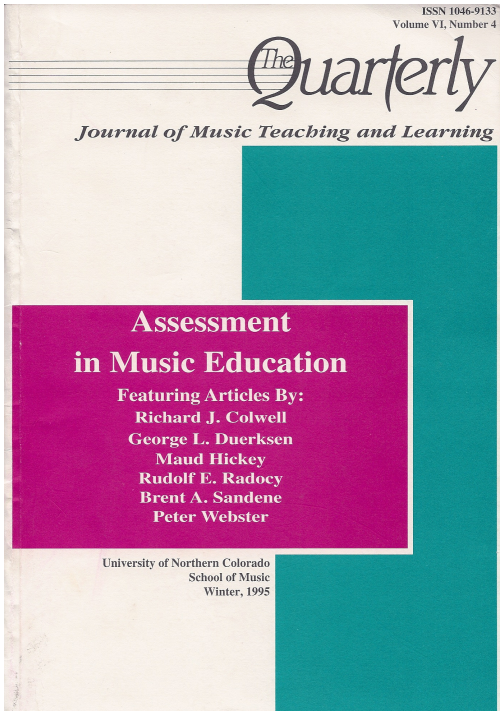

Title: Suggestions for Conducting Large-Scale Assessments in Music

Author(s): Brent A. Sandene

Source: Sandene, B. A. (1995, Winter). Suggestions for conducting large-scale assessments in music. *The Quarterly*, *6*(4), pp. 45-55. (Reprinted with permission in *Visions of Research in Music Education, 16*(6), Autumn, 2010*). Retrieved from* http://www-usr.rider.edu/~vrme/

Visions of Research in Music Education is a fully refereed critical journal appearing exclusively on the Internet. Its publication is offered as a public service to the profession by the New Jersey Music Educators Association, the state affiliate of MENC: The National Association for Music Education. The publication of VRME is made possible through the facilities of Westminster Choir College of Rider University Princeton, New Jersey. Frank Abrahams is the senior editor. Jason D. Vodicka is editor of the Quarterly historical reprint series. Chad Keilman is the production coordinator. The *Quarterly Journal of Music Teaching and Learning* is reprinted with permission of Richard Colwell, who was senior consulting editor of the original series.

Suggestions For Conducting Large-Scale Assessments In Music

By Brent A. Sandene

Educational Testing Service

Assessment is becoming increasingly important in all areas of education. As educators seek to further the place of music and the other arts in school curricula by adopting voluntary or mandatory standards ("AMC News," 1995), the need to assess student attainment of those standards will rise. Although large-scale (i.e., district or state level) assessment projects have been ongoing in subjects such as mathematics and reading for some time, large-scale assessment projects in the arts have been limited. Although music educators have dealt with some of the issues associated with testing and measurement for over seventy years, new generations of music educators at the state and district level may now be facing both practical and theoretical issues associated with large-scale assessment.

Although large-scale (i.e., district or state level) assessment projects have been ongoing in subjects such as mathematics and reading for some time, large-scale projects in the arts have been limited.

Despite these obstacles, and challenges inherent in assessing the subject of music, large-scale efforts to assess both cognitive and performance skills in music have been encouraging. For example, the National Assessment of Educational Progress (NAEP) assessments in music completed in 1971-72 (NAEP, 1974) and in 1977-78 (NAEP, 1981) have demonstrated that it is possible to conduct large-scale surveys that indicate what students know and can do in music. Other efforts have been encouraging as well, e.g., based upon a successful trial administration held in 1995, the Advanced Placement (AP) Exam in Music Theory now includes sight-singing.

The purpose of this article is to contribute to the continued successful implementation of large-scale assessments in music by summarizing and describing basic steps in conducting an assessment, addressing large-scale assessment issues, and providing practical suggestions that are based upon personal experiences in working with the NAEP Assessment in the Arts, the Advanced Placement (AP) Examination in Music Theory, and a review of relevant literature.

Planning and Designing an Assessment

Planners and designers should develop activities and outline schedules based on assumptions that assessment projects will most likely require refinement and a period of revisions before they can be constructed, pre-

Brent Sandene is a Program Administrator at Educational Testing Service. His research interests include student motivation, testing and measurement.

