt Lake City, UT 84112
(801) 581-7369

Northcentral
Susan Vaughan
State Supervisor of Music
550 Cedar Street
St. Paul, MN 55101
(612) 296-4075

In New York State: Assessment and "A New Compact for Learning"

Ross Miller, Assistant Professor of Music
Nazareth College of Rochester

Introduction

In 1991, "A New Compact for Learning" (NCFL) was adopted by the New York State Board of Regents (the state's education governing board) to revitalize public education in New York state. The NCFL calls for three kinds of assessment of student achievement in each subject area: portfolios, professional observations, and examinations. These three assessments are currently being investigated by the New York State School Music Association (NYSSMA) Evaluation Committee (of which the author is a member) and other NYSSMA members to determine if they can be used by the membership to improve music learning and teaching in the state.

The NCFL is a document of empirical educational ideas that establishes an outcomes model for education in New York. All children (within reasonable limits) are to meet the same high levels of achievement at the end of their schooling. It is the responsibility of the schools to provide the resources necessary to help all children achieve the outcomes that will be specified for each subject area. The NCFL is a planning document "which focuses on results, which promotes local initiative, and which empowers people at all levels of the system" (University of the State of New York, p. 2). Organizing principles of the NCFL assert that 1) all children can learn, 2) the aim should be for mastery and the equity of *outcome*, not of *input*, and 3) all participants in the educational system will be given the authority to discharge their responsibilities and be held accountable for the results. (University of the State of New York, p. 3)

Very general musical outcomes may be inferred from the *Regents Goals for Elementary, Middle, and Secondary School Students*. Goal 3 is most pertinent to the arts and particularly music. It states, "Each student will acquire knowledge, understanding, and appreciation of the artistic, cultural, and intellectual accomplishments of civilization, and develop the skills to express personal artistic talents. The areas of learning include:

3.1 Ways to develop knowledge and appreciation of the arts.
3.2 Aesthetic judgments and the ability to apply them to works of art.
3.3 Ability to use cultural resources of museums, libraries, theaters, historic sites, and performing arts groups.
3.4 Ability to produce or perform works in at least one major art form.
3.5 Materials, media, and history of major art forms.
3.6 Understanding of the diversity of cultural heritages. (University of the State of New York, p. 20)

The State Education Department presently publishes syllabi/handbooks containing suggested objectives and activities for state-required music courses, but music courses vary markedly from district to district. More uniform outcomes for music programs in New York may be proposed as various state and national groups work to develop standards and outcomes.

The NYSSMA evaluation committee is working to find tools for assessment in music that match the new compact's assessment plan. The assessment section of the NCFL is presently very general. No specific assessment tools are listed for any subject area. These guidelines are given:

The existing State testing program, except for the Regents examinations, will be revised. Assessments will not be limited to multiple-choice, standardized tests, but will include components such as the following:

(1) Examinations which measure problem-solving skills and the ability to analyze and synthesize, as well as recall facts;

(2) A portfolio of the pupil's best work, certified by his/her teachers and evaluated by qualified raters;

(3) A professional evaluation of the pupil's accomplishments, made by his/her teachers. This evaluation should extend not only to basic skills and knowledge, but also to desirable qualities (such as persistence, creativity, and sensitivity to others) not easily measured by conventional means.

"The revised state testing program will intrude as little as possible upon the time available for instruction. Where feasible, sampling techniques will be used. State assessments will be consolidated at grades 4, 8, and 12. Local assessments of programs will be made as appropriate throughout the grades." (University of the State of New York, p. 6)

Music teachers, like their colleagues in other subjects, may have to learn to use assessment measures with which they are unfamiliar but which have potential to result in important musical outcomes to meet the mandate of the new compact.

Portfolio

Three members of the NYSSMA Evaluation Committee participated in the New York State Education Department's *Institute on Assessment in the Arts: Portfolio*, during the summer of 1992. Later in that summer, training in the use of portfolio for the general NYSSMA membership began when one committee member developed an "Introduction to Portfolio" mini-workshop for music teachers attending NYSSMA state conventions. These sessions showcased several teachers' ways of working with portfolios in music classes.

One important portfolio concept is the classification of portfolios according to function. These classifications include:

(1) the *presentation/product* portfolio includes finished student work of consistent high quality. This kind of portfolio can serve a "gatekeeper" function where the evaluator assesses the work and decides whether the product is good enough to permit the portfolio owner to qualify for something, perhaps a job, admission to a school, advancement, and the like.

(2) the *product/performance* portfolio takes a "snapshot" of the work of a group of people by collecting the same product from each participant at the same point in time. This kind of portfolio can be useful for ranking and sorting students, teachers, school districts, and so forth.

(3) the *program portfolio* is a collection of the best work of a group of students from a particular program. This kind of portfolio might be assembled by a music teacher to help in the justification of funding for a music program or presenting a request for additional resources.

(4) the *process portfolio*, which is a collection of an individual student's works *including* his or her early and (perhaps) less successful attempts at production. Documenting the processes of revision and self-evaluation are pivotal in the *process portfolio*: students are taught to use the same criteria and processes that a "master" uses when evaluating and revising his or her own productions. The *process portfolio* figures prominently in the *Arts Propel* program in Pittsburgh where music students learn by completing a series of projects and have experiences in production, perception, and reflection.

The NCFL document calls for portfolios most closely resembling the *presentation/product* portfolio model. Collecting materials for this portfolio would be facilitated by the use of the *process portfolio* in music classrooms. The *process portfolio* can be the progenitor of any type of portfolio. How to implement the process *portfolio* will best be left to the individual school district or to the individual teacher until music curricula and school resources are more uniform throughout the state. The procedure for selecting works from the *process portfolio* for the *presentation/product* portfolio to be consolidated at grades 4, 8, and 12 and used for the state assessment program, remains unspecified. Music teachers will not be alone in learning about assessment, however, since the NCFL provides no specific details for any subject area on how presentation/product portfolios should be assembled.

The benefits of the portfolio have been previously discussed in this newsletter. The development of systems of scoring, and the sharing of the scoring criteria with students, is a particular strength of the process of portfolio development, enabling students to strive for skills of criticism using the same criteria for evaluation that an expert in the subject area uses. Musicians both young and old, amateur and professional, are well served by well developed critical evaluation skills in their daily practice, leisure and analytical listening, and creative musical work. Inclusion of the *process portfolio* model in music classrooms could be an exciting development.

Examinations

There is an understandable excitement among teachers about the *process portfolio* and the way it may invigorate music learning. But assessment that does not require substantial amounts of record keeping and scoring time, and which can provide information about large numbers of students, is also desirable. A short review of published music tests reveals several "examinations which measure problem-solving skills and the ability to analyze and synthesize, as well as factual recall" (University of the State of New York, p. 6). Selected questions from these music examinations and similar test items to be developed in the future could satisfy NCFL examination requirements. Suggested ways of asking musical questions found in presently published tests are presented in Figure 1.

