DOCUMENT RESUME

ED 401 294 TM 025 828

AUTHOR Roeber, Edward D.
TITLE Guidelines for the Management of Performance Assessments in Large-Scale Assessment Programs.
INSTITUTION Council of Chief State School Officers, Washington, D.C.; North Central Regional Educational Lab., Oak Brook, IL.
PUB DATE [96]
NOTE 23p.
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; *Performance Based Assessment; Sampling; *Scoring; *State Programs; *Test Construction; Testing Problems; *Testing Programs; Test Interpretation; Training
IDENTIFIERS *Large Scale Assessment; National Assessment of Educational Progress

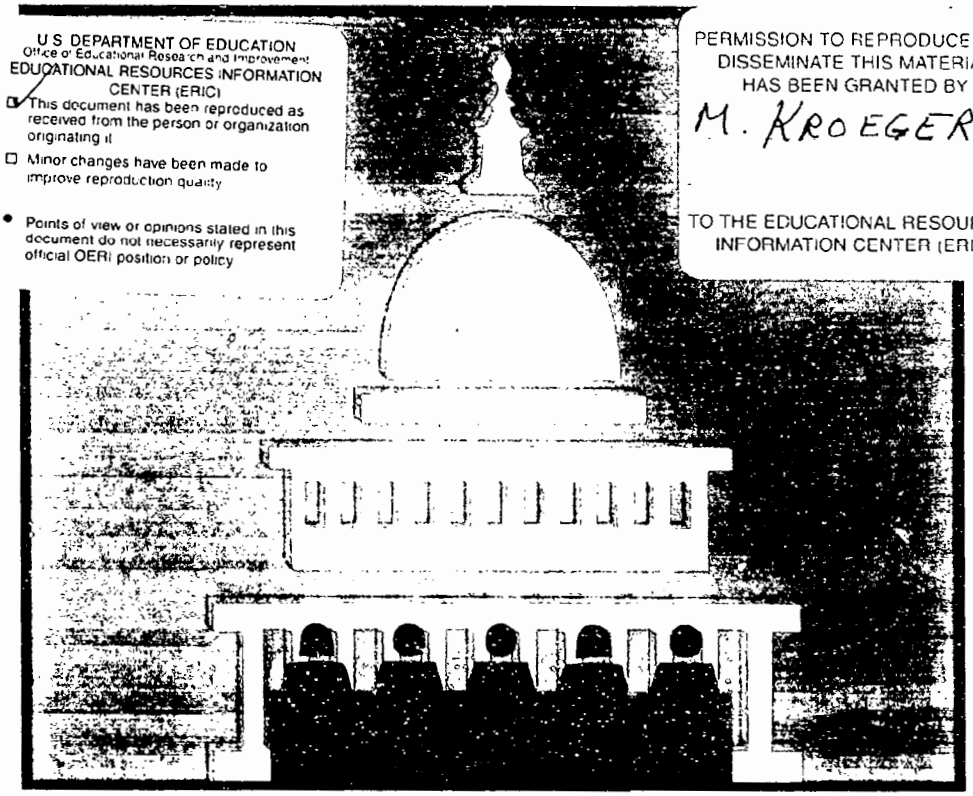
ABSTRACT

This paper is based on guidelines developed in 1989 for training workshops for state and local educators to demonstrate the processes by which performance assessments could be created, validated, and used in statewide assessment programs. These guidelines are based on work with the National Assessment of Educational Progress and several statewide assessments. Before assessment can begin, preassessment activities are necessary to establish the assessment framework, develop its plan, and determine assessment resources. Steps in the development of the assessment begin with the development of assessment prompts, and follow through editing of the developed exercises, development of administration procedures, exercise tryouts, development of scoring, resolution of statistical and technical issues, and refining the assessment after tryouts. Preparation for the assessment administration includes selecting samples and preparing schools and administrators for the assessment. Assessment administration includes notification of schools, monitoring the assessment, and evaluating the assessment process. Postadministration activities are important, from training scorers to summarizing, reporting, and interpreting results. These guidelines illustrate that performance assessment is feasible and manageable with proper preparation and evaluation. (SLD)

TM

Guidelines for the Management of Performance Assessments in Large-Scale Assessment Programs

REGIONAL POLICY INFORMATION CENTER



Edward D. Roeber
Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, D.C. 20001

NCREL

Guidelines for the Management of Performance Assessments in Large-Scale Assessment Programs

Edward D. Roeber
Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, D.C. 20001

Contents

Preface

This paper is based on guidelines developed in 1989 for use in training workshops for state and local educators to demonstrate the processes by which performance assessments could be created, validated, and used in statewide assessment programs. At the time, virtually all large-scale assessment programs consisted of paper-and-pencil, group administered assessments. The goal of the guidelines was to demonstrate the feasibility of administering performance assessment so that a broader range of student expectations could be assessed and reported on. This remains the goal of this paper.

A rhetorical war wages today over the advantages and disadvantages of performance assessment for large-scale assessment programs. Proponents often seem to advocate the use of performance assessment as the means of single-handedly improving school or student performance. In the extreme, these individuals condemn the use of multiple-choice assessments and indicate only those programs that are "pure" performance assessment-based are worthy. Critics, on the other hand, attack the impracticality and poor technical qualities of such assessments, often refusing to consider whether multiple-choice and even open-ended assessments would benefit from the addition of other assessment options and ideas.

This paper is written from the perspective that it is time for the rhetorical wars to end. The advantages and disadvantages of performance assessment can and should be viewed more objectively. Performance assessment is an important and unique tool available for measuring student performance at the state or local level and, as such, it should be used more frequently in large-scale assessment programs. This perspective is based on prior work carried out by the author in both the National Assessment of Educational Progress (NAEP) and in statewide assessment programs. The NAEP assessments, conducted in samples of schools and homes over two decades ago at a reasonable per-pupil cost, are perhaps the best demonstration that a mixed assessment program, combining multiple-choice, open-ended, and performance assessments, is both practical and feasible. The state-level work also demonstrates that it is possible to conduct valuable statewide sampling performance assessment for little cost.