Framework designers should investigate ways in which assessment results can be linked to policy-relevant school and student background variables in the reporting of assessment results.

tested, and administered operationally. Initiating assessment projects will most likely involve representatives from a wide range of constituencies, including teachers, administrators, elected officials, and parents. All policymakers involved in establishing and designing an assessment project should reach a consensus on the particular goals that will guide the assessment, and should work to establish its base of support. The purposes of an assessment should be stated explicitly so that evaluative procedures can be implemented to ensure that the goals of the assessment are being met prior to its being conducted.

If the assessment is to be a widely accepted and highly regarded tool, it must be managed by individuals who represent the political and philosophical perspectives of all constituencies involved. Typically, at a macro level, political considerations can involve advocating specific curricula, texts, or instructional methods, or anticipating how the results of the assessment will be used once it has been reported. At a micro level, views held by individuals, for example, regarding the selection of particular pieces of repertoire or advocacy of music of a certain style or genre, will inevitably arise and must be addressed. To enable everyone involved to have a sense of ownership in the project, all reasonable ideas or policies proposed by individuals should be open for consideration. Whenever possible, representatives should be fully informed about all phases of the project and should be consulted regularly.

Purposes of large-scale assessment projects can vary depending on the needs of the jurisdictions involved. Basic purposes of large-scale assessment projects can include: public reporting at the local, state, or national level about student knowledge and abilities; identifying needs and allocating resources within schools or districts; making judgments about promotion and graduation (Kearney, 1983); informing policy; assisting curriculum reform; and making schools accountable (Cooley, 1991). The degree to which any of these purposes should be advocated has profound implications for the ways in which the assessment should be designed, implemented, and reported. Without clearly understood goals, assessments can be directed towards purposes for which they are not well-suited. Different assessment goals lead to markedly different designs and pose varied sets of concerns. For example, Cooley (1991) decries the frequently ineffective use of assessment results to inform state policy:

> Unfortunately, this [clarity of purpose] seldom happens because state testing programs tend to be viewed as accountability mechanisms, rather than policy guiding mechanisms. As a result of the accountability emphasis, policy-relevant variables are neither collected nor integrated with test results, so that educational practices and policies are not easily linked to outcomes. Thus, what states tend to do is publish district or school results in the hopes that public display of low performance will "embarrass the inept into action." One problem, of course, is that there are many reasons for a school's low performance, and "ineptness" of staff is only one such possibility. Also, it is not clear from test results what a low performing school might do to improve performance, especially if ineptness happens to be the problem. The states tend to be monitoring in such a way that neither the state nor the districts learn how to improve schools. (p. 3)

Establishing an Assessment Design

Once a consensus advocating a rationale for the assessment has been reached, a detailed design framework for the development of the assessment should be created. An effective assessment framework should contain several elements, including a statement of purpose, an outline of the specific content area to be measured, examples of test exercises and scoring guides, and plans for the reporting of results. The NAEP Arts Education Assessment Framework (1994) can serve as a useful refer-

Considering the reactions of teachers is especially important, because teachers are especially affected by assessment projects and will help to inform the reactions of parents, students, and administrators to results.

ence or model for constituencies interested in developing assessments based on the National Standards for Music Education.

Assessment designers must address both practical and theoretical concerns. For example, it is important to consider the experience levels of the students that will be taking part in the assessment, including exposure to a common curriculum, and the characteristics of the music programs in which they are enrolled. In addition, issues such as the desirability of assessing whether students value music or use music for "aesthetic" means must be resolved.

Framework designers should investigate ways in which assessment results can be linked to policy-relevant school and student background variables in the reporting of assessment results. For example, when resources permit, and privacy issues can be addressed, school variables such as amount of instructional time devoted to music and the content of the school music curriculum, and relevant student variables regarding home and school experiences should be included in the assessment and reported along with student achievement variables.

Mehrens and Popham (1992) have outlined useful procedures that assessment developers should consider when implementing and evaluating the results of "high-stakes" tests (e.g., tests in which the results seriously impact the status of students, teachers, or music programs, such as those used to determine promotions, graduation, admission to honors programs, or decisions concerning teacher advancement). These include suggested procedures to follow when designing and conducting an assessment, the development of reliable and valid scoring criteria, setting of standards, and procedures to follow to ensure that tests are free of bias.