(S = the student, see References for complete test titles)

CONCEPT EXAMINED	MODEL FOR TESTING	REPRESENTATIVE TEST
Mode Identification	Chords, or phrases, or melodic patterns are played and S identifies whether major, minor, or neither is heard	*MAT 2, ITML 1*
Meter Discrimination	S listens to music and identifies whether music is in duple or triple meter	*MAT 1, ITML 1*
Skips and Steps	S listens to music and identifies whether music moves mostly by skip or step	*MAT 1*
Tonal Center Identification	A series of chords or a phrase is played and S identifies the key tone from among several tones played.	*MAT 2*
Error Identification in Notation	S listens to music while watching notation and indicates where the sound and notation do not match. (Separate exercises for pitch and rhythm).	*MAT 2 , ITML 1, Aliferis*
Interval Identification	S looks at two written pitches, the first pitch is played. S then imagines the second pitch, identifies from among three played examples which of the three is the pitch of the second written note OR an interval is played and the name of the interval is written.	*MAT 3, Traditional music theory test*
Musical Memory	A music fragment is played and then played again with or without changes. S must identify in some way *whether* a change has occurred or exactly *what* the change was (pitch, rhythm, or key) or *on what beat* the change occurred.	*Bentley, Drake, PMMA, IMMA*
Chord Analysis	A chord of two or more notes is played, S must tell how many notes are played OR a chord is played and S must tell which of three chords played afterwards is alike.	*Bentley, MAT 4*
Steady Beat	S listens to voice count 1, 2, 3, 4 in a certain tempo, S continues to count in tempo during silence which follows voice. Voice says "stop" and S writes down the total number of imagined steady beats.	*Drake*
Sight Reading	S plays music at first sight, playing progressively more difficult music until below a certain level of accuracy. Scored for pitch, rhythm, and expression separately or single combined score.	*Watkins-Farnum*

FIGURE 1

Many additional models have been used in published tests. The concepts tested and ways of answering are varied, but most use a "musical" stimulus to present the problem or a musical production is required in response. The questions represent "valid" musical problems to solve, similar to real musical problems encountered by musicians. These "musical" questions require an interaction with sound, present or imagined. In contrast, "When was Beethoven born?" could be called a "music" question, as it tests knowledge and requires no musical skill or interaction with sound to answer. Examinations comprised of "musical" and "music" questions would test both the skills and knowledge important to a music curriculum. Too often, the "musical" questions are neglected.

The evaluation of discrete musical skills represented by these kinds of problems would be useful to teachers, students, and for NCFL requirements. Some of the problems test prerequisite skills necessary to the achievement of more complex musical understandings and skills. Concepts such as higher and lower, steady beat, and longer and shorter (among others) are important enough to young musicians that valid and reliable evaluations of these skills should be conducted to monitor and encourage students' progress until the skills are mastered.

Authentic Testing

While the problems from these music tests may require musical thinking to solve, they fall somewhat short of the complex, multiple-skill tasks that musi-

cians must complete outside the testing room. As our schools move to implement "authentic" testing (also called performance-based assessment), sequential and complex musical problems appropriate to a variety of music achievement levels need to be developed along with the means for scoring student solutions to the problems. The writing of problems and developing scoring criteria require a thorough knowledge of one's music curriculum and demand both critical and realistic thinking.

There is an almost inescapable consequence to change the learning/teaching process when the evaluation model is changed. Most proponents of authentic testing believe that changes associated with the use of authentic tests will be beneficial for students and teachers alike. Authentic tests challenge students to analyze, synthesize, and evaluate. Teachers are challenged to train students for more independent, insightful, and self-critical music participation. One serious drawback of some testing is that creativity is not demanded in solving the problems; there is one correct answer. Authentic testing addresses this concern as teachers encourage students to create their unique solutions to complex musical problems.

A look at the scoring of questions used in the music portion of the 1971/72 National Assessment of Educational Progress provides a model for scoring authentic tests. In one of the NAEP tasks, students sang "America" along with a taped accompaniment and then sang it alone. Rhythm and pitch were scored separately as either Acceptable or Unacceptable. Overall quality was scored as Good, Acceptable, or Poor.

(1) Rhythm: The rhythm was considered to be incorrect if the performance would be notated differently to reflect the singing, although slight tempo changes were acceptable. To be acceptable in rhythm, a response could not have more than three rhythmic errors.

(2) Pitch: To be considered correct, a pitch must be closer to the right pitch than to the next half step. However, changes in register were not considered an error. An acceptable response was one that included fewer than four pitch errors.

(3) Overall Quality: An acceptable response was any response acceptable in both pitch and rhythm. Good responses not only met these two criteria, but maintained the correct pitch in all but the first two notes and included no more than one rhythmic error (National Assessment of Educational Progress, 1974, p 1).

While the NAEP standards may not have demanded a high level of performance for the top rating (written as they were for a national testing program of all students in all schools), by writing scoring statements locally and including factors emphasized in the local New York curriculum, students can be expected to meet the high standards suggested in the NCFL. For specific authentic tests to be valid, the design of tests should be local and based upon knowledge of the students, the educational setting and resources, community demands, and other contextual factors. Perhaps in the future, when (and if) national standards for music instruction are developed and implemented, and sufficient resources are provided to all school districts, more uniform models for the musical evaluation of students will be practical and similar complex tasks may be suggested for assessing music students nationwide.

While the exact design of performance-based evaluation may best be left to local decision-makers, the qualities of an "authentic evaluation" have been defined: "a true test of intellectual ability requires *the performance of exemplary tasks*. Evaluation is most accurate and equitable when it entails human judgment and dialogue, so that the person tested can ask for clarification of questions and explain his or her answers. Authentic tests recur, and they are worth practicing, rehearsing, and retaking. We become better educated by taking the test over and over. Authentic tests are not needlessly intrusive, arbitrary, or contrived merely for the sake of shaking out a single score or grade. Instead, they are "enabling" — constructed to point the student toward more sophisticated and effective ways to use knowledge" (Wiggins, 1989, pp 703, 704, 711). Other features of authentic evaluation include:

(1) the tasks require students to exercise judgments and to justify decisions made;

(2) there is little secrecy because the tasks are known ahead of time as are the scoring criteria;

(3) the tasks are like those encountered by a professional in that line of work;

(4) the tasks are complex, requiring the use of several skills and having many possible solutions; and

(5) authentic tests are a logical summative experience for students working with *process portfolios* since in evaluating *process portfolio* projects, the same kinds of evaluation criteria may be used.

What are exemplary music tasks? Charles Leonhard (professor emeritus, University of Illinois), promoted the idea that music teachers should strive to give music students the same kinds of involving experiences that music teachers have when learning and teaching music. As a starting point for developing local authentic tests, the teacher should think about what it means to be musical. What do musicians do that is musical?

Musicians perform, listen, and compose. This basic trio of musical activities subsumes a more extensive list of music activities sometimes used to organize planning for vocal-general music classes: sing, play, listen, move, read and write,

create. Either list can serve as a starting point for choosing exemplary tasks. Once students have been introduced to and practiced a chosen task, a performance-based evaluation can be conducted, testing the students in the same way that training was conducted. Exemplary tasks can be adapted to many different age levels and can serve in a "spiral curriculum" in the way that Jerome Bruner proposed basic ideas should (Bruner, 1960). Speculative authentic test models of exemplary music tasks follow.

The task: Real musicians practice music and prepare themselves for performances.

A possible evaluation model: Students are given a new piece of music and, with home practice or in-school practice, prepare the music for a performance by a specific date. Recordings of *practice sessions and the performance* are evaluated by the teacher and the student. Criteria for scoring each phase of the preparation process are known ahead of time. The performance assessment informs both the student and the teacher about the student's current ability to practice and perform. Done correctly, such an evaluation provides more feedback information than the single grade given for a typical playing test and should lead to improved practice and performance. (Sight-reading training and testing is another familiar musical task with potential for performance-based assessment of exemplary musical skills).

The task: Real musicians can make an arrangement of a melody (or create a new piece of music) for their own instrument (or other instruments with which they are familiar).

A possible evaluation model: Working with materials from lesson books, concert repertoire, radio and TV, traditional and folk sources, etc., students are assigned to arrange a melody in some specific and age-appropriate way. Elementary students could rewrite a melody in a new meter, new mode, or create a simple rhythmic or melodic variation. Secondary students could arrange the music for duet, trio, or larger group working with familiar instruments at first and later learning to work with instruments from other families. Projects would culminate in recitals of student works. Once again, a scoring plan would have to be developed that evaluates the critical steps of the project. Provided to students at the inception of the project, the scoring plan serves to guide the students throughout using the same criteria for self-evaluation that the teacher will use for the final evaluation. Tasks such as these are also appropriate for the *process portfolio* and can be scaled to an appropriate length to fit the situation: shorter for an evaluation taking a period or two, and longer for a portfolio project. The consistency of the teaching and evaluation process are a strength of the approach.

