Listening. Learning. Leading.®

Guidelines for Constructed-Response and Other Performance Assessments

2005

The 2005 Performance Assessment Team for the ETS Office of Professional Standards Compliance:

Doug Baldwin
Mary Fowles
Skip Livingston

An Explanation of These Guidelines

One of the most significant trends in assessment has been the recent proliferation of constructed-response questions, structured performance tasks, and other kinds of free-response assessments that ask the examinee to display certain skills and knowledge. The performance, or response, may be written in an essay booklet, word-processed on a computer, recorded on a cassette or compact disc, entered within a computer-simulated scenario, performed on stage, or presented in some other non-multiple-choice format. The tasks may be simple or highly complex; responses may range from short answers to portfolios, projects, interviews, or presentations. Since 1987, when these guidelines were first published, the number and variety of ETS performance assessments have continued to expand, in part due to ongoing cognitive research, changes in instruction, new assessment models, and technological developments that affect how performance assessments are administered and scored.

Although many testing programs have more detailed and program-specific performance-assessment policies and procedures, the guidelines in this document apply to all ETS testing programs. This document supplements *ETS Standards for Quality and Fairness** by identifying standards with particular relevance to performance assessment and by offering guidance in interpreting and meeting those standards. Thus, ETS staff can use this document for quality-assurance audits of performance assessments and as a guide for creating such assessments.

* The *ETS Standards for Quality and Fairness* is designed to help staff ensure that ETS products and services demonstrably meet explicit criteria in the following important areas: developmental procedures; suitability for use; customer service; fairness; uses and protection of information; validity; assessment development; reliability; cut scores, scaling, and equating; assessment administration; reporting assessment results; assessment use; and test takers' rights and responsibilities.

Contents

Introduction

Testing is not a private undertaking but one that carries with it a responsibility to both the individuals taking the assessment and those concerned with their welfare; to the institutions, officials, instructors, and others who use the assessment results; and to the general public. In acknowledgment of that responsibility, those in charge of planning and creating the assessment should do the following:

- *Make sure the group of people whose decisions will shape the assessment represents the demographic, ethnic, and cultural diversity of the group of people whose knowledge and skills will be assessed.* This kind of diversity is essential in the early planning stages, but it is also important when reviewing assessment content, establishing scoring criteria, scoring the responses, and interpreting the results.
- *Make relevant information about the assessment available during the early development stages so that those who need to know (e.g., sponsoring agencies and curriculum coordinators) and those who wish to know (e.g., parents and the media) can comment on this information.* The development of a new assessment should include input from the larger community of stakeholders who have an interest in what is being assessed and how it is being assessed.
- *Provide those who will take the assessment with information that explains why the assessment is being administered, what the assessment will be like, and what aspects of their responses will be considered in the scoring.* Where possible and appropriate, test takers should have access to representative tasks, rubrics, and sample responses well before they take the assessment. At the very least, all test takers should have access to clear descriptions of the types of tasks they will be expected to perform and explanations of how their responses will be assessed.

This document presents guidelines that are designed to assist staff in accumulating validity evidence for performance assessments. An assessment is valid for its intended purpose if the inferences to be made from the assessment scores (e.g., that a test taker has mastered the skills required of a foreign language translator or has demonstrated the ability to write analytically) are appropriate, meaningful, useful, and supported by evidence. Documenting that these guidelines have been followed will help provide evidence of validity.

Key Terms

The following terms are used throughout the document.

- *Task* = A specific item, topic, problem, question, prompt, or assignment
- *Response* = Any kind of performance to be evaluated, including short answer, extended answer, essay, presentation, demonstration, or portfolio
- *Rubric* = The scoring criteria, scoring guide, rating scale and descriptors, or other framework used to evaluate responses
- *Scorers* = People who evaluate responses (sometimes called readers, raters, markers, or judges)

Planning the Assessment

Before designing the assessment, developers should consult not only with the client, external committees, and advisors but also with appropriate staff members, including assessment developers with content and scoring expertise and statisticians and researchers experienced in performance assessment. Creating a new assessment is usually a recursive, not a linear, process of successive refinements. Typically, the assessment specifications evolve as each version of the assessment is reviewed, pretested, and revised. Good documentation of the process for planning and development of the assessment is essential for establishing evidence to support valid use of scores. In general, the more critical the use of the scores, the more critical the need to retain essential information so that it is available for audits and external reviews. Because much of the terminology in performance assessment varies greatly, it is important to provide examples and detailed descriptions. For example, it is not sufficient to define the construct to be measured with a general phrase (e.g., "critical thinking") or to identify the scoring process by a brief label (e.g., "modified holistic").

Because the decisions made at the beginning of the assessment affect all later stages, developers must begin to address at least the following steps, which are roughly modeled on evidence-centered design, a systematic approach to development of assessments, including purpose, claims, evidence, tasks, assessment specifications, and blueprints.

1. *Clarify the purpose of the assessment and the intended use of its results.* The answers to the following questions shape all other decisions that have to be made: "Why are we testing? What are we testing? Who are the test takers? What types of scores will be reported? How will the scores be used and interpreted?" In brief, "What claims can we make about those who do well on the test or on its various parts?" It is necessary to identify not only how the assessment results should be used but also how they should not be used. For example, an assessment designed to determine whether individuals have the minimum skills required to perform occupational tasks safely should not be used to rank order job applicants who have those skills.