Lehman (1992) has summarized potential design flaws frequently occurring in program evaluations that are also relevant in designing large-scale assessments. Flaws to be avoided include the use of inappropriate measures, basing the design on flawed assumptions about the curriculum or its implementation, or holding misleading assumptions about the population of students or teachers in the study.

Potential Effects of a Large-Scale Assessment Project in Music

Establishing an ongoing large-scale assessment project in music will undoubtedly have unintended as well as intended effects. Positive effects of an effective assessment can include sharing of information about the status of students' achievement, as well as providing informed guidance to assist the design and revision of curricula. Assessment designers should consider possible negative effects, however, and should plan strategies to prevent them from occurring. For example, studies in subject areas other than music have found that unintended effects of external or mandated testing in schools can include a reduction of time available for instruction, teachers' neglect of important curricular material that is not tested, development of classroom instructional methods that resemble tests, restriction of student access and opportunity within the program (Smith & Rottenberg, 1991), unethical conduct of teachers during the assessment, and a devaluing of the test by staff members (Moore, 1994). Considering the reactions of teachers is especially important, because teachers are especially affected by assessment projects and will help to inform the reactions of parents, students, and administrators to results.

Operational Issues

Planning and pilot testing is important in all stages of the assessment process. Fisher and Smith (1991) briefly outline issues for assessment planners to consider. Among their most important recommendations are to build sufficient time into each phase of the

Regardless of their format, performance tasks should be designed to give students enough directions to guide their performance and to let students know how they will be scored.

schedule, and to keep a log of problems that were encountered, their causes, and solutions. Perhaps the most crucial element to consider is the development of the overall schedule for assessment development, production, administration, scoring, and evaluation/reporting. The unique nature of assessment projects in music requires meeting scheduling challenges associated with the production of recordings, use of equipment, and the processing and storage of materials. It is crucial to have staff available to coordinate all phases of the assessment and to ensure that all important deadlines are met.

Pilot testing should include trial administration of a variety of item types and formats, difficulty levels, stimulus materials, and scoring methods. After pilot test data have been collected, a variety of methods for summarizing and analyzing the data can be tested in order to determine which is most effective. In addition to providing an opportunity to scrutinize general procedures used in conducting an assessment, pilot testing procedures are useful in determining the true costs associated with conducting the assessment at a full-scale level. As pilot testing proceeds, it is important to gain feedback from everyone involved in the project, including students, teachers, scorers, and policymakers.

Assessment Item Selection and Design

As planning proceeds, written exercises and performance tasks must be created or adapted from outside sources. Assessment developers should consider issues of copyright and potential costs of producing stimulus materials as exercises and tasks are being developed. A variety of item formats have been used in various large-scale assessment programs, such as multiple-choice questions, questions requiring written responses, performance-based tasks, and portfolio assessment. Issues associated with the development of various item types have been addressed throughout the field of educational psychology (Sax, 1989), as well as in music (Boyle & Radocy, 1987).

If implemented properly, performance tasks are appropriate to assess student achievement in a wide range of musical skills and abilities. Assessment developers should be aware of any need to create equivalent test forms or tasks, however, and should seek to avoid the problems identified with the use of performance assessment techniques in other curricular areas. These may include the lack of psychometric properties that have been identified in traditional multiple-choice tests (such as the opportunity to create scaled scores) (Miller & Legg, 1993), problems with test security (Mehrens, 1992), the tendency for tasks tend to be context-bound and not generalizable within a domain (Dunbar, Koretz, & Hoover, 1991), costs, logistical issues, technical (statistical) concerns, and the lack of support for their implementation among practitioners (Aschbacher, 1991). Stiggins (1987) and Baron (1991) offer useful general overviews and guides for the development of performance tasks in assessment projects.