The tasks: **(1)** Real musicians can listen to music, imagine how it is written, and then write it down, *or* **(2)** A real musician can look at printed music and imagine how it sounds, *or* **(3)** A real musician can listen to music and tell something about the style of the music, instrumentation, tonal or formal structures, compositional devices, and the like.

Models for evaluation: **(1)** Simple pitch and/or rhythm dictation exercises are possible and valuable even for our youngest students. Elementary age students can "write" what they hear using felt notes on felt staves or construct simple rhythm icons out of popcicle sticks after listening to a musical fragment. Progressively learning more difficult exercises and instruction in traditional notation skills can lead students to master relatively sophisticated dictation as they progress through junior high and high school. Dictation exercises could be "graded" like music festival solos from Level I to Level VI to provide festival-like incentives to progress and lead to recognition of music programs in which a certain number of students can take dictation at a certain level. Individual records of student progress can be passed from one music teacher to another to introduce the students and their achievement to their new teachers at each grade level or building. *or* **(2)** Students receive systematic training and an evaluation of their ability to identify errors in performance when comparing sound to notation. With live performances by the music teacher, a single piece of literature could be used several times with increasing subtlety in errors in secondary performance leading to skillful aural-visual discrimination. Students could take over some or most conducting responsibilities for selected literature in certain ensembles after appropriate training and evaluation in error identification and conducting and rehearsing techniques. *or* **(3)** Analytical listening tasks could also be started early in the music curriculum and lead to aural recognition, identification, and analysis of many forms and styles of music in middle and high school general (or other) music classes. If one believes in the ear as the primary avenue of musical experience, training and assessment of listening skills takes on central importance in music classes.

When students solve complex musical problems, a deeper understanding of their many musical abilities may emerge

Be sure to attend the Measurement and Evaluation Session devoted to ***Authentic Assessment*** at the Cincinnati MENC meeting

as more sophisticated musical behaviors are required of them than is typical in many school music programs. Contrast imitative singing in chorus with sight-reading music or independent preparation of a work. Contrast "drop the needle" identification of musical works with analytical listening to identify the form, instrumentation, and expressive devices used in a composition. Contrast never writing a piece of music (as in programs where singing and playing are the only activities) with writing, revising, and performing one's own music. Evaluation models can lead the way to improved musical outcomes simply because evaluation planning begins by looking at the desired outcomes of instruction. Authentic tests demand complex musical responses. Teachers will need to lead students to practice more complex musical behaviors to prepare for the tests. "Teaching to the test" need not be avoided when the test is a good one.

Professional Observations

Record-keeping models have been the primary focus of the NYSSMA evaluation committee's work in the area of professional observations, the third part of the NCFL's assessment model. At a recent NYSSMA state music convention, several workshop speakers presented their personal tools for recording their observations of students' music and social behaviors during large music classes. NYSSMA has not yet addressed the assessment of "desirable qualities (such as persistence, creativity, and sensitivity to others) not easily measured by conventional means" (University of the State of New York, p. 6) nor located a convenient assessment tool to use. To evaluate students' personal behavior, as this statement from the NCFL requires, it may be possible to use the same assessment tool in all subject areas. In music classes though, observations should also include the assessment of affective response to music. Journals written during portfolio work might contain an individual's account of his or her response to music. Teacher observations that identify correlated behavior could confirm the student's self-evaluation.

Conclusion

The NCFL includes many of the assessment reforms suggested in education circles in the last few years (and assessment circles for many years). Although some teachers from New York are skeptical that the NCFL will be fully implemented, the evaluation plan could benefit music learning and teaching if accepted and used by music teachers. By looking for models of assessment which have been previously developed and adapting them to the needs of our individual programs, we may be able to avoid wasting time "reinventing the wheel." Examinations using "musical" problems could evaluate skills of many students quickly and efficiently. Indeed, preparing all of our students to master the musical problems presented in the best of the existing objective tests presents a formidable challenge. More complex tasks and the means for scoring them can be developed to challenge students to go beyond discrete musical problems to solve problems that working musicians encounter. If the implementation of the NCFL can create a climate amenable to improved musical outcomes for our students, the efforts required of teachers to develop powerful new teaching and evaluation methods will be worthwhile.

References

National Assessment of Educational Progress. (1974). *The first national assessment of musical performance.* (Report 03-MU-01). Washington, D. C.: U.S. Government Printing Office.

University of the State of New York, The. (1991). *A new compact for learning.* Albany: The State Education Department.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-714.

Published Tests referred to in chart

Aliferis: Aliferis, J. (1954). Aliferis Music Achievement Test. Minneapolis: University of Minnesota Press (Now available from J. E. Stecklin, 1988 Wheeler Street, Roseville, MN, 55113).

Bentley: Bentley, A. (1966). Measures of Musical Abilities. Berkshire, Great Britain: NFER- Nelson.

Drake: Drake, R. (1954). Drake Musical Aptitude Tests. Chicago: Science Research Associates. (Now available from the author, 711 Beach Rd., #106, Sarasota, FL 34242).

Colwell MAT 1, 2, 3, 4: Colwell, R. J. (1986). Music Achievement Test 1, 2, 3, and 4. Music Achievement Test. (Now available from the author, 12 Peacock Farm Road, Lexington, MA, 02173, 617/ 862-1709).

Gordon ITML: Gordon, E. (1970, 1991). Iowa Tests of Music Literacy: Levels 1-6, Chicago: G.I.A. Publications, Inc.

Gordon PMMA: Gordon, E. (1979). Primary Measures of Music Audiation. Chicago: G.I.A. Publications, Inc.

Gordon IMMA: Gordon, E. (1983). Intermediate Measures of Music Audiation. Chicago: G.I.A. Publications, Inc.

Watkins-Farnum: Watkins J. and Farnum, S. (1954). The Watkins-Farnum Performance Scale Forms A and B. Winona, MN: Hal Leonard Music.

The Assessment of Music Performance: Techniques for Classroom and Rehearsal

T. Clark Saunders, Associate Professor of Music
Hartt School of Music

Introduction

There is an episode early in my teaching career that is firmly etched into my memory. I had just completed the annual ritual of listening to each of my concert band students for the purpose of assigning seats within each section of the ensemble. Each student prepared a selected excerpt from their band music and then performed it individually as I sat in my small cluttered office and listened. As each student performed with their instrument, I made brief notes quickly indicating how much I liked each performance. After hearing all seventy-some students, the next morning I posted on the band room bulletin board a new seating chart, indicating the seating position of each student member within their instrument section.

It was later that evening at home that I received a phone call from a *concerned* parent. "Why was my son placed in the fourth-chair first-part in the trumpet section?" I began to explain in the most diplomatic manner, that her son's performance was not as proficient as the other trumpet performers that I placed higher in the section. I used delicate verbal expressions that indicated that, in my opinion, her son's trumpet performance was not as good as some other student performances.

The parent persisted in her inquiry, "In what way was my son's performance inferior to the other trumpet performances?" Having heard many different performances without any specific method of assessment, I was unable to provide specific information about the character of her son's performance. Reaching for something to say, I told her that some of the rhythm patterns within certain passages were not as accurate as the other student performances. "Could you be more specific, in what way were the rhythms inaccurate?" I was unable to be any more exact in my characterization of the performance. The parent at that point became exasperated and began to express frustration. "I am trying to help my son understand what specific things in his performance need greater preparation and you are unable to help me! It's apparent to me that you are somewhat haphazard in your method of evaluating the instrumental performances of your students." We both hung up the phone feeling less than satisfied. The parent was looking for specific information to improve her son's music performance skills and I was unable to provide specific information.