2. *Define the domain (content and skills) to be assessed.* Developers of assessments often define the domain by analyzing relevant documents such as textbooks, research reports, or job descriptions; by working closely with a development committee of experts in the field of the assessment; by seeking advice from other experts; and by conducting surveys of professionals in the field of the assessment (e.g., teachers of a subject or workers in an occupation); and of prospective users of the assessment.

3. *Identify the characteristics of the population that will take the assessment and consider how those characteristics might influence the design of the assessment.* Consider, for example, the academic background, grade level, regional influences, or professional goals of the testing population. Also determine any special considerations that might need to be addressed in content and/or testing conditions, such as physical provisions, assessment adaptation, or alternate forms of the assessment administrator's manual.

4. *Inform the test takers, the client, and the public of the purpose of the assessment and the domain of knowledge and skills to be assessed. Explain how the selection of knowledge and skills to be assessed is related to the purpose of the assessment.* For example, the assessment of a portfolio of a high school student's artwork submitted for advanced placement in college should be directly linked to the expectations of college art faculty for such work and, more specifically, to the skills demonstrated by students who have completed a first-year college art course.

5. *Explain why performance assessment is the preferred method of assessment and/or how it complements other parts of the assessment.* Consider its advantages and disadvantages with respect to the purpose of the assessment, the use of the assessment scores, the domain of the assessment, other parts of the assessment (where relevant), and the test-taker population. For example, the rationale for adding a performance assessment to an existing multiple-choice assessment might be to align the assessment more closely to classroom instruction. On the other hand, the rationale for using performance assessments in a licensure examination might be to require the test taker to perform the actual operations that a worker would need to perform on the job.

6. *Consider possible task format(s), timing, and response mode(s) in relation to the purpose of the assessment and the intended use of scores.* Evaluate each possibility in terms of its aptness for the domain and its appropriateness for the population. For example, an assessment of music ability might include strictly timed sight-reading exercises performed live in front of judges, whereas a scholarship competition that is based on community service and academic progress might allow students three months to prepare their applications with input from parents and teachers.

7. *Outline the steps that will be taken to collect validity evidence.* Because performance assessments are usually direct measures of the behaviors they are intended to assess, content-related evidence of validity is likely to receive a high priority (although other kinds of validity evidence may also be highly desirable). This kind of content-related evidence often consists of the judgments of experts who decide whether the tasks or problems in the assessment are appropriate, whether the tasks or problems provide an adequate sample of the test taker's performance, and whether the scoring system captures the essential qualities of that performance. It is also important to make sure that the conditions of testing permit a fair and standardized assessment. See the section Using Statistics to Evaluate the Assessment and the Scoring at the end of this document.

8. *Consider issues of reliability.* Make sure that the assessment includes enough independent tasks (examples of performance) and enough independent observations (number of raters independently scoring each response) to report a reliable score, given the purpose of the assessment.

   A test taker's score should be consistent over repeated assessments using different sets of tasks drawn from the specified domain. It should be consistent over evaluations made by different qualified scorers. Increasing the number of tasks taken by each test taker will improve the reliability of the total score with respect to different tasks. Increasing the number of scorers who contribute to each test taker's score will improve the reliability of the total score with respect to different scorers. (If each task is scored by a different

scorer or team of scorers, increasing the number of tasks will automatically increase the number of scorers and will therefore increase both types of reliability.) The scoring reliability on each given task can be improved by providing scorers with specific instructions and clear examples of responses to define the score categories. Both an adequate sample of tasks and a reliable scoring procedure are necessary; neither is a substitute for the other.

In some cases, it may be possible to identify skills in the domain that can be adequately measured with multiple-choice items, which provide several independent pieces of information in a relatively short time. In this case, a combination of multiple-choice items and performance tasks may produce scores that are more reliable and just as valid as the scores from an assessment consisting only of performance tasks. For example, an assessment that measures some of the competencies important to the insurance industry might include both multiple-choice questions on straightforward actuarial calculations and more complex performance tasks such as the development of a yield curve and the use of quantitative techniques to establish investment strategies. Many academic assessments include a large set of multiple-choice questions to sample students' knowledge in a broad domain (e.g., biology) and a few constructed-response questions to assess the students' ability to apply that knowledge (e.g., design a controlled experiment or analyze data and draw a conclusion).

Writing the Assessment Specifications

Assessment specifications describe the content of the assessment and the conditions under which it is administered (e.g., the physical environment, available reference materials, equipment, procedures, timing, delivery medium, and response mode). For performance tasks and constructed-response items, the assessment specifications should also describe how the responses will be scored. When writing assessment specifications, be sure to include the following information:

1. *The precise domain of knowledge and skills to be assessed.* Clearly specify the kinds of questions or tasks that should be in the assessment. For instance, instead of "The student reads a passage and then gives a speech," the specifications might say "The student has ten minutes to read a passage and then prepare and deliver a three-minute speech based on the passage. The passage is 450–500 words, at a tenth-grade level of reading difficulty, and about a current, controversial topic. The student must present a clear, well-supported, and well-organized position on the issue."

   As soon as possible in the item development process, create a model or shell (sample task with directions, timing, and rubric) to illustrate the task dimensions, format, appropriate content, and scoring criteria.

2. *The number and types of items or tasks in the assessment.* Increasing the number of tasks will provide a better sample of the domain and will produce more reliable scores but will require more testing time and will increase scoring costs.