A delicate balance between standardization and authenticity must be achieved when designing performance tasks. Many useful classroom-based assessment activities, although they are authentic and appealing, may not be appropriate for a given assessment. For example, some tasks that are integral to a well-balanced music curriculum based on the National Standards for Music Education, such as improvisation and composition tasks, may be relatively unfamiliar to many students. If possible, varied types of tasks and music should be pilot-tested in order to determine which repertoire and task design are most appropriate for the purposes of the assessment. Other techniques, such as portfolio assessment, may be useful if the appropriate teacher and student resources are used in preparing them and there are avail-

able funds and expertise to ensure the implementation of appropriate scoring criteria. Regardless of their format, performance tasks should be designed to give students enough directions to guide their performance and to let students know how they will be scored. Care should be taken, however, not to overwhelm students with unduly complex vocabulary or reading burdens, or instructions that may be intimidating or that take excessive amounts of assessment time to administer.

Arter and Spandel (1992) suggest that the design of exercises should stem from the results of a combination of *top-down* and *bottom-up* processes. Tasks that are *top-down* stem primarily from curricular objectives, state mandates, and newly implemented standards, and may primarily reflect the influences and interests of measurement experts, administrators, or those in higher education. Although *top-down* tasks may offer the advantage of standardizability, they may lack authenticity from the point of view of students and teachers. Assessment tasks that come from the teachers (*bottom-up*) may be authentic but often lack constraints of standardization necessary for assessment items. As Aschbacher (1991) has pointed out, caution should be used in implementing assessment techniques that are "too far ahead of instructional practices" (p. 285) because teachers may react to them unfavorably and may perceive the assessment as a threat to their programs. Therefore, it is useful to involve the input of both measurement experts and classroom practitioners throughout the design and selection process. This will ensure that items and tasks will have sufficient levels of standardizability but will still be reflective of the activities in which students most frequently engage.

Careful attention should be devoted to writing and reviewing scoring guides for non multiple-choice questions at the same time that exercises are being developed and reviewed. Frequently, issues arising from the review of scoring criteria enable item developers to prevent potential problems when developing items. Preliminary scoring guides for performance tasks and open-ended questions that outline ranges of proficiency can be written based on general hypotheses of what the most likely student responses will be. Quellmalz (1991) has outlined several points of consideration for the development of appropriate scoring criteria, including: significance, fidelity to the task, generalizability within the domain, developmental appropriateness, the degree to which criteria are accessible and clear to participants, and the overall utility of the criteria results with regard to the improvement of instruction.

Scoring guide developers should consider which dimensions (e.g., rhythm, tempo, creativity) should be scored on each student response and how many scoring levels should be assigned to each dimension. It is useful to consider the general purposes of the assessment, the practicality of the number of scoring levels in making reportable distinctions, the number of student responses that are likely to achieve each of the score points on the scale, the degree to which dimensions scored are correlated with each other, and the number of meaningful distinctions between levels of achievement that scorers can reasonably discern from the samples of students' work. In general, using a wider possible range of score points per item may offer statistical advantages if the policy does not result in undue difficulty in training scorers to make reliable distinctions between levels.

Judging multiple criteria or dimensions may be more time-consuming than assigning global ratings or assigning holistic scores to performances or written responses; however, it may be desirable for certain types of tasks or items. Depending on the purpose of the assessment and the resources that are devoted to scoring, a variety of item and task scoring rubrics can be included in any single assessment.

Controlling Assessment Costs

Assessment projects must be designed and conducted carefully in order to use allotted funds and staff time efficiently. Design of the assessment should include consideration of available funds for development, administration, and scoring. Costs of designing and pilot-testing items and tasks can be reduced by relying on those developed for use by other jurisdictions or organizations. Assessment items and tasks should be reviewed or revised, however, to ensure that they meet

...group performance tasks can be especially cost-effective and can also serve as models for classroom assessment activities. Group performance tasks, however, limit the ability of assessment developers to make judgments about an individual student's level of achievement.

the unique needs of the jurisdiction that will adopt them. In any case, decisions made for reasons of economy should be balanced with assessment goals. For example, as Hardy (1995) points out, group performance tasks can be especially cost-effective and can also serve as models for classroom assessment activities. Group performance tasks, however, limit the ability of assessment developers to make judgments about an individual student's level of achievement (Webb, 1993).