Music teachers for the most part are well trained and adequately equipped to teach students to perform music accurately and artistically. They are effective in instructing students to acquire an accumulation of: 1) observed music performance skills (performance from recall, from notation, improvisation, and the like), 2) musical understanding (phrasing, musical line) and 3) technique (fingerings, articulations, breath control, and so forth). Many music teachers are not as well prepared, however, to assess students' acquired ability levels in each of those music performance learning areas. Missing from individual music teaching methods are adequate strategies to measure the specific areas of music performance sought from the students. Music teachers are effective teachers but not necessarily effective in assessing the music performance skills, content, and techniques they teach.

Traditional Rating Scales

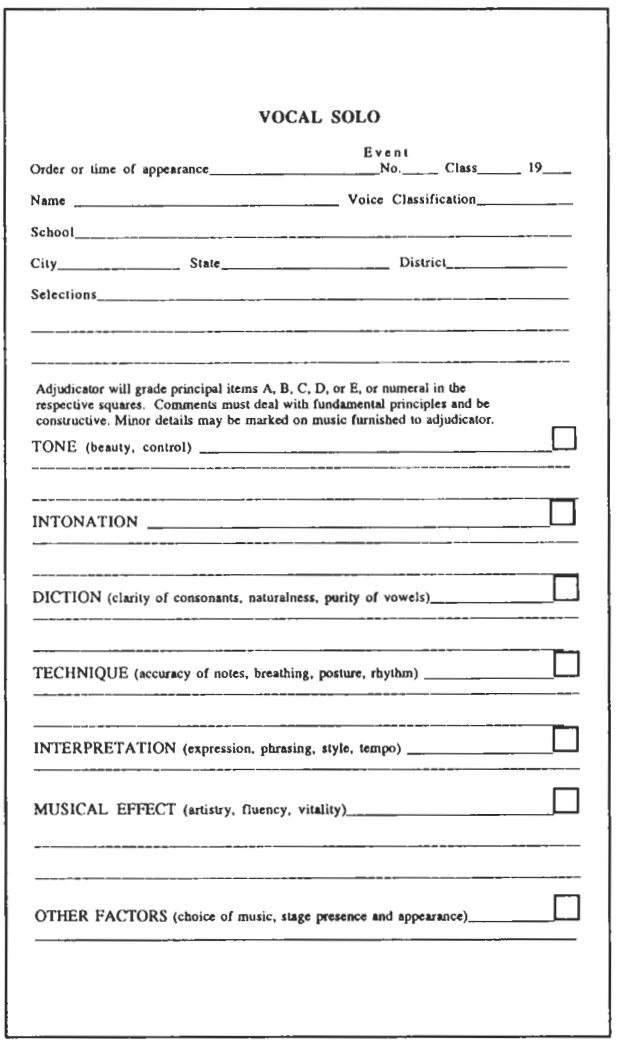
Attempts at bringing a more objective level of measurement to music performance have been made when panels of judges are assembled to evaluate bands or choruses participating in an adjudication festival, or when individual students attend performance auditions for participation in select regional or state ensemble groups. In such cases professional judges are hired and asked to listen to student performances and to use a prepared rating sheet. A typical rating sheet (see Figure 1) includes written terms that correspond to selected dimensions of music performance (rhythm, tone quality, etc). As the judges listen to a student or ensemble's performance, they are asked to assign, for each performance dimension, a number to indicate the performance level of the student or ensemble. A composite rating score is derived from the sum of the scores assigned for each of the different performance dimensions.

Typically the numbers (1-5) assigned by the judges indicate performance ability that can also be described by descriptors of sequential levels of excellence (outstanding, very good, good, poor, very poor). These are not often equal intervals, thus the judges are being asked to listen and determine, for each music performance dimension, a quality of "performance goodness." Scoring variations between two or more judges often occur, however, when using this method due to the fact that descriptors are primarily a judge's subjective opinion of goodness for a particular category. The judges bring differing performance backgrounds and expectations to the judging task. Their criteria of performance goodness is therefore quite different as there is seldom a training session for judges during which time models of various levels of performance are demonstrated and the judges consulted about severity of errors. The resulting lack of

agreement in judges' scores can cause confusion and difficulty in interpretation of the relative level of student performance, and in the determination as to which of the scores is the most accurate measure of an ensemble's or individual's level of performance ability. Fiske (1983), in an examination of traditional contest adjudication procedures, found low levels of consistency among judges' results and questioned the fairness of many traditional performance evaluations.

Criteria-Specific Rating Scales

In an attempt to minimize the inclusion of judge's subjective opinion during the act of music performance assessment, rating scales that include carefully constructed descriptions of differing levels of acquirement have been developed. With the use of these criteria-specific rating scales, the intent is to determine what specific levels of performance dimensions are a perceptible part of a whole music performance (Boyle, 1992). Judges document the level at which a musician is performing by listening and indicating particular aspects that are audible in the performance.



VOCAL SOLO

Event

Order or time of appearance_____ No.____ Class_____ 19____

Name _____ Voice Classification_____

School_____

City_____ State_____ District_____

Selections_____

_____

_____

Adjudicator will grade principal items A, B, C, D, or E, or numeral in the respective squares. Comments must deal with fundamental principles and be constructive. Minor details may be marked on music furnished to adjudicator.

TONE (beauty, control) _____ ☐

_____

INTONATION _____ ☐

_____

DICTION (clarity of consonants, naturalness, purity of vowels)_____ ☐

_____

TECHNIQUE (accuracy of notes, breathing, posture, rhythm) _____ ☐

_____

INTERPRETATION (expression, phrasing, style, tempo) _____ ☐

_____

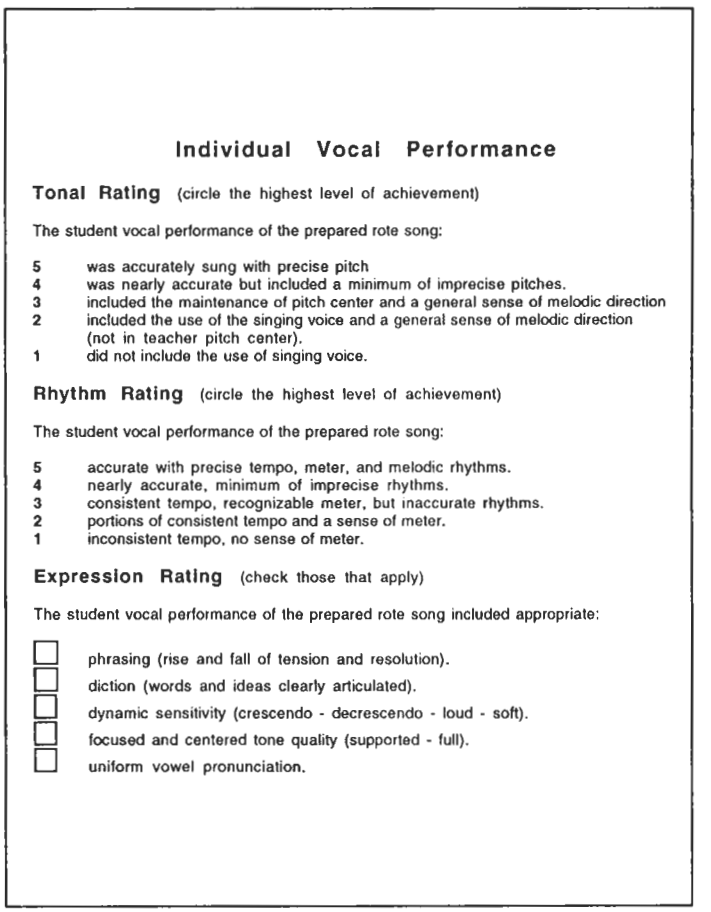
MUSICAL EFFECT (artistry, fluency, vitality)_____ ☐

_____

OTHER FACTORS (choice of music, stage presence and appearance)_____ ☐

_____

FIGURE 1



Individual Vocal Performance

Tonal Rating (circle the highest level of achievement)

The student vocal performance of the prepared rote song:

5 was accurately sung with precise pitch
4 was nearly accurate but included a minimum of imprecise pitches.
3 included the maintenance of pitch center and a general sense of melodic direction
2 included the use of the singing voice and a general sense of melodic direction (not in teacher pitch center).
1 did not include the use of singing voice.