The author is grateful for the peer reviews on the original and revised versions of this paper. The reviewers challenged the incomplete thoughts and smug assumptions written into previous versions. The paper has benefited from their critical review. The opinions expressed in this paper are those of the author, however, and are not necessarily those of the organization for which he works nor agencies that use or reprint this article.

Introduction

These guidelines are provided to offer guidance to district and state policymakers and assessment directors concerning some of the issues of managing the development, administration, and use of performance assessments in large-scale assessment programs. Just as it is not possible to devise the best way to develop or administer an assessment in general, it is not possible to do so for this more specialized type of assessment. However, many of the issues that builders and users of such instruments should consider are known. It is important to think carefully about these issues as decisions are made relative to the scope and types of assessments to be used. Alternatives depend on the level of resources that are available, the areas in which assessment will occur, the staffing present, and so forth. The purpose of these guidelines, then, is to help the user of performance assessments consider in advance some of the issues to be faced and help plan the manner in which such assessments will occur.

Before embarking upon a description of the procedures for developing and using performance assessments, it is important to consider what the term *performance assessment* does and does not mean as used in this paper. It is not uncommon to hear the term used to describe almost any non-multiple-choice assessment: short-answer questions, essays, hands-on assessments with manipulatives, group activities in which students are observed and rated individually or as a group, and even responses for which students are given extended periods of time to respond. While the most frequently used non-multiple-choice assessment is the essay assessment in the area of writing (and such assessments do constitute a writing *performance* assessment), the term performance assessment as used in this paper is reserved primarily for those assessments that go beyond paper-and-pencil, group-administered assessments, whether multiple-choice or open-ended. While the multiple-choice and the open-ended exercise format are valuable tools for the large-scale assessment programs, the purpose of this paper is to describe assessment methods that go beyond these means of assessing student achievement.

Pre-Assessment Development Activities

Before assessment development can or should occur, several important planning activities set the stage for the assessment. These steps take place at the outset so that the assessment is developed in a manner that fits the content area to be assessed and is within the resources available. It is presumed that the agency sponsoring the assessment already will have determined:

- The assessment framework. It is assumed that the assessment framework calls for assessment strategies other than multiple-choice and open-ended. If the framework does not call for the use of performance assessments, there may be little reason to develop them.
- The grade levels at which the assessments will take place. Performance assessments at the elementary level may pose tasks or use response formats that are beyond students. At the older grade levels, student apathy may pose challenges in using such assessments.
- The purposes for which the overall assessment program, as well as the performance assessments to be developed, will serve. Are these high- or low-stake uses, both for students and for the professionals that work in the system? High stakes uses of the assessment will require more rigor in the development, validation, the administration, and the scoring of the assessment. In some high stakes situations, such as high school graduation tests, it may not be feasible to use performance assessments. In low stakes assessments, it may be desirable to assess only samples of students statewide, or in a sample of schools. The stakes accompanying the assessment will help determine the design of the assessment.
- Whether the results will be reported and if so, how results will be reported. Will the results be combined across the performance assessments? Will the performance assessment results be combined with other types of assessment results (multiple-choice or open-ended assessments) and reported in an overall fashion? If so, there are significant technical issues to be considered.
- To whom results will be provided and for what purposes. Will the student,teacher, school building, or school district, receive results or will just statewide results be reported? Will group results be reported to the public? Will individual results be reported to parents? What statements about the results are to be made based on these assessments? The answers to this set of questions and the previous set will help determine the number of performance assessments to be used, how they might be scaled, and the level of detail needed for reporting.
- Whether the results will be compared from year to year. The reporting of results longitudinally requires the equating of different forms of the assessment so that comparisons can be made across years, an added technical requirement.

Once these issues have been determined, the program is ready to embark on the development of the performance measures. This paper presents the development ideas rather informally. It is important to note, however, that the more "high stakes" the program, the more it will be necessary to make certain that technical and policy issues have been formally addressed.

The following sections present information about how performance assessments can be developed, administered, scored, and reported. While the paper suggests informal, less costly means to develop and use these types of measures, keep in mind that some of the uses outlined above might require the use of external contractors or technical advisors.

Development of the Assessment Framework

The framework for the assessment, whether constructed specifically for the assessment or put together for more general reasons (i.e., as a document to suggest to schools the standards that students should achieve at various points of time), serves as the guide to the entire assessment. It is from the framework that the assessment is developed. Much has been written about the process of writing performance objectives and it is not the intent of this document to duplicate that work. However, such documents can be written in several ways, and there are a couple of important points about such lists of objectives.

The traditional manner in which sets of objectives are written is to gather a group of content area experts and classroom teachers and ask them to indicate what students should know and be able to do by the end of particular points of instruction (e.g., by the end of third, sixth, and ninth grades). These descriptions can vary in number and may be briefly or extensively written up. However, the intent is to provide at least the most important outcomes and to describe these in more or less behavioral or observable terms. An alternate procedure is to allow the framework of expectations for students to evolve out of research and practical experience of educators on what outcomes students are capable of at particular times in their school career. In essence, classroom teachers and curriculum specialists explore the outcomes that students appear to be capable of and use this as the basis for the descriptions of the outcomes that students should be able to accomplish by that point. Either approach will work, with the former method probably having the advantage of both speed and lower cost, while the latter has the advantage of being more behaviorally anchored.