   For example, suppose that a state plans to assess the writing skills of all eighth-grade students. To find out "how well individual students write," the state would need to assess

students in several different types of writing. If time and resources are sufficient to assess only one type of writing, the state has a number of options. It can narrow the content domain (e.g., assess only persuasive writing). Alternatively, it can create a set of writing tasks testing the different types of writing and administer them to different students, testing each student in only one type of writing; some students would write on task A (persuasive), some on task B (descriptive), some on task C (narrative), and so on. The resulting data would enable statisticians to estimate the performance of all students in the state on each of the tasks. Another option would be to administer only one or two writing tasks in the state's standardized assessment but to evaluate each student's writing more broadly through portfolios created in the classroom and evaluated at the district level by teachers who have been trained to apply the state's scoring standards and procedures for quality control.

From the point of view of reliability, it is better to have several short tasks than one extended task. However, if the extended task is a more complete measure of the skills the assessment is intended to measure, the assessment planners may need to balance the competing goals of validity and reliability.

3. *Cultural and regional diversity.* Specify, where appropriate, what material will be included to reflect the cultural and regional background and contributions of major groups within both the population being tested and the general population. For example, an assessment in American literature might include passages written by authors from various ethnic groups representing the population being assessed.

4. *Choice, where appropriate, of tasks, response modes, or conditions of testing.* On some assessments, the test takers are allowed to choose among two or more specific tasks (e.g., choosing which musical selection to perform). On some assessments, test takers are allowed to choose the response mode (e.g., writing with pencil and paper or word processing at a computer) or to choose some aspect of the conditions of testing (e.g., what car to use in taking a driving assessment). Whether or not to allow these kinds of choices will depend on the skills the assessment is intended to measure and on the intended interpretation of the assessment results: as a reflection of the test takers' best performance or of their typical performance over some domain of tasks or conditions. Although test takers do not always choose the tasks, response mode, or conditions in which they perform best, the test takers are likely to perceive the assessment as fairer if they have these choices.[1]

5. *The relative weight allotted to each task, to each content category, and to each skill being assessed.* There is no single, universally accepted formula for assigning these weights. Typically, the weights reflect the importance that content specialists place on the particular kinds of knowledge or skills that the assessment is designed to measure. One common practice is to weight the tasks in proportion to the time they require. Since more complex tasks require more time than simpler tasks, they receive greater weight in the scoring. However, the assessment makers must still decide how much testing time to allocate to each skill or type of knowledge to be measured.

1 For a more in-depth discussion of this issue as applied, for instance, to writing assessment, see "Task Design: Topic Choice" in *Writing Assessment in Admission to Higher Education: Review and Framework*, Breland, Hunter, Brent Bridgeman, and Mary Fowles, College Board (Report #99-3) and GRE (#96-12R).

Another common approach is to assign a weight based on the importance of the particular task or, within a task, on the importance of a particular action, regardless of the amount of time it requires. For example, in an assessment for certifying health-care professionals, the single most important action may be to verify that a procedure about to be performed will be performed on the correct patient. In this case, an action as simple as asking the patient's name could receive a heavy weight in the scoring.

In some assessments, the weights for the tasks are computed by a procedure that takes into account the extent to which the test takers' performance tends to vary. If Task 1 and Task 2 are equally weighted, but the test takers' scores vary more on Task 1, then Task 1 will account for more of the variation in the test takers' total scores. To counteract this effect, the tasks can be assigned scoring weights that are computed by dividing the intended relative weight for each task by the standard deviation of the test takers' scores on that task, and then multiplying all the weights by any desired constant.

6. *The timing of the assessment.* In most assessments, speed in performing the tasks is not one of the skills to be measured. Occasionally, it is. In either case, it is important to set realistic time requirements. The amount of time necessary will depend on the age and abilities of the test takers as well as on the number and complexity of the tasks. The time allowed for the total administration of the assessment must include the time necessary to give all instructions for taking the assessment. An assessment may have to be administered in more than one session, especially if its purpose is to collect extensive diagnostic information or to replicate a process that occurs over time.

7. *The medium and format of the assessment and the response form.* Specify how the directions and tasks will be presented to the test takers (e.g., in printed assessment booklets, on videotape, or as a series of computer-delivered exercises with feedback). Also specify how and where the test takers will respond (e.g., writing by hand on a single sheet inserted into the assessment booklet, word processing on a computer, speaking into an audiotape, making presentations in small discussion groups in front of judges, or submitting a portfolio of works selected by individual students or candidates).

8. *Permission to use responses for training purposes.* It may be necessary to ask test takers to sign permission statements giving the program the right to use their responses or performances for certain purposes (e.g., using their responses for training raters or to provide examples to other test takers or for research).[2] To the extent appropriate, clearly explain to the test takers why the information is being requested.

9. *Measures to prevent the scorers from being influenced by information extraneous to the response.* Seeing the test taker's name, biographical information, or scores given on this or on other tasks could bias the scorer's evaluation of the response. It is often possible to design procedures that will conceal this information from the scorers. At the very least, the scorers should be prevented from seeing this information inadvertently. Many performance assessments involve live or videotaped performances. In these situations, programs may need to take special steps in training scorers and monitoring scorer performance to ensure that scorers are not biased by irrelevant information about the test takers.

2 If possible, assessment information bulletins should provide actual responses created under normal testing conditions. However, because the responses must not reveal any information that could identify the specific test taker, it may be necessary to edit the sample responses. (In the case of videotaped responses, for instance, the program could hire actors to reenact assessment performances so that test takers could not be identified.)

10. *The intended difficulty of the tasks.* The ideal assessment difficulty depends on the purpose of the assessment. It may be appropriate to specify the level at which the assessment should differentiate among test takers (e.g., chemistry majors ready for a particular course) or the desired difficulty level of the assessment (e.g., the percentage of first-year college students expected to complete the task successfully). To determine that the assessment and scoring criteria are at the appropriate level of difficulty one should pretest the tasks or questions and their scoring criteria or use collateral information such as previous administrations of similar items.