Rhythm Rating (circle the highest level of achievement)

The student vocal performance of the prepared rote song:

5 accurate with precise tempo, meter, and melodic rhythms.
4 nearly accurate, minimum of imprecise rhythms.
3 consistent tempo, recognizable meter, but inaccurate rhythms.
2 portions of consistent tempo and a sense of meter.
1 inconsistent tempo, no sense of meter.

Expression Rating (check those that apply)

The student vocal performance of the prepared rote song included appropriate:

☐ phrasing (rise and fall of tension and resolution).
☐ diction (words and ideas clearly articulated).
☐ dynamic sensitivity (crescendo - decrescendo - loud - soft).
☐ focused and centered tone quality (supported - full).
☐ uniform vowel pronunciation.

FIGURE 2

Criteria-specific rating scales (see Figure 2) include two distinct types: a) **continuous** (sequentially more demanding performance criteria) and b) **additive** (nonsequential performance criteria) rating scales. Each of the criteria in both types of rating scale describes a specific level of perceivable music skill/content/technique achievement.

Criteria-specific rating scales have been shown to yield a high inter-rater reliability ($r > .90$) in recent music education research (Azzara, 1992; Feierabend, 1983; Levinowitz, 1989; Rutkowski, 1987 & 1990; and Saunders, 1990). Criteria-specific rating scales provide greater consistency of results of evaluation because they require specific information about a performance product. Determining through evaluating specific criteria what a student is capable of, or not capable of, in his or her music performance is a more reliable and, therefore, a more valid assessment technique than recording opinions about the level of "performance goodness."

Additional recent research (Rothlisberger, 1992) demonstrates the reliability among judges when using criteria-specific rating scales as a part of a student audition process to select honor

instrumental ensembles (see Figure 3). Judges who listened to individual student performances with criteria-specific rating scales had an inter-rater reliability > .80. Currently, Maryland and Connecticut instrumental All-State ensemble selection committees are considering discontinuing the use of long-held assessment instruments which solicit judgements of "performance goodness" and are developing rating procedures which include the use of criteria-specific rating scales.

Instructional Basis of Criteria-Specific Rating Scales

Criteria-specific rating scales provide instructional benefits not available with traditional rating scales. By examining the criteria (levels of performance) for each performance dimension marked by the judges, in addition to those criteria not marked, a student receives direct feedback in regard to areas of proficient performance capability and also performance areas that need additional attention for improved artistic and accurate performance execution. Moreover, the students receive positive/proactive information about their performance. Criteria-specific rating scales provide a clear indication of what was indeed *present* in a student's performance. In contrast, traditional rating scales contain vague information about a judge's opinion accompanied with hastily written comments, almost always noting what was *missing* or substandard about the performance (negative/reactive).

In the same way, a music teacher who creates and uses criteria-specific rating scales for the periodic assessment of student music performance skills would have specific feedback information to offer a student or to use to design long-term teaching strategies to meet a student's individual music performance learning needs. Criteria-specific rating scales would then be used to their greatest advantage: the improvement of music performance instruction.

SLOW, LYRIC PREPARED SELECTION
First Performance

TONE QUALITY OF FIRST PREPARED SELECTION (**Circle only one number**)
The student's tone quality as demonstrated in the first prepared selection:
5 was full, rich, and **characteristic of the tone quality** of the instrument in **all ranges and registers.**
4 was of a **characteristic tone quality** in **most ranges**, but distorts in occasional passages (i.e., loud or soft, high or low tessitura, etc.).
3 has **some flaws** in basic tone production (i.e., thin sound, spread tone, unfocused tone, forced tone, air in sound not contributing to the tone).
2 has **several major flaws** in basic tone production.
1 is not a tone quality characteristic of the instrument.

[ ] Number of the item circled

NOTE ACCURACY OF FIRST PREPARED SELECTION (**Circle only one number**)
The student's performance of the first prepared selection included:
5 **all** pitches/notes **performed accurately.**
4 **most** pitches performed accurately (three or four pitches/notes missed).
3 **many** pitches performed accurately but **five or more pitches/notes missed.**
2 a number of missed pitches with several starts and stops.
1 **a large number of missed** pitches/notes in the music with many starts and stops.

[ ] Number of the item circled

RHYTHM OF FIRST PREPARED SELECTION (**Circle only one number**)
The student's performance of the rhythm patterns of the first prepared selection:
5 was **accurate throughout.**
4 was **nearly accurate** but lacked precise rhythmic interpretation of some rhythm patterns.
3 **approximated the notated rhythms**, but lacked accuracy in performance of **some** rhythm patterns.
2 demonstrated an **inconsistent** performance of **most** rhythm patterns.
1 **was not accurate.**

[ ] Number of the item circled

INTONATION OF FIRST PREPARED SELECTION: (**Circle only one number**)
The intonation of the student's performance of the first prepared selection:
5 is **accurate throughout**, in all ranges and registers.
4 is accurate, but student fails to adjust on isolated pitches, or **demonstrated minimal intonation problems.**
3 is **mostly accurate**, but has notes that are played out-of-tune including some significant problems, and did not adjust on some out-of-tune pitches.
2 has a **basic sense of intonation, but did not adjust pitches** to an acceptable standard of intonation and had significant intonation problems.
1 **is not accurate.** Students performance is continuously out-of-tune.

[ ] Number of the item circled

FIGURE 3

Rating Scale Construction

A. Determine age/grade and area of instruction rating scale would be used.
- Early Childhood
- Elementary General Music
- Middle School General Music
- Middle/High School Choral
- Elementary Instrumental
- Secondary Instrumental
- Adult Music Education

B. Determine voice/instrument performance you wish to assess.
- solo voice
- vocal ensemble
- solo instrument
- instrumental ensemble

C. Determine type of performance skill to be assessed.
- performance from memory
- echo-performance with voice and/or instrument
- perform from notation (prepared/familiar music)
- perform unfamiliar music from notation (sightreading)
- create/improvise (vocally or instrumentally)
- performance from listening (movement)

D. Determine the dimension of the performance product to be assessed
- tonal/intonation
- rhythm
- expression/interpretation
- creativity
- executive techniques
  - holding/hand position
  - breath control
  - embouchure/bow
  - articulation
  - fingerings/bow technique
  - vibrato

E Write an appropriate rating scale stem (incomplete sentence format) which reflects the performance dimension (skill/content) to be assessed.

F. Determine rating scale criteria, within a selected dimension, by choosing levels of music skill, content, and technique difficulty or observed levels of performance competency. Each criteria written in a format to complete the incomplete sentence of the rating scale stem.

G Determine the type of rating scale to be constructed (continuous or additive).

H. Arrange and number **five** selected criteria as appropriate for the type of rating scale constructed (additive or continuous). For a continuous rating scale, the third criteria should be selected to correspond to the majority performance level of the assessment group. In that way, there will be room for the measurement of higher levels of performance achievement as well as lower levels of achievement from among the group.