Regardless of the manner in which the assessment framework is developed, one important thing to wrestle with is the extent and types of performance assessments that will be used. It is possible to devise assessments that vary along the lines of the type of stimulus material to be used (audio, video or other specialized types), the types of responses (written, oral or other performance), group or individual performance, and so forth. This needs to be determined at the outset so that as the outcomes are being described, they are captured using action verbs that will be consistent with the manner in which assessment likely will occur. When a performance assessment is contemplated, it is perfectly acceptable to use action verbs that might suggest more performance assessment than is feasible (e.g., "student will describe the relationship between the post-WW I environment in Europe and the causes of WW II"). However, it is not desirable to describe the outcomes that would suggest a more conventional assessment (using verbs such as "select the correct answer") . Ideally, the assessment framework can be devised without consideration of the modes of assessment to be used.

Development of the Assessment Plan

Once the assessment framework has been developed, creating the actual assessment plan should be straightforward. The assessment plan provides an overview and description of the types of assessments to be developed and used, as well as the manner in which assessments will be carried out. The plan serves the useful purpose of describing the types of assessments that are envisioned and how such assessments will be administered, scored, and reported. It is important to have this plan at the outset so that the actual assessment materials to be developed and used also are known at the outset. In essence, the assessment plan serves as the introduction to the assessment blueprint, which is described in a following section.

The assessment plan should break down the assessment framework into the various types of assessments to be conducted and the types and numbers of assessment packages to be used. Thus, these decisions will be based on the purposes for the assessment, the manner in which results are to be reported, and the uses to be made for the assessment. Since it is feasible to package some of the assessments together, it is important not only to estimate the number of exercises but

also the number of assessment packages and the assessment time per package. Doing this planning at the outset of the project will make obvious to all involved just how extensive the non-traditional component of the assessment will be and should go a long way to prevent the development of elaborate assessment methods that cannot be used because of the lack of assessment administration or scoring resources.

Determination of Assessment Resources

The assessment plan should force the agency sponsoring the assessment development to carefully consider the resources needed and available in order to make decisions about the assessment at the outset. This is a critical step, often overlooked until after the assessment development is completed and the agency is considering just how feasible it is to administer what has been developed. Since the postponement of such decisions about resources is likely to lead to the development of assessments that cannot be used, it is far better to consider the resources that will be needed and that are available at the outset of the project. This will allow three things to occur. First, the sponsoring agency can seek additional resources in order to develop and administer the assessment desired. Second, the sponsoring agency can change the manner in which the assessment is developed or administered (e.g., can develop or administer the assessment using volunteers rather than a paid staff or a contractor). Third, it can also change the scope of the assessment to fit within the available resources.

Several types of resources should be considered at the beginning of the project. Who will construct the assessments, and who will develop the assessment administration and scoring procedures? Will a contractor be used, or will the assessment development be carried out by volunteers? Who will do the editing and the preparation of materials for tryouts? What resources are available to support the development of the assessments? Additional questions can be asked about how the assessments will be administered. Will classroom teachers administer the more conventional assessments? Will the performance assessments be administered by in-school personnel, or will outsiders need to be recruited to go to the schools to administer the assessments? Will these outsiders be individually recruited (and will they be volunteer or paid staff), or is a contractor going to be used? If volunteers are to be used, how will quality control be maintained where it might not be feasible or permissible to dismiss a volunteer who cannot competently carry out assessment administration or scoring? How will the assessment administrators be trained? What expenses will be paid? Will student responses be scored as students perform, or will they be recorded in some manner and scored later? Who will conduct the scoring and who will participate in it? Will these be volunteers or paid scorers? Will the sponsoring agency conduct the scoring sessions, or will a contractor be used for these activities.

Once the assessment plan has been written and the assessment resources have been determined, it is possible to make any adjustments needed in the assessment plan. At this point it is safe to assume that the plan that has been devised is a feasible one, that the assessments to be developed are known, as is the manner in which the assessments will be administered and scored. Hence, it is now possible to move safely into the next phase of the assessment development, writing the assessment blueprint. However, it should be kept in mind that the assessment plan may change during the subsequent steps in the development process, either because of unforeseen aspects to the assessment of particular outcomes or because of changes in the level of available resources. If this occurs, the agency responsible for the assessment should return to the assessment plan and revisit the answers to the above questions in order to assure that the needed assessment will be developed.

Development of the Assessment Blueprint

Once the assessment plan has been written and accepted, it is time to develop the assessment blueprint. The blueprint differs from the assessment plan in both purpose and level of detail. The blueprint will describe the characteristics of an adequate assessment for each content area of the assessment framework. Rather than focusing on the assessment administration mode or the manner in which students respond (as the assessment plan does), the blueprint describes the characteristics of the assessment for each area of the framework.

The blueprint, which should begin with a description of the answers to the questions posed at the beginning of this section, would have to describe the characteristics of an adequate assessment for each student outcome. For example, the assessment framework might indicate that students are to devise an instrumental accompaniment to a song which they hear sung to them. The blueprint for a music assessment would need to include the range of songs that a student should be familiar with at particular grade levels, the manner in which the song(s) will be presented to them (live or on a tape recorder), the manner in which students will be asked to respond, the materials to be used (e.g., stick and wood block), the manner in which student responses will be recorded and scored (scored live or tape recorded for later scoring), the criteria by which students' responses will be scored, as well as how the scores will be reported.

The blueprint will contain this information for all areas of the assessment framework. Once completed, this should guide the development of the assessments that are needed given the resources available. The final step in the pre-assessment development activities is for the assessment plan and the assessment blueprint to be approved by the sponsoring agency and any additional advisory groups or individuals. While gaining approval may appear obvious, this should not be considered a trivial step. Such approval should be viewed as a promise of resources and the sponsoring agency's commitment to see that the needed assessments are developed and actually used.