    For certain criterion-referenced assessments, however, a task may be exceedingly difficult (or extremely easy) for test takers and still be appropriate. For example, an assessment of computer skills might require test takers to cut and paste text into a document. The standard of competency is inherent in the task; one would not necessarily revise the task or scoring criteria depending on how many test takers can or cannot perform the task successfully.

11. *The way in which scores on different forms of the assessment will be made comparable.* Often it is necessary to compare the scores of test takers who were tested with different forms of the assessment (i.e., versions containing different tasks measuring the same domain). The scoring criteria may remain the same, but the actual tasks—items, problems, prompts, or questions—are changed because of the need for security. On some assessments, it may be adequate for the scores on different forms of the assessment to be only approximately comparable. In this case, it may be sufficient to select the tasks and monitor the scoring procedures in such a way that the forms will be of approximately equal difficulty. On other assessments, it is important that the scores be as nearly comparable as possible. In this case, it is necessary to use a statistical adjustment to the scores to compensate for differences in the difficulty of the different forms. This adjustment procedure is called equating.

    To make the unadjusted scores on different forms of the assessment approximately comparable, the tasks on the different forms must be of equal difficulty, and the scoring procedure must be consistent over forms of the assessment. To select tasks of approximately equal difficulty, assessment developers need an adequate selection of tasks to choose from, and they need information that accurately indicates the difficulty of each task—ideally, from pretesting the tasks with test takers like those the assessment is intended for. If adequate pretest data cannot be obtained, the use of variants—different versions of specific tasks derived from common shells—can help promote consistent difficulty. Consistency of scoring requires that the scoring criteria be the same on all forms of the assessment and that they be applied in the same way. Some procedures that help maintain consistency of scoring include using scored responses from previous forms of the assessment to establish scoring standards for each new form (when the same item or prompt is used in different forms) and including many of the same individual scorers in the scoring of different forms. However, at best, these procedures can make the scores on different forms of the assessment only approximately comparable.

The score-equating approach requires data that link test takers' performance on the different forms of the assessment. These data can be generated by at least three different approaches:

a) Administering two forms of the assessment to *the same test takers*. In this case, it is best to have half the test takers take one form first and the other half take the other form first. This approach produces highly accurate results without requiring large numbers of test takers, but it is often not practical.

b) Administering two or more forms of the assessment to large groups of test takers *selected so that the groups are of equal ability* in the skills measured by the assessment. This approach requires large numbers of test takers to produce accurate results, and the way in which the groups are selected is extremely important.

c) Administering different forms of the assessment to different groups of test takers *who also take a common measure* of the same or closely related skills. This common measure is called an anchor; it can be either a separate measure or a portion of the assessment itself.

The anchor-equating approach requires that the difficulty of the anchor be the same for the two groups. If the anchor consists of constructed-response or performance tasks, the anchor scores for the two groups must be based on ratings generated at the *same scoring session*, with the responses of the groups interspersed for scoring, even though the two groups may have taken the assessment at different times.

Analysis of the data will then determine, for each score on one form, the comparable score on the other form. However, the scale on which the scores will be reported limits the precision with which the scores can be adjusted. If the number of possible score levels is small, it may not be possible to make an adjustment that will make the scores on one form of the assessment adequately comparable to scores on another form.

It is important to remember that equating is meant to ensure comparable scores on different versions of the same assessment. Equating cannot make scores on two different measures equivalent, particularly if they are measures of different knowledge or skills.

12. *The response mode(s) that will be used.* Whenever there is more than one possible mechanism for having test takers respond to the task—whether the test takers are offered a choice of response modes or not—it is important to examine the comparability of the scores produced by the different response modes. Different response modes may affect the test takers' performance differently. For example, if a testing program requires word-processed responses, the test takers' ability to respond to the task may be affected (positively or negatively) by his or her keyboarding skills. Different response modes may also have different effects on the way the responses are scored. For example, word-processed responses may be scored more stringently than handwritten responses.

Writing the Scoring Specifications

When specifying how the responses should be scored, the assessment planners should consider the purpose of the assessment, the ability levels of the entire group being tested, and the nature of the tasks to be performed. With the help of both content and performance-scoring specialists, specify the following general criteria:

1. *The method to be used for scoring the responses to each task.* One important way in which methods of scoring differ is in the extent to which they are more analytic or more holistic. An analytic method, including skill assessments with a series of checkpoints, requires scorers to determine whether, or to what degree, specific, separate features or actions are present or absent in the response. Holistic scoring, as currently implemented in most large-scale writing assessments, uses a rubric and training samples to guide scorers in making a single, qualitative evaluation of the response as a whole, integrating discrete features into a single score. Trait scoring, a combination of analytic and holistic scoring, requires the scorer to evaluate the response for the overall quality of one or more separate features, or traits, each with its own rubric. Still another combined approach, core scoring, identifies certain essential traits, or core features, that must be present for a critical score and then identifies additional, nonessential features that cause the response to be awarded extra points beyond the core score.

   The scoring method should yield information that serves the purpose of the assessment. For example, a global or holistic scoring method might not be appropriate for diagnosis, because it does not give detailed information for an individual; an analytic method might be better suited to this purpose—assuming it is possible to isolate particular characteristics of the response.