The high costs of scoring performance assessments can be easily overlooked, but should be given special consideration as items are developed and reviewed (Hardy, 1995). This is especially important in conducting performance assessments because, at the present time, the opportunity to use machine scoring of performance assessment tasks is limited. Scoring of an assessment can be completed by either participating classroom teachers or consultants hired on the basis of their expertise. As Hardy (1995) has written, devoting a large portion of assessment costs toward compensating teachers, rather than consultants, for their time and expertise as scorers can be cost-effective. Allocating paid staff or inservice time to scoring projects can increase teacher involvement, raise enthusiasm, and increase support of assessment efforts. Moreover, by examining a wide range of student responses, teachers can gain a perspective on their students' achievement and can better prepare their students for subsequent assessment activities.

Some types of items are relatively easier, and therefore cheaper, to score. For example, scoring short answer questions or assigning holistic ratings to extended constructed-response questions may be cheaper than evaluating students' portfolios, which require examination of a wide range of diverse materials. Asking scorers to provide specific comments regarding the strengths and weaknesses of each student's responses or classifying the types of errors observed as scoring proceeds can also increase scoring costs. However, providing such diagnostic information can be extremely useful to students and teachers taking part in an assessment program if the program aims to generate data at these levels.

Testing only a sample of students in a jurisdiction offers another way to reduce assessment costs while still providing information, although as Hardy (1995) points out, it limits the degree to which feedback can be provided to individual teachers and students. In addition, Hardy states, "... policymakers at the state level often do not trust the results unless every student is tested. Aschbacher reports that, in at least one state, the sampling of students for performance assessment 'led teachers and administrators to think of performance based activities as enrichment, not as mainstream assessment and instruction.'" (p.8)

In addition to considering the political implications of a sample size, it is also important to consider its technical effects upon measurement error when designing assessment projects. Acceptable rates of measurement error may vary, depending on the purpose of the assessment and the scope of the project. Testing larger sample sizes of students leads to lower rates of standard error and increased measurement power; however, in some cases the precision obtained with lower rates of standard error may be largely outweighed by the prohibitive additional scoring and administration costs. Policymakers should consult with measurement experts in order to determine the needed sample size in order to ensure ac-

...it is important that the material that is used in an assessment be examined specifically for issues of potential bias and use of stereotypes, or for the use of items which could offend students' religious, cultural, moral, or political sensibilities.

ceptable rates of standard error for any given assessment task or item. As Koretz et al. (1994) have pointed out, when scores are unreliable, or too few students are sampled per classroom, school, or district, serious challenges to the integrity of the assessment and its usefulness as an evaluation or accountability mechanism can be raised.

Addressing Issues of Cultural Sensitivity and Test Bias

Large-scale assessment projects in music present numerous challenges with regard to the selection of potential stimulus materials (e.g., audiotapes or videotapes, written materials, administration scripts). A comprehensive music curriculum and a diverse student body require the use of a wide range of music. Therefore, it is important that the material that is used in an assessment be examined specifically for issues of potential bias and use of stereotypes, or for the use of items which could offend students' religious, cultural, moral, or political sensibilities.

Issues of sensitivity can involve the selection of music, the wording of the questions, the lyrics of songs, the depiction of individuals in photographs, the roles in which members of various groups are depicted, and the design of music creation and performance tasks. Cultural norms, for example, may influence the degree to which students will engage in various music performance tasks. Solutions to potentially problematic stimulus materials include dropping the stimulus material and the items associated with it entirely, rewording questions, revising song lyrics, or adding additional stimulus materials throughout the assessment in order to remedy the balance of cultures and/or representation of groups depicted.

After the assessment has been scored and if the number of students sampled is sufficiently large, calculation of DIF (Differential Item Functioning) statistics for each item or task in the assessment may be completed. By using DIF statistics to detect discrepant rates of performance with regard to gender or ethnicity, potential hidden item biases may be detected and remedied.