Assessment Development Steps

Once the assessment framework, plan, and blueprint have been developed and approved by the sponsoring agency and others, it is now time to move into the actual development of the needed assessment materials. As indicated earlier, it may be necessary to make changes in either the assessment plan or blueprint at this point. Such changes should be made as the assessment development process unfolds.

Development of Assessment Prompts

Given a well-defined assessment blueprint, it should be relatively easy to develop the actual assessment prompts. For multiple-choice exercises, this consists of writing the stem and responses. With open-end exercises, this consists of writing the prompt and determining different categories of correct and incorrect responses. With performance exercises, it is even more complex. Here the developer must consider how the prompt will be presented to the student, what additional stimulus materials will be needed (and where such materials can be located or developed), how students will respond and how such responses will be recorded and scored, what criteria will be used to judge student responses, the number of scale points for scoring student responses (a four- or six-point scoring scale is typical), and samples of each level of response. Obviously, the task of devising adequate performance exercises is considerably more complex than writing adequate multiple-choice exercises.

One useful technique for developing such assessments involves a combination of individual and group effort. Since there are so many aspects to the development of just one assessment prompt, it may be difficult for one individual to devise the entire exercise. However, it is also difficult for committees of exercise writers to work creatively. One strategy, therefore, is to have individuals do the initial development work on assessment prompts alone and then present their work to an exercise writing committee. The committee can then question the individual, make editorial suggestions in the prompt, suggest additional or alternative stimulus materials, provide additional guidance on the recording and scoring of students' responses, as well as possible types of students' responses in order to refine the scoring guides. The original exercise writer or another individual can continue to refine the assessment prompt. This process can be repeated if needed or desired.

Another useful technique is to try out the performance exercises with a few students during the initial developmental phase. Although this may be difficult with exercises that require elaborate stimulus materials (such as specially recorded music), it is feasible with many exercises. In addition, it is most valuable if the actual exercise writer is the one who administers the exercise in this informal tryout, so that the exercise developer can see first-hand how students respond. Gathering a few student responses can assist the exercise developer in writing appropriate assessment administration directions, wording the exercise in an understandable manner, and devising suggested scoring criteria and guides.

Editing of Assessment Exercises

Once the exercises have been written, the next step is to edit them. The editorial work for performance exercises is complex because there are assessment administration procedures to be followed, material to be read to the student, directions for recording student responses, scoring criteria and so forth. This is an important step in assuring that the exercise and assessment administration process are understandable, that the exercise warrants the special time and attention that performance measures require, and that there is consistency among the performance exercises.

For performance exercises, the editor will assure not only that each exercise is clearly worded, but that the "package" of exercises fit together and flow from one exercise to another. The editor can assure that the apparatus needed for one exercise will fit with the materials needed for the next one. The editor also can check to make certain that the stimulus materials fit with the exercises. Finally,

the editor should assure that each performance exercise has a preliminary scoring guide that provides examples of the different levels of correct and incorrect responses, a rationale for what constitutes an appropriate response, and suggestions for reporting the responses to the exercise. Ideally, this will have been developed by the exercise writer during the exercise writing stage and may reflect informal tryouts done by the exercise writer. In almost no case should exercises proceed to tryouts without such a scoring guide, since trying out such exercises may be a waste of valuable resources, particularly when the exercise writer and editor can think of no responses which may be judged as correct or not correct!

Developing Assessment Administration Procedures

One of the more difficult aspects of performance exercises is writing the assessment administration directions. Fortunately, not all performance exercises require such special directions. Open-end exercises that call for students to write an essay on a topic are an example of exercises that may not require special assessment administration procedures. However, many performance exercises are administered to students individually or in small groups. These exercises require the assessment administrator to set up a standard situation for each student to respond to, as well as to read a standard set of directions to each student. In addition, the directions may provide a standard set of probes in case the student's response is vague or incomplete. The emphasis is on the word "standard" because it is necessary to provide such standardized assessment administration procedures if the results from different students, classrooms, schools, or districts are going to be combined for reporting. Although this depends on the answers to the questions posed in the Pre-Assessment Development Activities section, it is something that the assessment designer should consider. To the extent that results across students are to be combined and these results publicly reported, the procedures for collecting the information will need to be standardized across the data collection sites.

In writing the assessment administration directions, there are several things to keep in mind. First, overall directions to the assessment administrator, including such things as drawing the sample of students within the school, locating a suitable room for testing, and calling the students to be assessed, need to be written. Second, the step-by-step process of administering the exercise needs to be written. This may begin with setting up any apparatus or materials needed in the assessment, correctly positioning the student(s) for the exercise, the actual material to be read to students (shown in a manner different from the directions to the assessment administrator), the probes to be used should students give no response or one that is not complete, how to move from one exercise to the next, and so forth. The trick is to "think through" the administration process one step at a time, continually asking yourself "what if?" types of questions. Once the directions are written, it may be helpful to administer them to someone else or have someone else administer the exercises to you.

Developing standard assessment administration directions that are complete and accurate is usually the result of trying the exercise out one or more times and noting areas of student confusion, responses that students provided which are vague or incomplete (and devising appropriate probes in order to elicit additional information from the student), and ways in which some or all students responded that were not anticipated. While some of this can be caught at the editing stage, tryouts of performance exercises are critical.

Trying Out the Assessment Exercises

As mentioned above, trying out performance exercises is critical to the development of sound performance exercises. Unlike group-administered, paper-and-pencil exercises, where having sufficient numbers of students respond to each exercise in tryouts is critical, the most important part of trying out the performance exercise is to make sure that it is given to students by one or more well-trained individuals. In essence, it is important to train the assessment administrator as well as possible, certainly as well

as will be done in the actual assessment. This means that the tryout assessment administrator should be the same type of individual as will be used in the actual assessment, since a major purpose of the tryouts is to determine whether the exercise can be properly administered. It also means that draft assessment administration directions should be written and used for training the tryout assessment administrators, so that the effectiveness of the written procedures for training can be tried out and refined.