2. *The number of score categories (e.g., points on a scale, levels of competency, rubric classifications) for each task.* In general, one should use as many score categories as scorers can consistently and meaningfully differentiate. The number of appropriate score categories varies according to the purpose of the assessment, the demands of the task, the scoring criteria, and the number of clear distinctions that can be made among the responses. An analytic method is typically based on a list of possible features, each with a two-point (yes/no) scale. A holistic method typically uses four to ten score categories, each described by a set of specific criteria. A typical trait method might use anywhere from three to six categories for each trait.

   Pilot test sample tasks or items with a representative sample of test takers and evaluate the responses to confirm that the number of score categories is appropriate. For example, suppose that a constructed-response item requires test takers to combine information into one coherent sentence. At the design stage, a simple three-point scale and rubric might seem adequate. However, when evaluating responses at the pilot test stage, assessment developers and scorers might discover an unexpectedly wide range of responses and decide to increase the score scale from three to four points.

3. *The specific criteria (e.g., rubric, scoring guide, dimensions, checkpoints, descriptors) for scoring each task.* Once again, consider the purpose of the assessment, the ability levels of the test takers, and the demands of the task before drafting the criteria. It is important

that the scoring criteria be aligned with the directions and task so that the scorers are using appropriate criteria and so that scorers do not reward certain formats or specific kinds of information not explicitly required by the item or task, thereby penalizing otherwise competent responses.

For example, suppose that test takers were asked to write an essay explaining their own views on an environmental issue. Appropriate criteria on a writing assessment might require test takers to present their ideas on the issue, support their ideas with relevant details and clear examples, organize the information logically, and communicate their ideas clearly and correctly. Inappropriate criteria on such an assessment might specify that for the essay to receive the highest score, a certain number of examples must be presented, or a certain organizational strategy must be followed. These latter criteria—although easy for scorers to follow—would shift the emphasis of the assessment from evaluating the quality of the test taker's thinking and writing to simply measuring the quantity of examples used or the ability of the test taker to apply a predetermined format. Another type of inappropriate criterion is evaluating test takers on how well they address issues not mentioned in the directions or on how well they follow a specific format not delineated in the directions.

There are several ways to establish scoring criteria. For a job-related, criterion-referenced performance assessment, a committee of experts might specify a set of behaviors that the test taker must demonstrate in order to perform the task satisfactorily. For other criterion-referenced assessments, a committee might conduct research (e.g., opinion surveys, observations, pilot tests) in order to develop a set of appropriate criteria. No matter what methods are used, it is usually necessary to try out the criteria, revise them on the basis of test taker responses and the ability of scorers to apply the criteria to the responses, and then try out the criteria again before using them operationally.

4. *The number of scorers who will evaluate each response.* Several factors should influence this decision. In addition to considerations of schedule and budget, these factors should always include at least the following:

    - The importance of the decisions that will be based on the scores
    - The reliability of the process of rating the individual responses
    - The number of different scorers whose ratings will contribute to a single test taker's total score on a given assessment (this number will depend on the number of tasks to which each test taker responds, the number of scorers rating each response, and the number of an individual test taker's responses that are rated by the same scorer)

If a performance assessment or a constructed-response assessment is to be used for a high-stakes decision, it is important that the process of rating the responses be highly reliable. A test taker's score must be—as nearly as possible—the same, no matter which individual scorers rate that test taker's responses. If the assessment results will be used only for diagnostic feedback or guidance (e.g., to make the test taker aware of his/her strengths and weaknesses), a lower level of reliability may be adequate.

Some kinds of responses can be scored very reliably. These tend to be the kinds of responses for which the scoring criteria are highly explicit and the relevant characteristics of the response are easily observed. For these responses, a single rating may be adequate even if the assessment scores are used for important decisions. Other kinds of responses are more challenging to score reliably. If a substantial portion of a test taker's score depends on a response of this latter kind, that response should receive at least two independent ratings, and there should be a procedure for resolving any significant disagreements between those two ratings.

If an assessment includes several performance tasks, with no single task accounting for a large proportion of the test taker's score, a program might decide to single-score the majority of responses to each task and second-score a specified number of the responses in order to monitor inter-rater reliability. Some performance assessments include very few separately scored tasks—in some cases, only one or two. If an assessment consists entirely of a single exercise, with each response scored by a single scorer, that individual scorer will determine the test taker's entire score. If the scorer reacts in an atypical way to the response (e.g., the response may have an unusual approach to the task), the test taker's score for the entire assessment will be inaccurate. Scorer training can reduce the frequency of these anomalous ratings, but it cannot completely eliminate them. The safest way to minimize this effect is to provide thorough training *and* to increase the number of different scorers whose ratings determine an individual test taker's score. If the number of separate exercises is small, it will be necessary to have each response rated independently by at least two different scorers. If the number of exercises is large enough, a single rating of each response may be adequate, even for high-stakes decisions. For example, suppose that a social studies assessment consists of twelve separate exercises and each of the test taker's twelve responses is evaluated by a different scorer. In this case, an individual scorer can influence only one-twelfth of the test taker's total score.

Other factors can also influence scoring decisions. If the assessment is used with a cut score, the program might, for instance, decide to second-score all responses from test takers whose total scores are just below the cut-point. Or suppose that a school district requires all students to pass a critical reading and writing assessment consisting of two tasks as a requirement for graduation. The district might decide to double-score all of the responses (two different scorers per response, for a total of four different scorers per student). Then, because of the importance of the assessment results, the district might specify that all responses from failing students whose total score is within a specified distance of the cut-point be evaluated by yet another group of scorers, who would need to confirm or override the previous scores (the program would need to have clear guidelines for resolving any such overrides). This procedure helps ensure that the scoring is fair and reliable.