Training of Assessment Administrators

After assessment items have been developed and/or selected, individuals must be trained to administer the assessment using standardized procedures. Although it is possible to hire individuals from outside a jurisdiction, in many situations, using classroom teachers as assessment administrators may be an advantageous option. As Hardy (1995) notes, using classroom teachers to conduct assessment activities can inform staff about useful classroom assessment techniques and instructional practices. The training of teachers to administer an assessment can be integrated with other assessment-related activities and will be informative and useful if organized and implemented effectively.

Administration manuals and scripts must be developed that provide clear directions to administrators and students. It is useful to have training sessions with all test administrators involved. This gives the assessment planners the opportunity to explain the rationale for each of the tasks, to demonstrate the administration of the tasks, to answer questions about the procedures involved, and for everyone involved to gain a common understanding of the goals of the project. Meetings can also address important details of standardization and logistical procedures to be followed. Two issues of standardization that are relevant to assessments in music are the amount of time that students are given to complete their work, and the ways in which students' questions are answered. In addition, attention should be paid to securing the minimally acceptable space, acoustic condi-

The setting of standards when scoring and evaluating the results of large-scale assessments is enormously important and requires that serious issues of semantics and pedagogy be confronted.

tions, and equipment for the assessment.

It is important that administrators are trained to use equipment properly in order to make quality audiotape or videotape recordings of student performances. Directions for the labeling and processing of all materials need to be carefully planned so that no materials are misplaced. Developers should consider the need to have students' schools or teachers be identifiable during scoring and should create the directions for the labeling of materials accordingly.

Scoring

Unlike machine-scored multiple-choice items, scoring students' creative products and performance tasks presents unique challenges. The potential to use technology for scoring depends on the way in which tasks and questions have been designed. Successful computer scoring of student essays created using computers has already been demonstrated (Page & Petersen, 1995), and researchers most likely will develop easier ways of scoring students' handwritten responses for future assessments. Music software may be useful for scoring certain types of music performance tasks.

When using expert scorers to evaluate students' work, a crucial step is locating and selecting actual student *anchor,* or sample, responses from the pool of responses. Student anchor responses should provide clear examples of the characteristics and information that should be present in each student response in order to receive credit for each level or point on the scoring scale. Whenever possible, a full range of diverse examples for each item or task should be selected so that scorers can gain familiarity with each of the various types of student responses. Because of their complexity and diversity, scoring music performance and creation tasks may require additional training, sample responses, and more detailed scoring guides than cognitive written items. After the full range of samples has been examined and prior to actual scoring, scoring guides can be revised in order to give specific information regarding unusual cases that arise.

The setting of standards when scoring and evaluating the results of large-scale assessments is enormously important and requires that serious issues of semantics and pedagogy be confronted. For example, if a performance task is to measure students' abilities to play improvisations, how good is good enough? What specific elements must a student display in order to receive a passing score? Or, for example, if an assessment task is designed to measure students' skill in composition as described in National Standard #4, scorers may raise the issue — what exactly constitutes "composing music creatively in using the elements of music for expressive effect?" How creative must a student be to meet the standard? Although each situation is unique, in general, scoring guides that are descriptive and that list characteristics present in performances or creative products may be more useful than scoring guides that use adjectives such as poor, fair, good, or excellent.

Training of Scorers

Although on-site scoring of performance tasks can give immediate feedback to students and teachers, in most situations it will be preferable to record performances on audiotape or videotape to be scored at a later time at a central site. Among the advantages of scoring at a central site are the potential to maintain higher levels of score reliability and an overall reduction in training time.

Competent supervisors/trainers should provide comprehensive training for the staff responsible for scoring the assessment. Training should include a complete introduction to the entire assessment, the specific items or tasks being scored, and the rationales upon which they are based. If possible, scorers

should actually complete the item or task (e.g., perform the sight-reading melody, write out the answer to the question) in order to become acquainted with it. Scorers should review the scoring guides developed for the item, the anchor or reference student papers or performances for each scoring point on the item, and then practice scoring a full range of actual student responses in order to check for accuracy and understanding.

It is important for scorers to be able to raise issues for discussion as training proceeds and once actual scoring begins. Frequently, scorers raise the "What if...." question. For example — what if the student starts the sight-reading melody over? What if the student answers the first part of the question in great detail but does not address the second part of the question? What if a student did not precisely follow the directions for a task but ended up with a superior musical product? General policies on how such divergent responses are to be scored must be articulated and documented.