In addition, it is important to determine just how assessment administrator specific the assessment administration directions are. This means that it may be more important to have 10 to 15 assessment administrators administer the performance exercise to five students each than to have five assessment administrators administer the exercise 15 times each. While this will require extra work, both in terms of recruiting tryout assessment administrators and assessment administration, the payoff will be receiving a more accurate picture of the feasibility of the assessment administration process.

If it is feasible and the resources permit, it is desirable to assess as many students as possible. If a sample of schools and students can be drawn and multiple assessment administrators used as suggested above, having a sample of as few as 300 students at the tryout stage may be sufficient. However, larger sample sizes are always desirable, particularly to assure that the range and types of responses to the exercise will represent the range and type of responses to be collected in the actual assessment, as well as to examine the differences in performance among groups of students. The more complex and varied the types of responses anticipated, the larger and more varied the tryout sample should be.

Developing Scoring Guides

Responses to performance exercises may range from psychomotor responses that are observed and rated live (or on videotape), recorded on audiotape, observed, recorded by the assessment administrator, or written by the student being assessed. The first essential element to developing sound scoring guides is having obtained an adequate number and range of responses in the tryout phase of the project. As mentioned above, the number and types of responses anticipated will help determine the sample size required to achieve this goal.

Once the student responses are gathered, someone needs to review the responses and attempt to score them according to the criteria and preliminary scoring guide developed by the exercise writer. This may be the exercise editor who does this step, or it may be another individual associated with the project. However, this step is generally best handled by one or two individuals. They should try to verify the preliminary scoring guide by finding samples of each type of anticipated response, as well as samples of student responses that do not fit the preliminary scoring guide.

Once the initial work on the scoring guide is complete, an expert panel of judges should be convened. This expert panel may be composed of classroom teachers in the area, curriculum specialists, and subject-matter experts from the university level. A substantial amount of training may be necessary for the scoring panel to learn to use the scoring guides developed for each exercise reliably. This training should be documented so that it can be replicated when the subject area is reassessed in the future.

This panel should be asked to review each exercise, confirm the preliminary judgments for each student response selected to illustrate each score scale point, and to discuss those student responses that did not appear to be scorable according to the preliminary scoring guide. The panel may note changes that need to be made in the assessment administration process, such as the addition of probes for particular student responses or places where some of the students being assessed appeared confused. Once completed, the expert panel review should confirm the scoring guide or serve as the basis for changes in it. The result will be having an exercise that is feasible to administer and which can be scored appropriately.

Statistical and Technical Issues

Performance assessments (whether open-ended exercises or ones requiring more intricate, individual supervision) require time to administer. Hence, in a fixed period of time fewer performance assessments than conventional, multiple-choice exercises can be administered. In addition to concerns about time, there are several technical issues with which the developers should be concerned. These include generalizability, bias, equating, and scaling. Each of these issues should be considered during and after the tryouts.

Generalizability concerns whether the set of exercises chosen represents the total domain under investigation. Since relatively few performance assessments are typically given, the tryouts should investigate how many exercises of the type(s) used are needed to obtain stable estimates of student achievement.

Several types of bias must be considered. The first is whether there are differences in performance by various sub-groups of students. During tryouts, it will be important to examine the differential item functioning (DIF) of the exercises by gender and racial-ethnic group. It also may be important to examine the performance of the exercises by region and community type. Both statistical DIF analysis and expert judgment should be used to review the exercises. As the exercises are being tried out the issue of response bias should also be investigated. This refers to the differences among students with the *mode of response* used in the exercises. For example, does the written or performance response format disadvantage certain students who would be able to perform adequately on the exercise if presented in a different mode? This should also be investigated during the tryouts by interviewing students following the tryouts, as well as trying out the exercise with different response modes.

Scaling and equating are also issues that should be investigated following tryouts. Scaling is important because the exercises will undoubtedly be reported together as an overall level of performance, either the performance assessments alone or in conjunction with other assessments used. In addition, due to the unique, "memorable" nature of these exercises, it will be important to use new exercises during each actual administration. Each of these will require that the set of performance assessments be scaled together, at least among the performance assessments (and with the other exercises to be used). Placing the exercises on the same scale will make it possible to select subsets of the exercises for use each year, as well as to equate the different forms of the assessment.

Refining the Exercises After Tryouts

Once tryouts are complete, the student samples have been scored, and the scoring panel has made any suggestions for improvement of the exercise, the exercise editor can make any final changes. Presumably, at this stage the changes proposed are minor in nature. If they are not, as is sometimes the case, then the process given above really should be repeated. While there may be a temptation to assume that any major changes in the exercises have been made and will correct any deficiencies in the exercises, this is really not known until new student data is collected. After all, if it didn't work the first time, why are you certain that it will work this time?

The assumption is that it is far cheaper to try out the exercise one more time than to put a faulty exercise in the large-scale assessment program and not have the exercise work. At best, this will be a waste of money and effort; at worst, the exercise may work so poorly that it may disrupt students' responses to other exercises. Hence, major changes in the assessment administration directions, student tasks, stimulus materials, or scoring guides suggest additional tryouts are needed before the exercise is used in the large-scale program.

Preparation for Assessment Administration

Once the performance exercises have been developed and tried out, the next important set of activities is getting ready for the actual assessment. This involves selecting the schools that will participate, preparing the schools for participation in the assessment, and training the individuals who will gather the data from students. This is a stage that is often not given enough attention in planning and conducting performance assessments, yet is a vital one for the gathering of useful information about students.