TABLE 1

Conclusion

The successful creation of criteria-specific rating scales primarily involves a careful determination and definition of the type and the variation among levels within a dimension (see Figure 4).

Following the creation and development of effective, reliable, and valid rating scales, research should be conducted to develop a greater understanding of the

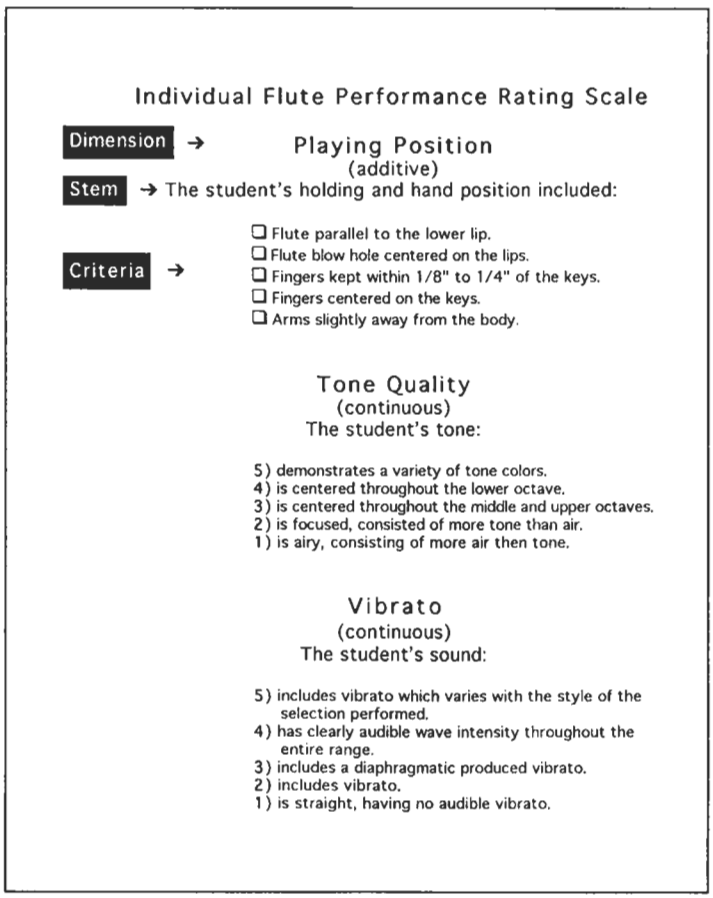


FIGURE 4

of performance to be assessed (see Table 1). The age/instructional level, type of performance skill, performance dimensions and performance criteria within each dimension must clearly be defined in order to develop appropriately stated unambiguous levels of performance competency. A performance *skill* is the exhibited student music performance behavior; performance *dimensions* are the result of the delineation of a performance skill into distinctive elements of music expression or aspects of physical execution demands; and performance *criteria* are levels of performance proficiency dimension of the human judgement of process in the evaluation of music performance behaviors.

The psychology of the assessment decision-making process, the psychometric characteristics of rating scales, and a comparison of the effectiveness of various rating scales have yet to be examined thoroughly. Possible questions to be addressed include: to what degree would there be contrasting measurement results with the use of different rating instruments of different procedural design and music dimensions and criteria? Are particular performance rating scales and corresponding music skills, contents, and techniques more susceptible to systematic error of measurement than others? Are there certain constructs of music performance measured in an optimal manner with rating scales constructed with a different number of criteria? How are criteria-specific judgements made? Does experience and expertise influence the mode of operation and consistency of performance judgements?

When I remember that awkward phone conversation with the parent of my band student, I often wish that I had been able to quickly and accurately supply specific information to characterize the quality of her son's trumpet performance. If I had been able to show that parent a performance rating procedure that systematically measured music performance skills, content, and executive technique, I would have been able to communicate directly to her what aspects of her son's trumpet performance were performed in a satisfactory manner and what aspects of his performance needed improvement. That parent would have gained confidence to know that I determined band seating assignments based on carefully collected information and I would have clear-cut knowledge to guide my instrumental music instruction.

References

Abeles, H.F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, 21, 246-255.

Azzara, C. D. (1992). *The effect of audiation-based improvisation techniques on the music achievement of elementary instrumental music students.* Unpublished doctoral dissertation, Eastman School of Music, Rochester.

Boyle, J.D. (1992). Evaluation of music ability. In R. Colwell (Ed.), *Handbook of research of music teaching and learning.* (pp. 260-262). New York: Schirmer.

(pp. 260-262). New York: Schirmer.

Cooksey, J. M. (1977). An application of the facet-factorial approach to rating scale for high school choral music performance. *Journal of Research in Music Education,* 25, 100-114.

Feierabend, J.M. (1983). *The effects of specific tonal pattern training on singing and aural discrimination abilities of first grade children.* Dissertation Abstracts International, 45, 110A. (University Microfilms No. 8410243)

Fiske, H. (1983). Judging Musical Performances: Method or Madness. *Update* 1 (3), 7-10.

Gordon, E.E. (1990). *Learning sequences in music: Skill, content, and patterns.* Chicago: GIA.

Levinowitz, L.M. (1989). An investigation of preschool children's comparative capability to sing songs with and without words. *Bulletin of the Council for Research in Music Education,* 100, 14-19.

Rothlisberger, D. (1992). *The effects of video tape preparation on performance anxiety and its relationship to instrumental music achievement among high school children.* Unpublished doctoral dissertation, The University of Maryland, College Park.

Rutkowski, J. (1987, March). *A study to investigate the relationship between kindergarten children's use of singing voice and developmental music aptitude.* Presented at the Eastern Regional Meeting of the Music Educators National Conference, Baltimore, MD.

Rutkowski, J. (1990). The measurement and evaluation of children's singing voice development. *The Quarterly,* 1, 81-95.

Saunders, T.C. (1990). A preliminary investigation of the suitability of selected rating scales used to measure student music performance skills. *Missouri Journal of Research in Music Education,* 27, 15-29.

Content Area Performance Assessment in Music: California's New Test for Prospective Music Teachers

Carolynn A. Lindeman, Professor of Music
San Francisco State University
and
Linda L. Tyler, Associate Examiner
Educational Testing Service

Introduction

The California Commission on Teacher Credentialing (CTC), the state governmental agency certifying teachers to work in the public schools, recently created new tests for assessing the subject matter knowledge of candidates who have not completed a CTC-approved program of study (within California) or who were educated out of state. Featuring constructed-response questions, the new Content Area Performance Assessments (CAPAs) are better able to measure the examinee's depth and breadth in the subject area as well as involve candidates in more complex performance tasks and cognitive processes.

The CAPA tests were developed after an extensive validity study was conducted in 1986/87 of the fifteen National Teacher Examination (NTE) Specialty Area Tests. Since California uses the fifteen NTE tests to make decisions regarding the subject matter competency of many credential candidates, such a review was deemed necessary to determine how well and to what extent these specialty area tests assess subject matter knowledge. The study revealed that important content knowledge in ten subject areas (English Language and Literature, French, Life Science, Mathematics, Music, Social Science, Spanish, Physical Education, Physical Science, and Art) could not be assessed adequately through multiple-choice questioning. It was recommended that supplementary tests be developed that would assess this important content knowledge through constructed-response questioning. Given the state's licensure structures, there were two additional recommendations — 1) to assess only subject-matter knowledge that candidates should have mastered by the time they earned their bachelor's degree and 2) to exclude pedagogical information that candidates would typically study in California's professional teacher education programs (one-year programs that follow undergraduate work).

The Content Area Performance Assessment program was launched in 1989 with advisory panels established in each of the ten subject areas. The panels included teachers currently working in the schools, university teacher educators, subject matter specialists from the Department of Education, and CTC consultants. In addition, since Educational Testing Service (ETS) was to be involved in the development of the CAPA tests, an ETS specialist also worked with each panel. The panels were challenged to review the validity study in the subject area, develop new standards for the academic preparation of teachers, draft specifications for their CAPA test, and write and review test questions.