The types of scoring issues that arise frequently are directly related to the type of assessment item or task being scored. Scoring items or tasks which have a relatively limited range of definitively right or wrong responses can raise issues of standards or interpretation (i.e., "How specific must the student's answer be in order to be scored as correct?" "How much of a dynamic contrast does the student need to show in order to gain full credit for 'dynamics'?"; "Does a student who slides into a pitch still gain full credit for performing the note correctly?"), whereas responses that are by their nature more creative or open-ended are likely to raise issues associated with classifying many diverse responses into a limited number of scoring classifications.

Scorers should demonstrate competency to score each item of the assessment that they will encounter. It is wise to implement an ongoing monitoring procedure led by a supervisor/trainer who can verify that standards are being maintained and who can provide remediation or intervention if necessary. Scorer competency can be determined through the use of either informal or formal procedures. Informal procedures include discussion of ratings awarded to the same item by scorers working independently. Formal methods can include techniques such as administering qualifying exams demonstrating competency. Policies regarding scorer reliability must be clearly articulated and implemented in order to ensure the usefulness of the data collected. The benefits of implementing formal procedures must be evaluated in terms of the added costs needed to accomplish them, and should be considered in planning scoring budgets.

Periodic review of scoring policies, anchor performances, and scoring guides during scoring is warranted. This is important to keep scores from drifting either higher or lower, due to scorer fatigue, or as a result of experiencing a series of student responses that are either consistently strong or weak. If scorers encounter an item or task that is rather easy or difficult for students to complete, scorers may tend to shift their expectations or begin to compare students' responses to each other, rather than to an a priori set standard that reflects a specific curricular objective or assessment specification.

Distributing all student samples of the same item to the same subgroup of scorers is efficient for training, processing materials, and providing for the highest possible rates of score reliability. If a primary purpose of the assessment is teacher inservice, however, and if the number of different responses to be scored is not excessive, scorers can rate a variety of different tasks or items in an assessment.

As scoring begins, careful records should be kept documenting all issues that arise and how they are resolved. Documentation is particularly important during pilot testing, if jurisdictions wish to complete quasi-experimental design assessments that seek to measure changes in students' achievement over time (analysis of trend), if scoring is not completed at a central time and location, and for the reporting of results. It is also important to keep an accurate log of the amount of time that is spent training and scoring each item in the assessment. This information is useful for planners responsible for calculating budgets and in allocating the use of staff time for future assessment activities.

As a result of examining a wide range of student responses in a pilot test, scorers may

Ultimately, the most important purpose of conducting any assessment in music is to inform a given group or constituency of the educational progress being made by students in any given jurisdiction.

be able to offer helpful ways of rewording questions, restructuring tasks, or suggesting what types of items seem to elicit the best and most informative student responses. In cases in which assessment activities are audiotaped or videotaped, scorers can document the performance of individual test administrators, ensure that all script directions have been followed, and verify that administration conditions were adequate.

Test Security Issues

Large-scale assessments present many challenges of test security. Regardless of the scope and purposes of an assessment, procedures should be implemented so that the opportunity for unethical conduct on the parts of students and teachers is limited. During all phases of the assessment, procedures must be developed to ensure that access to the items and to student scores is restricted to individuals who have a need to see the information. Should an assessment be high-stakes, developers can consider pilot testing items or tasks in jurisdictions that are not to be included in the main assessment. This procedure, however, can add to overall development costs.

Access to computer databases and copies of test materials also should be restricted. If items and/or tasks are going to be readministered in the same jurisdictions from year to year, procedures should be implemented to guarantee that the test conditions are not intentionally or unintentionally compromised. For example, Maeroff (1991) has described problems arising from teachers coaching students inappropriately for specific tasks in assessments that are repeated.

Dissemination of Results

Ultimately, the most important purpose of conducting any assessment in music is to inform a given group or constituency of the educational progress being made by students in any given jurisdiction. Assessment results should be presented in ways that avoid using jargon and that provide clear examples. Depending on the resources available, and when issues of student and teacher privacy do not prevent it, providing examples of actual student responses is especially useful. It is particularly important to provide parents and community members the contextual information they need to interpret the results of an assessment properly. Sufficient interpretation and explanation of the data should be provided so that assessment results are not used inappropriately.