5. *Policies and procedures to follow if a test taker formally requests that his or her response be rescored.* According to the ETS Constructed-Response Score Review Policy, each program is required to develop a detailed and reasonable plan for reviewing scores of constructed responses. This plan should specify how long test takers have to challenge their reported scores and what they must do (and how much they must pay) to have their responses rescored. The plan should establish procedures for rescoring the responses, including the

qualifications of the scorers (many programs decide to use their most experienced and reliable scorers for this procedure) and should specify rules for using the results of the rescoring (possibly in combination with those of the original scoring) and the conditions under which a revised score will be reported. The plan should also define procedures for the reporting of scores that have been revised as a result of rescoring.

6. *Policies and procedures to follow if scorers encounter responses that contain threats, admissions of wrongdoing, reports of abuse or violence, references to personal problems, or other emotionally disturbing material.* In some programs, especially in K–12 assessments, these procedures (including timelines for alerting appropriate agencies) may be state mandated.

Reviewing the Tasks and the Scoring Criteria

All tasks and rubrics should be created and reviewed by qualified individuals: content and assessment-development specialists as well as educators, practitioners, or others who understand and can represent the test taker population. Reviewers should evaluate each task together with its directions, sample responses, and scoring criteria so that they can determine that the test takers are told to respond in a way that is consistent with the way their responses will be evaluated. Reviewers should also assess each task in relation to its response format; that is, the space and structure in which the test taker responds or performs. The reviews should address at least the following questions:

1. *Is each task appropriate to the purpose of the assessment, the population of test takers, and the specifications for the assessment?*

2. *Does the assessment as a whole (including any multiple-choice sections) represent an adequate and appropriate sampling of the domain of knowledge and skills to be measured?*

3. *Are the directions for each task clear, complete, and appropriate?*

   Test takers should be able to understand readily what they are to do and how they are to respond. (Often, practice materials available to test takers provide sample tasks, scoring guides, and sample responses.)

4. *Is the phrasing of each task clear, complete, and appropriate?*

   The tasks need to be reviewed from the perspective of those taking the assessment to make certain that the information is not confusing, incomplete, or irrelevant. Occasional surveys of test takers can provide feedback that can answer this, as well as the previous, question.

5. *Are the scoring rubrics for each task worded clearly, efficient to use, and accompanied by responses that serve as clear exemplars (e.g., prescored benchmarks and rangefinders) of each score point?*[3]

   When the overall quality of the response is being evaluated, as in holistic scoring, each score level usually describes the same features, but with systematically decreasing or increasing levels of quality. The scoring criteria should correspond to the directions.

3 Benchmarks and rangefinders refer to sample responses preselected to illustrate typical performances at each score point. In training sessions, scores typically appear on benchmark responses but not on rangefinders. The primary purpose of benchmarks is to show scorers clear examples at each score point. The purpose of rangefinder sets is to give scorers practice in assigning scores to a variety of responses exemplifying the range of each score point.

For example, if the directions tell test takers to analyze something, the scoring rubric should include analysis as an important feature. However, no matter how well crafted a scoring rubric may appear, its effectiveness cannot be judged until it has been repeatedly applied to a variety of responses.

6. *Are the formats of both the assessment and the response materials appropriate*?

   Both the demands of the task and the abilities of the test takers need to be considered. For example, secondary school students who must write an essay may need two or more pages of lined paper and perhaps space for their notes. Elementary school students will need paper with more space between the lines to accommodate the size of their handwriting.

7. *Is the physical environment appropriate for the assessment and the test takers*?

   For example, dancers may need a certain kind of floor on which to perform, speakers may have certain acoustical needs in order for their responses to be recorded, and elementary students may need to conduct their science experiments in a familiar and comfortable setting (e.g., in their classroom instead of on stage in a large auditorium).

8. *Do the materials associated with the scoring (e.g., scoring sheet or essay booklet) facilitate the accurate recording of scores? Do they prevent each scorer from seeing ratings assigned by other scorers and, to the extent possible, from identifying any test taker*?

   The scoring format should be as uncomplicated as possible so that the scorers are not likely to make errors when recording their scores. Also, the materials should conceal or encode any information that might improperly influence the scorers, such as scores assigned by other scorers or the test taker's name, address, sex, race, school, or geographic region.

   For programs in which responses will be electronically scanned for scorers to read online, it is imperative that assessment booklets be designed to minimize the chances of test takers writing their responses outside the area to be scanned or putting their answers on the wrong pages.

9. *Do the tasks and scoring criteria meet ETS standards for fairness*?

   Specially trained reviewers should examine all tasks and scoring criteria to identify and eliminate anticipated sources of bias. Sources of bias include not only racist, sexist, and other offensive language but also assumptions that the person performing the task will hold certain attitudes or have had certain cultural or social experiences not required as part of the preparation for the assessment.[4] Reviewers should ensure that the tasks do not present unnecessarily sensitive or embarrassing content or require test takers to reveal their individual moral values or other personally sensitive information.

   If feasible, programs should survey scorers at the end of a scoring session (or on a regular basis, for programs with continuous scoring) to see if they have fairness concerns with the tasks and/or the scoring criteria.

4 All ETS assessments must be approved by trained fairness reviewers. The ETS 2003 *Fairness Review Overview* is available at **http://www.ets.org**.