CAPA Test in Music

One of the ten subject areas tapped for this new performance assessment was music. The music advisory panel first reviewed the California Validity Study of the NTE Specialty Area Examination in Music Education. The validity study revealed that the test was seriously out of alignment with California's standards and curriculum. Specifically, the report noted that the NTE in Music Education did not adequately assess performance skills in conducting, primary and secondary instruments, nor pertinent information in vocal and instrumental music. Recommendations for new versions of the NTE or a California measure included:

Don't Forget to Submit your papers for the Measurement and Evaluation Session in Cincinnati!!

- placing more emphasis on performance skills;
- providing separate assessment measures in vocal and instrumental music; and
- incorporating pedagogy and methodology as they apply to music teaching.

Test Specifications for the CAPA Test in Music

Based on the validity study recommendations and current developments in the field, the music advisory panel drafted new standards for the academic preparation of music teachers in California. Since approximately one-half of California's music teachers enter professional teacher education programs by taking the NTE in Music Education rather than through an academic program, the panel recommended that the NTE in Music Education and the CAPA test must assess the same understandings and skills articulated in the standards. The panel proposed a two-hour constructed-response test (CAPA) to cover those performance and methodology areas not measured in the multiple-choice questions in the NTE Specialty Test in Music Education. Specifically, the panel recommended that the constructed-response component cover instrumental or choral music developmental learning processes, a topic in classroom/general music, evaluation of musical scores for age appropriateness and essential characteristics of the music, and error detection (related to technical interpretation) of recorded performances. The panel also recommended that the test be aimed not at the scope of each examinee's preparation, which is assessed by the multiple-choice Specialty Area Test, but at the depth of his or her knowledge in crucial areas. Thus, in two of the questions, the panel proposed that examinees be allowed to choose from several topics, either from choral or instrumental music, or from choral, instrumental, or general music. This choice focuses the task on a topic within the examinee's area of specialization but, in turn, can require more from the examinee than a cursory summary of steps or points. These recommendations resulted in the following test specifications for the Content Area Performance Assessment in Music.

I. One essay question designed to assess the examinee's understanding of performance techniques. The examinee is asked to describe correct performance techniques related to a specified topic or remedial techniques appropriate for a specified performance problem. Examinees have a choice of two prompts, one relating to instrumental and one relating to choral music (30 minutes).

II. One essay question on general music designed to assess the examinee's ability to design and describe a step-by-step plan focusing on a music concept (30 minutes).

III: Two questions designed to assess the examinee's ability to aurally detect technical interpretation errors in performance. The examinee, using the musical score for reference, is asked to locate errors by measure number and to describe the errors. One question is based on an instrumental excerpt and the other on a choral excerpt (15 minutes each—total of 30 minutes).

IV. One essay question designed to assess the examinee's ability to evaluate the appropriate age-group level or level of difficulty of two musical scores (printed in the test book); to describe the essential cultural, historical, or stylistic characteristics of each piece; and to identify particular performance challenges in each piece or features in each piece that could appropriately demonstrate certain music concepts. Examinees choose a pair of musical scores from one of three areas: instrumental music, choral music, or general music (30 minutes).

Presenting the CAPA Test

After drafting the specifications, the panel prepared three forms of the test for pretesting. During the 1991/92 academic year, California examinees taking the NTE Music Education test could also elect to take the CAPA test. After each administration, the music panel read examinee responses to evaluate whether the questions were appropriate. The panel evaluated the questions for their appropriateness in terms of: the level of difficulty; the time allotted for responding; the clarity of the directions and the topics; the familiarity of the topic to examinees from a wide range of curricular preparations; and the comparability of each question to other questions. Finally, the panel decided whether each question elicited information that would help an evaluator determine whether an examinee was adequately prepared in the subject area.

In reviewing the pretest results, the music panelists found ways to refine the assessment in almost all areas of the specifications. Question type IV, for example, underwent substantial revision. The original wording of the question asked examinees choosing instrumental or vocal scores to identify the level of difficulty of each score and defend their choice, then discuss potential performance challenges each score might pose for a group of student musicians. After reading the pretest responses, the panel found that examinees were spending too much time defending their choice with rather mundane observations about the music, then skimming from detail to detail in the score, neither projecting a compelling overall conceptualization of the piece nor focusing enough on its most significant stylistic and performance elements. The panel revised the question, requiring examinees to focus on the key stylistic influences exemplified in the piece, two important performance challenges, and rehearsal techniques that would help players meet the two challenges.

Question type II, asking examinees to plan a demonstration of a music concept, also required refinement in order to produce responses that better reflected each examinee's preparation in the general music area. The initial wording of the question asked for a plan of a music-learning experience focusing on a specified concept. Responses ranged from elaborate curriculum plans for sequential lessons on the concept to detailed expositions of a ten-minute demonstration of one aspect of the concept. The panel revised the question, to the planning of one lesson. This lesson to include: at least one musical selection, at least one participatory experience, and suggestions for future reinforcement of the concept. In this way, all examinees demonstrate their skills and knowledge within the same parameters.

Scoring the CAPA Test

The questions on the new performance assessment were not merely a challenge to create and refine, but also a challenge to score. The panel discovered that the scoring guides for readers who were to score the responses needed to reflect the types of tasks that were asked of examinees, but the types of questions that were asked, in turn, had to be compatible with a scoring guide that was useful for discriminating among various levels of examinee preparation. The panel found itself working back and forth from question type to scoring guide to reach this delicate but critical balance.

The panel devised three different scoring guides: a holistic guide for question types I and II, an analytical guide for question type III, and a different analytical guide for question type IV. The holistic guide describes each score point from 0 to 5 in terms of how well the examinee demonstrates understanding of the concepts and skills presented in the question, how correctly and thoroughly the examinee answers the question, the appropriateness of the musical examples cited by the examinee, the variety of cultural origins and musical traditions represented in the musical examples cited, the soundness of the sequence in which the musical concepts and/or strategies for skill development are presented, and the accuracy and spelling of musical terms.

The analytic scoring guides for question types III and IV instruct the reader to assign points for each task correctly completed: for example, the examinee receives one point for each correctly located and described error in question type III. The panel members believe that these scoring guides appropriately parallel the tasks asked of the examinees and will ensure greater consistency than would a holistic scoring guide.

The panel also found it helpful to draft question-specific scoring guides to supplement the general scoring guide for each question. In this supplement, the panel identifies the key concepts or content areas that must be considered in determining whether the question has been answered completely, used appropriate details and examples, or referred to significant parts of the musical scores. While question-specific guides help define the scoring dimensions for particular questions, the general scoring guides are essential for ensuring comparability of test scores from one test form to another and from one scoring session to another.

Beginning in November 1992, when the CAPA music test became operational, a procedure for scoring the tests was established. After each administration of the test, ETS invites a number of music specialists—both university teacher educators and teachers currently working in the schools—to attend a two-day scoring session. A Chief Reader, presently a university teacher-educator who met with the panel while the test was being developed, leads the scoring session by introducing the readers to each topic, explaining the scoring guides, preselecting sample papers and presenting them to the readers as benchmarks at particular score points, and checking readers' work as the

scoring proceeds. Any time a particular question is reused in a later form of the test, the same scoring guides and the same benchmarks are used to ensure comparability of scoring across different administrations.

During the reading, each examinee's identification is masked, and each book is given to as many different readers as possible. Each response receives two independent readings. If the two scores differ by more than one point, the response receives a third independent reading by the Chief Reader. This procedure is intended to protect examinees from erratic or miswritten scores. The reading procedure is flexible in that it can accommodate small or large volumes: the number of readers need only be increased.