As results are disseminated and if reports are issued, assessment developers should consider means of data summarization, analysis and interpretation they believe are most useful to each varied constituency. For example, if an assessment is designed to give feedback to individual students or teachers, reports can describe specific techniques to improve performance in areas that are seen as weak.

Ensuring Quality

Ensuring quality in assessment projects includes the establishment of procedures to ensure that all aspects of the assessment are completed with as few errors as possible. In assessments, this includes reviews of the test items to check for content accuracy and appropriateness, reviews of the production of all stimulus materials, such as audio or videotapes, sheet music, and other printed materials, and reviews of instructions, administration scripts, and the processing of materials.

Ensuring quality of assessment administrations includes on-site visits to actual administrations to ensure that all scripts are being used correctly, that instructions are administered to students in a standardized fashion, and that adequate space, acoustics, and facilities have been used in the administration of the assessment. Ensuring quality during scoring includes monitoring of training procedures, scorers' reliability rates, and the accuracy of coding data for analysis.

Summary

Music educators interested in developing assessment programs in their jurisdictions are faced with many new and interesting challenges. Planners of large-scale assessments in music should consider the resources and procedures that may already be in place in their jurisdictions for other types of programs that are conducted regularly, such as all-state and regional honor ensembles requiring auditions. Many of these programs offer potential procedures to resolve logistical problems associated with large-scale assessment projects. Further, issues relevant to large-scale assessment in music are being dealt with by educators in other fields. As projects in large-scale music assessment are developed, music educators are urged to form collaborative relationships with researchers conducting assessments in other curricular areas and to learn from their efforts.

References

"AMC news: Special report of the American Music Conference," *Teaching Music 3*(2), (October, 1995), 47-50.

Arter, J. A., and Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice 11*(1), 36-44.

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education 4*(4), 275-288.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education 4*(4), 305-318.

Boyle, J. D., and Radocy, R. E. (1987). *Measurement and evaluation of musical experiences.* New York: Schirmer Books.

Cooley, W. W. (1991). State-wide student assessment. *Educational Measurement: Issues and Practice 10*(4), 3-6, 15.

NAEP Arts Education Assessment Framework, pre-publication edition (1994). Council of Chief State School Officers. National Assessment Governing Board, Washington, D.C.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education 4*(4), 289-303.

Fisher, T. H., and Smith, J. (1991). Adventures in implementing a testing program. *Educational Measurement: Issues and Practice 10*(1), 24-26.

Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education 8*(2), 121-134.

Kearney, C. P. (1983). Uses and abuses of assessment and evaluation data by policy makers. *Educational Measurement: Issues and Practice 2*(3), 9-12, 17.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice 13*(3), 5-16.

Lehman, P. R. (1992). Curriculum and program evaluation. In (Colwell, ed.) *Handbook of Research on Music Teaching and Learning.* Reston, VA: Music Educators National Conference, pp. 281-294.

Maeroff, G. I. (1991). Assessing alternative assessment. *Phi Delta Kappan 73*, 272-281.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice 11*(1), 3-9, 20.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education 5*(3), 265-283.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice 12*(2), 9-15.

Moore, W. P. (1994). The devaluation of standardized testing: One district's response to a mandated assessment. *Applied Measurement in Education 7*(4), 343-367.

National Assessment of Educational Progress. *The first music assessment: An overview.* (1974). Denver: Educational Commission of the States.

National Assessment of Educational Progress. *Music 1971-79: Results from the second national music assessment.* (1981). Denver: Educational Commission of the States.

The National Standards for Arts Education. What every young American should know and be able to do in the arts. (1995). Reston, VA: Music Educators National Conference.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading. *Phi Delta Kappan 76*(7), 561-565.

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education 4*(4), 319-331.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation,*. Third Edition. Belmont, CA: Wadsworth Publishing Co.

Smith, M.L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice 10*(4), 7-11.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice 6*(3), 33-42.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment 1*(2), 131-152.