Drawing the Samples of Schools and Students

Since the assessment materials that have been developed and refined are intended for large-scale assessment use, it is presumed that the exercises will be given to some or all of the students at one or more grade or age levels. The original assessment plan will help to determine how and to whom the exercises will be administered. If each of the exercises is to be administered to all students at a grade or age level, no sampling is needed and the assessment designed can move on to the next stage. However, if all students at a particular grade or age level will not be assessed, or if they will not all take the same assessment, then some type of sampling procedure will be needed.

Sampling is an activity that most large-scale assessment programs have a contractor skilled in sampling design perform for them. It is a process that is relatively easy to carry out at the school level, particularly if the state or district already has some process of stratifying the state or district. If this is the case, then it should be relatively easy to draw a sample of schools. Given below is one example of how sampling at the school and student levels can be carried out. The actual procedures used, and the complexity of these, is highly dependent on the design and purposes for the assessment. The example shown is for a low-stakes assessment in which only samples of schools and students will participate and results will be reported only at the statewide level.

It is better to include as many schools as possible in the sample by keeping the number of students to be assessed in any one school relatively low. Using a figure of 20 students per school and an overall minimum figure of 500 students per assessment package requires a minimum of 25 schools per package. One way to economize on travel expenses is to administer more than one assessment package in a school, with 20 different students in each school taking each assessment package. Once the total number of students per school is decided, the number of schools to be selected from each sample stratum can be determined.

Spaced sampling is one easy procedure for selecting the actual schools and students to participate. Starting with a complete list of eligible schools, the assessment administrator first determines the total number of eligible schools. Then the administrator divides this by the number of schools to be assessed. This results in a constant, which can be labeled **c**. The assessment administrator then picks a random number between **1** and **c**; this is the first school included in the assessment. Next the assessment administrator counts down the list of schools, selecting every **cth** school; these are the remaining schools that will be assessed on the assessment package. Since there are cases where selected schools may not have any students of the eligible grade or age level, or the school may refuse to participate, it will be helpful if a set of alternate/replacement schools is selected for each stratum at the time that the original sample of schools is drawn.

At the student level, drawing a spaced sample of students is equally easy and can be handled by a trained assessment administrator. To reemphasize how a spaced sample works—starting with a complete list of eligible students, the assessment administrator first determines the total number of eligible students. The administrator then divides this by the number of students to be assessed on a assessment package in the school. This results in a constant, which can be labeled **c**. Then the

assessment administrator picks a random number between 1 and c; this is the first student included in the assessment for that assessment package. The assessment administrator then counts down the list of students, selecting every **cth** student; these are the remaining students who will be assessed on the assessment package. Since some of these students may be absent or unavailable for assessment, the assessment administrator should also select some additional, alternate students. Since this process is so easy, the assessment administrator can select the sample of students on the day of assessment, presuming that a complete list of students (in any order) is available on the day of assessment.

Preparing the Schools for Assessment Administration

The preparation of the school for the assessment begins with the notification that the school has been selected for participation in the assessment. Letters should be sent to the school coordinator and to the district coordinator if applicable. It will help the school if the notification letter includes such details as when the assessment will occur, what will be involved in the assessment, what the school's responsibilities will be, who will administer the assessment, and so forth. The more specific this notification letter can be, the fewer questions will be raised and the easier it will be to receive school cooperation.

Schools should be told the approximate time when the assessment will occur, who will be contacting them to make additional arrangements, what process will be used to draw the sample of students (and why they need to have a complete listing of all eligible students on the day of the assessment), and what facilities and equipment they will need to provide (e.g., a quiet room, two chairs, a table, and a VCR and television monitor).

Training the Assessment Administrators

A major key to the success of the entire performance assessment project is the quality of the individuals who are selected to conduct the assessment. If the resources are available, an organization that provides such services can be contracted to handle most of the details of selection of the assessment administrators and their training to administer each assessment package. Even in these cases, however, reviewing the following guidelines may be helpful to assure that all important points have been considered. If the resources for an assessment administration contract are not available, the sponsoring agency will need to carry out the following activities directly.

Locating suitable assessment administrators may be the most challenging aspect of carrying out a performance assessment without having a contractor to administer it. One technique that has worked is to hire persons who would normally substitute teach in the subject area(s) to be assessed. These individuals may be interested in the assessment activity and be willing to be trained to administer the assessment. Another technique that works is to use the colleges and universities in the state as a network from which to draw. In this case, each university is asked to name a team made up of one or more faculty members, plus one or more graduate students. This works particularly well in the cases where the state has a regional university system. For example, in one assessment situation, university curriculum teams not only carried out the assessment, they also were available to other local districts not in the state sample that wished either to be assessed or to be trained in administering the tests. In addition, since the administration of a performance assessment can serve as an excellent professional development activity, some local districts will support their staff in learning to administer the performance tests. In other contexts, this might not be possible if potential assessment administrators demand substantial honoraria or union contracts do not permit volunteer work.

Once the assessment administrators have been identified, training should occur. Depending on the nature of the assessment, as well as whether student responses are simply going to be recorded in some fashion for later scoring or will be scored on the spot during the assessment, the training may be rather extensive or relatively brief. Such training sessions have been as short as a day and

as long as four days. In any case, the same material is reviewed.

First, the assessment administrators are briefed on the project and the steps that have taken place prior to the training session. Second, they are informed how the state sample was drawn and how they are to draw the sample of students (see above). Next they are given their assignment of schools and told who and how to contact the district assessment coordinator and the building assessment coordinator. Fourth, each assessment exercise that they are to administer at each grade level is reviewed. This is done first by literally reading a copy of the assessment administration guide to them. Then each exercise is demonstrated to them, and finally, they have a chance to practice administering each exercise to one another. They are also reminded to re-review and practice the assessment administration prior to administering the assessment the first time. This practice is particularly important when the assessment administrator must coordinate the use of manipulatives or equipment during the assessment. If possible, plenty of time is allowed for this practice assessment administration. Where time permits, arrangements are made to have either adults or children available in order to give the assessment administrators simulated practice in assessment administration.