Pretesting the Tasks

Whenever possible, programs should pretest all performance tasks and directions on a group of people similar to the test takers who will take the operational form of the assessment. Although the purpose of pretesting is to evaluate the tasks, not the test takers, experienced readers should score the pretest responses as they would score responses from an actual administration of the assessment. Pretesting allows you to answer such questions as these:

1. *Do the test takers understand what they are supposed to do?*
2. *Are the tasks appropriate for this group of test takers?*
3. *Does any group of test takers seem to have an unfair advantage—did any test takers earn higher scores for skills or knowledge outside the domain of the assessment?*
4. *Do the tasks elicit the desired kinds of responses?*
5. *Can the responses be easily and reliably scored? Can they be scored with the intended criteria and rating scale?*
6. *Are the scorers using the scoring system in the way it was intended to be used?*
7. *Do the scorers agree on the scores they assign to the responses?*

Pretesting poses special security concerns for performance assessments. Because test takers usually spend considerable time and effort on only a few tasks, they are likely to remember the specific tasks. One solution is to pretest the tasks with an alternate population who will not be taking the assessment. For example, tasks intended for a national population of college-bound high school seniors may be pretested on college freshmen early in the fall. Tasks designed for students in a particular state may be pretested on students of the same age in another state in an effort to keep the assessment content secure. However, it is not always possible to find a comparable population for pretesting purposes, especially if the assessment involves content or skills that are highly specialized or specific to a particular group of test takers. For example, an essay assessment on the geography of Bermuda, for students in Bermuda secondary schools, might cover material taught only to those students. In this case, there would be no other comparable population on which to try out the questions (although it might be feasible to try out parallel questions that assess knowledge of local geography in other areas).

Some writing-assessment programs have addressed the security problem by prepublishing a large pool of topics (from 50 to over 200 for a given task). Before taking the assessment, test takers can read and even study the topics, but they do not know which of those topics they will encounter when they take the assessment. This approach is valid only when the pool of topics is so extensive as to preclude memorizing responses to each one. The required size of the pool will vary, depending on the assessment and on the test-taker population.

Even when the pretest and assessment populations are thought to be comparable, differences between them in demographics, curriculum, and culture can make comparisons between them less useful. Another factor is the difference in the motivation of the test takers. Pretest participants are not as highly motivated to do their best as are test takers in an operational,

high-stakes situation. Thus, pretesting is valuable for trying out new tasks and scoring criteria, but it may be misleading for decisions such as setting the standard for passing the assessment.

If pretesting is not possible, reviewers must specifically address the seven questions listed above and consider carefully how test takers might respond; reviewers should themselves produce responses in order to help evaluate the tasks and scoring criteria.

Scoring the Responses

Successful scoring is the result of careful planning and preparation. That planning should include the following essential steps:

1. *Specifying both the qualifications and characteristics of the scorers, and recruiting scorers who meet those specifications.* Typically, an important qualification is the person's experience in observing the kind of performance being assessed. For example, a middle school music assessment might specify that all scorers have taught instrumental music in grades five, six, or seven for at least five years. An evaluation of a writing-across-the-curriculum program in a secondary school might specify that the group of scorers include representatives of all the disciplines taught in secondary schools. Another important qualification may be other characteristics of the scorers such as demographic group, geographic region, or professional background.
2. *Determining the responsibilities of the scoring leaders (the people who will conduct the scoring), specifying their qualifications, and recruiting qualified people for this role.* Highly competent and experienced scoring leaders are essential to the success of any performance assessment. They have the responsibility for making certain that scorers are thoroughly trained, that the scoring is carefully monitored, and that the entire scoring process is as valid and reliable as possible. They should help plan the scoring.
3. *Specifying and conducting adequate training for the scorers.* It is important that all scorers be trained, carefully and thoroughly, to apply the same scoring criteria in the same way, consistently and accurately. Ideally, training should be interactive, although it need not be conducted face-to-face in a centralized scoring session. Often, scorers come to the training session with preconceptions about what constitutes a good response, and sometimes their preconceptions are not completely consistent with the scoring criteria. In that case, the scorers must learn to put aside their preconceptions and score according to the rubric, so that all test takers' responses will be rated on the same criteria.

   Adequate scorer training, whether conducted in person or online, includes at least the following activities:

   - Showing each scorer the exact assignment and directions that were given to the test takers.
   - Explaining the scoring method for each task and giving explicit instructions and criteria for scoring responses to that task.
   - Giving the scorers sufficient practice in interpreting the rubric for a particular task and applying it to a varied sample of responses representative of those they will be evaluating.

- Giving the scorers feedback on the accuracy of the ratings they assigned to the practice responses.
- Explaining how scorers should handle responses that are not accounted for in the scoring rubric (e.g., responses that are off-topic, blank, written or spoken in a foreign language, or emotionally upsetting and responses that take possibly valid but unanticipated approaches to the topic).