CAPA's Future

The CAPA music test and the other nine CAPA tests were developed at the same time that ETS was developing its own new generation of teacher tests, called The Praxis Series: Professional Assessments for Beginning Teachers. The Series consists of three stages which parallel the stages in a prospective teacher's preparation: Praxis I: Academic Skills Assessment, tests basic skills in reading, writing, and mathematics; Praxis II: Subject Assessments, tests subject-matter knowledge; and Praxis III: Classroom Performance Assessment, evaluates new teachers in the classroom. For the Praxis II tests, the offerings in each subject area consist of several modules rather than a single multiple-choice test. The Core Module is a two-hour multiple-choice test intended to cover the knowledge and skills essential for all teacher candidates in a given subject area. Optional one-hour modules, consisting entirely of essay or other free-response questions, are also offered: they focus on particular aspects of the subject or on ways of teaching the subject. A state can decide on the best combination of modules from the "menu" of available modules in a given area.

For seven of the ten subject areas covered by the CAPA tests in California, ETS began its own independent development of Praxis II modules in 1989. But for Music, French, and Art, the lowest-volume tests of the ten areas, Praxis development did not begin in earnest until 1992. It was decided that the CAPA tests in Music, French and Art might fit into the Praxis Series as free-response modules, so ETS examined the CAPAs from a national perspective as well. They were deemed valid according to the job analysis surveys that were conducted nationwide in each subject area. Since the Praxis Series had developed free-response modules in lengths of one hour, each two-hour CAPA test was split into two one-hour modules. Thus, the CAPA music test was divided into "Concepts and "Processes," including question types I and II, and Analysis," encompassing questions types III and IV. The national Committee of Examiners in Music for the Praxis Series is currently developing the Core Module and the Music Pedagogy Module and has requested that work begin immediately on a Performance Module.

References

Educational Testing Services. (1992). *Content Area Performance Assessments: Music leaflet.* Princeton, NJ: Educational Testing Service.

Educational Testing Services. (1992). *The Praxis Series: Professional assessments for beginning teachers: 21st century teacher assessments—Leaflet.* Princeton, NJ: Educational Testing Service.

Wheeler, Patricia, Judy Bonnett Hirabayashi, Jean Martinson, and Richard W. Watkins. (1988). *A study on the appropriateness of fifteen NTE Specialty Area Tests for use in credentialing in the state of California, Vol. II.* Emeryville, CA: Educational Testing Service.

Editor's Desk

Richard Colwell

In a previous newsletter, your officers asked for information about the use of portfolio evaluation in music classrooms. Questions that have come into this office are representative of those raised with the use of any evaluation device. What evidence is there about usability, reliability, and validity of the portfolio for the various types of music instruction? Is the internal reliability affected by weighting of items within the portfolio? How does one establish weighting and does this weighting change between first and twelfth grade? If the portfolio has content validity, to what extent does it have criterion related and construct validity? How should grades be derived from a portfolio when every student's portfolio differs in content? Is a portfolio based on six years of instruction more or less valid than a portfolio based on one year? If portfolios are used in teacher evaluation, what is the potential for use of portfolios in undergraduate teacher education programs? Would not a portfolio about music theory instruction be one of the most useful and would teacher evaluation portfolios be based on individual courses or on a year's program of integrated experiences and courses? Can college professors advocate portfolios to the public schools if they have not demonstrated their usefulness in college music education programs? The number of questions coming into the office exceed the number of answers. The Measurement and Evaluation SRIG session at the Cincinnati conference will be devoted to issues in authentic evaluation. If you know of someone who has given serious attention to the questions being raised and would like to be on the convention program, have them send some supporting data to: Richard Colwell, 12 Peacock Farm Road, Lexington, MA, 02173 as soon as possible.

In the last newsletter, your editor's comments expressed enthusiasm for the evaluation component of the proposed national standards. Nothing would please the SRIG members more than to have the profession give serious attention to the ways of providing decent feedback to teachers and students (diagnostic evaluation) and to establishing indicators of acceptable and unacceptable performance as a result of music instruction. As far as we know there is no consensus on either task.

For music education to benefit from this tremendous effort, we hope that instruction will be improved as well as programs "saved." We hope that the members of the philosophy SRIG are working on issues such as what sorts of programs should be saved, issues that extend beyond the immediate concerns of the evaluation SRIG. We have expressed four cautions: (1) obtaining data useful for standards in a voluntary evaluation program, (2) the difficulties of incorporating an evaluation program into a curriculum that has no tradition of systematic evaluation other than public performance, (3) the possible misconceptions arising out of an arts evaluation, and (4) issues arising from evaluating untaught objectives. There are other, perhaps more important cautions as well - for example, will the instruction to be evaluated be conducted by specialists or teachers, and with what frequency, moral support, and equipment and facilities.

To accomplish the objectives related to arts advocacy, the profession has united with visual arts, drama, and dance teachers to convince policy makers of the importance of the arts, of having suggested national standards, and of proceeding with a national assessment program. In a united effort, the effort is no stronger than its weakest link. Our narrow experience would indicate that there are many school systems with uneven arts education programs. The recurring nightmare is that the results of this national standards project will be released to the press as a collective effort, not indicating the strength of the music program and/or the needs of the dance program. Release of collective data is becoming more likely as our philosophers are suggesting to school administrators that they should think in terms of goals for the arts rather than in subject matter goals for music, visual arts, theater, and dance. We have to be convinced of any commonality in the teaching of objectives of the various arts programs; our uneasiness about evaluation data is a natural result. Differences abound. Classroom teachers appear more willing to supplement instruction in visual arts than in music and there are more products from most visual arts programs than there are from dance and music programs. I sup-

The *SRIG Measurement and Evaluation Newsletter* is moving. Please address all correspondence to:

Richard Colwell
12 Peacock Farm Road
Lexington, MA 02173

pose there are school systems where the arts teachers plan collaboratively but we have become accustom to the expressed concern that all of the music teachers cannot find time to work and plan together. Although it may not be in keeping with whole language philosophy, we would suggest that music educators carefully assess their own needs in a national standards and evaluation process before undertaking the ambitious project suggesting learning commonalities among the arts. The sciences and math can unite under an umbrella such as astronomy, a subject that uses knowledge from the separate disciplines, but the arts are not uniting under such a structure.

The fourth caution concerns the careful identification of those skills, knowledges, attitudes, understandings, and the like, that are taught systematically in the schools as opposed to attainments gained through the culture, private lessons, and community activities. Many of the outstanding music programs are located in supportive communities where students learn outside of school. There are many ramifications in this situation for an evaluation program. It is not inconceivable that students will be more adept at untaught objectives by the school than with objectives systematically taught. How does one separate failure at objectives attempted from failure at objectives not attempted when the culture can be such a powerful influence? It seems to us that students learn louder and softer, faster and slower, what a trumpet sounds like, and the distinction between male and female voices outside of school although music teachers often take credit for accomplishing these instructional objectives. Many of the items in a portfolio are subject to the same concern. Society's objective is to have musically educated citizens, with less concern for the derivation of that ability.

Our point is that a major objective of the proposed project is to establish that the arts are basic and need to be systematically and sequentially taught. Interpreting data is always a challenge; with national visibility on our efforts the task is daunting. What did we learn about interpreting data from the first two national assessments and who has that knowledge that will enable us to build upon our past successes and failures? Music educators are not building upon the same level of experience in evaluation as are the other basic subjects and the task of evaluation appears to at least, if not more, complex.

Boston University
School for the Arts
Department of Music Education
855 Commonwealth Avenue
Boston, MA 02215
(617) 353-6888
(617) 353-6808 (FAX)

Address Correction Requested

Attention: Richard Colwell and Robert Ambrose, Editors