Once the assessment administrators are comfortable with the administration of the tests and the manipulatives and equipment that are used in them, the focus of the training shifts to the recording and scoring of student responses. In some cases, this training may be relatively simple. For example, in an assessment of music, if student responses are tape-recorded this step probably can be omitted, since the assessment administrators have already practiced using tape recorders as part of learning to administer the exercises. However, in the case where the assessment administrator must score student responses as they occur, training may take one or more days.

In the case of physical fitness, for example, the assessment administrator needs to learn to score about 20 different exercises. Two techniques can be used. First, various student responses are videotaped, both at regular and slow motion speed. This allows the development of both a training tape and a validation tape for each exercise. The training tape shows regular and slow motion performances for each score scale point. Second, accompanying the training videotape is a training manual that uses stick figures to illustrate each criterion for an acceptable or unacceptable performance. For example, there are several criteria for throwing a ball, including cocking the arm, the elbow leading the hand, breaking the wrist, and the follow through. Each of these is illustrated in the training manual, and the assessment administrator can see examples of both good and poor performances on the videotape. The written criteria are reviewed extensively with the assessment administrator. Then they practice scoring other samples of student performance from the training videotape for each exercise.

Once the assessment administrators have been trained on each exercise separately and are able to properly score each of the samples on the training videotape, they are given the validation videotape to score. This tape presents several student responses in the order in which they occur in an assessment package. Since an assessment package could contain several different exercises in a row and the assessment administrator would need to remember the different scoring criteria for each, the validation videotape presents the samples for each student together, just as they would appear to the assessment administrator as they administer the exercises live. Trainees with acceptable scoring prowess are certified as assessment administrators; others must be either cycled back through more training or dismissed from the assessment project.

Assessment Administration

It may seem that a lot of work has already gone into the development and training for the performance assessment. Yet, the next phase is also a crucial one in assuring the feasibility and accuracy of performance assessment in the schools. The manner in which schools are contacted, students are selected and assessed, and the assessment process is reviewed for improvement in the future can determine the success or failure of the entire process. Therefore, this is not a time to relax or not pay attention to what the assessment administrators or schools are doing.

Notification of Schools About the Assessment

As mentioned earlier, the notification to the schools that were selected to take part in the assessment usually takes the form of a letter to school coordinator and to the district coordinator, if applicable. Once the assessment administrators are trained, they also should contact the appropriate assessment coordinators. Thus, they might remind the district assessment coordinator that one or more schools were selected to take part in the performance assessment, indicate just what the assessment will consist of and determine who the school assessment coordinator is. Or, if within a district, they would contact the school coordinators. If some time has elapsed from the time that the initial letters were sent, it may be helpful to send a reminder note shortly before the assessment administrator is to contact the district.

If applicable, it is critical that the district contact occurs before the building contact. Then the school assessment coordinator is contacted. Hopefully, this individual is aware that the performance assessment will be occurring in the building. The purpose of this contact is to schedule the date(s) and times for the assessment administration to occur, to remind the school coordinator to have a complete listing of students at the appropriate grade level(s) available on assessment day and specify what facilities and equipment will be needed. These contacts should be made before any assessment administration starts.

Any problems that are uncovered by the field assessment administrators should be relayed to a designated contact person. These problems might require the selection of a replacement school or the designated person to contact the school in order to secure cooperation and so forth.

Monitoring the Assessment

Once all contacts have been made, the field assessment administrators can begin the process of assessment administration. This will start with the drawing of the sample of students to be assessed and a list of alternates. About an hour should be allowed for this to take place. Once the student list has been compiled, it may be helpful to have the school assign an aide or a student who can work with the assessment administrator to locate the students when needed for the assessment and bring them to the assessment administration site. A person who is familiar with the teachers, the class schedule, and the physical layout of the school will be the most helpful.

This aide can bring students to the assessment administration site as needed and can help assure that each student is returned to class without undue disruption. The next student can be brought to the room just before the previous student is finished so that the assessment administration can flow quickly and efficiently from one student to another.

As the assessment administration process is taking place in the schools, it will be most helpful if the designated contact person and others associated with the project but not involved directly in the assessment administration would select some schools in which to observe one or more students taking part in the performance assessment. Not only will the observers have an opportunity to observe assessment administration taking place, they also can discuss the assessment with some of the students following the assessment. This can provide invaluable insight on why and how students responded and their motivation and interest in the assessment, as well as ways in which the

assessment administration and assessment administrator training processes can be improved in the future.

Evaluating the Assessment Process

The monitoring process outlined above will be helpful in the evaluation of the performance assessment administration process. It also may be helpful to solicit comments directly from the assessment administrators, school coordinators, and district coordinators. This evaluation process is important in order to determine whether there were any factors that affected student performance and, therefore, ought to be considered when interpreting the data that results from the assessment. It also can be helpful for the future when other assessments are being designed so that problems encountered in this assessment are not repeated (or, at least, are better anticipated) in the next assessment.

Post-Assessment Administration Activities

Once the assessments of students on the performance exercises is completed, there still may be considerable work ahead in order to score, report, and interpret the assessment information. If the performance assessment was scored while being administered, some of these steps can be omitted. Otherwise, these steps will be essential for compiling useful information.

Training the Scorers of Open-End Exercises

The first step in the scoring phase is to locate people who can be trained to be scorers. In some cases, it is desirable to select and train classroom teachers in the scoring routines and processes. Such scorer training can be an excellent professional development activity. If this is done at a time when school is in session, then the number of days available from any scorer may be limited and it may be necessary to train more scorers. Also, it may be advantageous to conduct the scoring at a regional level rather than at a central location, so that the necessity for overnight accommodations is limited. This will allow a teacher, for example, to drive to the scoring center each day and return home in the evening. One way in which regional scoring centers can be "housed" is on college campuses.

The next step is the development of the training package. This may begin with the development of a training package comparable to what was developed to train the assessment administrators who would have scored student responses in the field. This all takes place before the scorers are convened for the first time. First, various student responses are selected by an expert in the area to represent the various types of responses that students may have given (in the case of "holistic" scoring) or that illustrate each type of response (in the case of "primary trait" scoring). This also may lead to refinements in the scoring guide and rubrics.

Such expert judgments are next confirmed by an expert panel of judges; some of the samples will appear in the scoring guide prescored and will be used to train the scorers. The others (that are also prescored) will be used to judge the accuracy (reliability) of the scorers following the initial training. Depending on the nature of the assessment packages and how student responses were collected, it may be desirable to have the scorers score just one exercise at a time, or to be able to score all of the exercises in an assessment package for students one at a time. This will help to dictate the structure of the training sessions and the scoring process.

Once the scorers have demonstrated their ability to reliably score the exercise(s), training is complete. Scorers are then ready to begin the scoring process.

Conducting the Scoring of Open-End Exercises

The scoring process requires a number of things to be prearranged for the scoring to flow smoothly. First, a determination would have to be made about whether there will be one or two scorers for each response. If more then one scorer will be used, what process will be used to resolve differences? How will the first scorer's scores be kept from the second scorer? Second, arrangements will need to be made to distribute booklets and other materials to be scored, plus rating sheets, to each of the scorers. It will undoubtedly be desirable to have runners available to distribute and collect materials. In addition, it is critical that people who are well organized be used to keep track of each booklet and other material to be scored to make certain that it has received all needed ratings.

Third, routine reliability checks should be built into the scoring process. Each scorer should be given a few of the prescored exercises without notice to score during the actual scoring session. If their ratings have "drifted" from those of the expert panel, it may be necessary for on-the-spot

review of the scoring guides. It is also important for each scorer, later on in the scoring, to rescore a few responses that they scored earlier to make certain that scoring subsequent student responses has not caused their scoring standards to change. Again, it may be necessary to conduct an on-the-spot review of the scoring criteria.

Once the scoring design is determined, then the scoring training and actual scoring can commence. This may be completed in one session, or subsequent sessions may be required. If more than one session is needed, subsequent review of the scoring standards may also be needed.

Summarizing the Assessment Results

Once the students' responses to each of the performance exercises have been scored, the various scores need to be summarized and prepared for reporting. The summaries will be most efficient if the individual(s) who will be doing the reporting and those people directing the scoring discuss the needed and desired data summarization process before the scoring is conducted. This will allow the most efficient use of the scorers and the scoring contractors in preparing the data in the format(s) needed to summarize the data in ways that will lead most efficiently to the types of reports to be issued. Careful thought about this in advance can save hours of re-coding or re-entry of data.

Reporting the Assessment Results

There are probably at least as many ways of reporting on the performance of students on performance exercises as there are number of exercises. However, one thing to keep in mind is that the typical reader of a report will probably be unfamiliar with the performance exercise, the subject area, the scoring criteria, and so forth. What the typical reader is likely to be interested in is whether or not many students were able to correctly perform the required task, and if not, what were the typical performances that students gave. Explanations for low or unexpected performance are also usually of interest. If the reports of results—at least for the public and for classroom teachers—focus on these points, such reports will be of interest to the widest audience.

There will be others who are interested in more detail about how students performed on the exercises, and, for these groups, reports of a more complete and/or technical nature will be most beneficial. Others may be interested in obtaining the actual samples of student responses in order to conduct secondary analyses of the data. With proper safeguards for student and teacher privacy, these data can be made available and subsequent reports that these researchers make also can be publicized.

Interpreting the Assessment Results

The reports alluded to above suggest that more than mere numbers will be provided. Most audiences want to know not only how well students performed, but also why students performed as they did, were the experts surprised in any way by the level or types of student performances, and perhaps most importantly, what do the experts think needs to be done to help students improve. These important questions should be covered in the interpretation of the performance assessment results.

It may also be helpful to integrate the interpretation of both the performance and non-performance exercises into a single interpretative report. This could allow, for example, knowledge questions from a multiple-choice assessment to be reported in the context of whether or not students were able to apply that knowledge in various applications assessed in the performance exercises. This may lead to clearer interpretations of the results without diminishing the importance of the performance measures. Just as in reporting, keep in mind that different depths of interpretation may be desired and needed by different audiences so that multiple interpretive reports may be best for meeting different information needs.

Summary

This guide to performance assessments has illustrated the steps in creating, trying out, administering, scoring, and reporting performance assessments for large-scale assessment use. Although such assessments are not easy to develop, administer, score and interpret, much is known about the steps that need to be carried out, and many different organizations at the national, state and local levels have already conducted successful assessments in a wide variety of subject areas. These assessments range from the simple to the complex and from ones that cost little (a few thousand dollars) to a considerable amount (hundreds of thousands to millions of dollars).

While performance assessment may be new to some people (news stories have characterized how it has been recently "discovered" by some), it is not new nor is it untried. National and statewide performance assessments were successfully conducted in a reliable and cost-efficient manner decades ago. As this guide has illustrated, performance assessment is feasible and manageable. Such assessments are vitally needed in the assessment landscape so that those interested in assessing what students are capable of doing have access to more complete information on student performance. Although the steps are more complex and more involved, such assessments are important in the determination of what skills our students need to have and whether or not they do in fact have them. Performance assessment is an important adjunct to overall large-scale assessment strategies.

North Central Regional Educational Laboratory
1900 Spring Road, Suite 300
Oak Brook, IL 60521-1480
(708) 571-4700
Fax (708) 571-4716