4. *Planning and conducting a process to confirm that the scorers are able to score consistently and accurately before they begin scoring operationally.* In centralized scoring sessions, this confirmation step may be incorporated informally into the training process. When scoring is decentralized, via an online scoring network or similar system, a more formal procedure may be necessary. One such procedure is to require every scorer trainee to pass a certification assessment as a precondition to operational scoring. In the certification assessment, the scorer evaluates a certain number of sample responses for which the correct rating has been predetermined. The scorer must meet a predefined standard for agreement with the correct scores. For example, if the rating scale contains six possible scores, the scorer may be required to assign the correct rating on at least 25 of 50 responses, with no more than two ratings that differ by more than a point from the correct rating. It may be reasonable to allow scorers who fail this assessment to retrain and be retested with another set of sample responses, but with the same requirement for accuracy.
5. *Specifying and carrying out procedures to maintain consistent and accurate scoring throughout each scoring session and, when appropriate, from one scoring session to another.* These procedures should include at least the following steps:
    - Where possible, recalibrating the scorers (that is, retraining them with a new set of prescored responses and/or reviewing the training materials) on a regular basis (e.g., at the start of each new scoring session) to confirm that they are able to resume scoring accurately.
    - Monitoring the agreement of ratings assigned to the same responses (a check on scoring reliability). If the number of inconsistent ratings is unacceptably high, it may be necessary to stop the scoring and retrain the scorers until consistency is established. However, with qualified scorers, careful training, clear tasks, and appropriate scoring criteria, that situation is unlikely to arise.
    - Monitoring the accuracy of scoring (a check on validity and reliability). In addition to agreeing with each other, the scorers must assign scores consistent with those assigned by master scorers, a scoring committee, or some other group that determines the validity or correctness of a score. This can be done by introducing prescored responses into operational scoring, making certain that the scorers do not know which responses are part of this monitoring process, and/or by ongoing backreading of randomly selected responses to check the readers' scores.
    - Establishing standards for accuracy of rating and removing from the pool of scorers any individual scorer who consistently fails to meet these standards.

6. *Designing and conducting a scoring process that will minimize inappropriate influences on the scorers.* Inappropriate influences (see also #9 under Writing the Assessment Specifications) include:

    - Penalties, rewards, or pressures that might jeopardize the accuracy or consistency of scoring. For example, scorers should not be either pressured into or financially rewarded for scoring so rapidly that they cannot make sound judgments about the responses. Nor should the evaluation of scorers (i.e., how their exact and adjacent scores are calculated) result in rewarding scorers for tending to avoid scoring at the extremes of the scoring scale.

    - Fatigue induced by scoring for too long a period; at inappropriate hours; in uncomfortable, noisy, or poorly lighted rooms; or with low-resolution computer-delivered images that cause eyestrain. Scorers need to follow a realistic schedule, one that includes frequent breaks; typically this is no more than seven hours of scoring time per day.

7. *When responses are each rated by two or more scorers, defining discrepant scores, specifying how they will be resolved, and seeing that those specifications are followed at the scoring sessions.* Discrepant scores are unacceptably different scores assigned to the same response. Tolerance for score differences depends on such factors as the total number of points or score classifications for the task, the importance of the score, and the purpose of the assessment.

8. *If the scale used by the scorers has a specified critical score (e.g., a cut-point or passing score), reevaluating any response that one scorer placed at or above the critical score and another scorer placed below the cut-point.* This step is essential when an assessment consists of only one task, the two initial scorers assign scores on opposite sides of the cut-point, and important decisions are based on the results (in general, this is not a recommended assessment design). If possible, the program's most experienced and reliable scorers should make these final evaluations.

9. *Ensuring that the scoring is sufficiently reliable to support the intended interpretation of the scores.* Steps to do this may include:

    - Having more than one person independently evaluate each response. Multiple ratings increase the reliability of the scoring process (i.e., the inter-rater reliability), because the sum or the average of two independent ratings is more reliable than a single rating. Similarly, the sum or average of three independent ratings is more reliable than the sum or average of only two, and so on. The more independent ratings that are included in the sum, the greater the reliability of the scoring process. At least two independent ratings of a sample of responses are necessary for a statistical estimate of the reliability of the scoring process.

    - Having people other than the original scorers reevaluate responses with discrepant ratings. Preferably, the program's most experienced and reliable scorers should resolve any discrepancies. The definition of a discrepancy will generally depend on the rating scale and on the number of responses it is feasible to rescore.

One possibility is to apply a stricter definition of discrepant ratings to the responses of test takers whose total scores are near the cut-point for a pass/fail decision, to minimize the chance that a rater's misjudgment will affect a test taker's pass/fail status.

- Having different scorers score each part of a test taker's assessment, so that no single scorer can have a large effect on the total score. Having different scorers for each task also prevents a halo effect—the tendency of a scorer to be influenced by a test taker's performance or score on another part of the assessment. This step is nearly always a good practice, but it is particularly important if each response will be scored only once.
- Scrambling the order of the responses before each round of scoring. Scorers need to see a representative sample of responses, to reduce the context effect. A context effect is the tendency of a scorer to compare a response with those recently preceding it. For example, a response may receive an inappropriately low score if it follows several very good responses. It may receive an inappropriately high score if it follows several very poor responses. Scrambling ensures that a scorer will not score responses from only one area or from only a particular group of test takers.

10. *For programs that use automated essay scoring software, clearly documenting how the scoring engine is used, on what criteria its scores are based, and how its scoring is monitored.* Whether computer scoring is used in conjunction with human scorers or as a standalone evaluation tool, the software should to the extent possible meet the same performance guidelines as described above for human scorers.

Administering the Assessment

1. *Well in advance of the administration of the assessment, prospective test takers and, as appropriate, others (parents, teachers, school counselors, managers, the client) need to be provided with the following information*:
    - *The knowledge or skills the assessment is designed to assess and the intended use of the scores*
    - *Clear directions for taking the assessment*
    - *Descriptions and typical examples of all types of tasks in the assessment*
    - *A description of the criteria and standards for scoring the responses and of the procedure for converting the ratings of individual responses into the reported scores as well as sample scored responses*
    - *Information about where and how long the responses will be kept (and in what form), whether the test taker can obtain a copy of his/her responses (and if so, how), and how a test taker can request that his/her responses be rescored*
    - *A clear explanation of the correct use of the assessment scores*

2. *Security of the assessment needs to be maintained, which may possibly mean restricting the dates for administration.* Because performance assessments often consist of only a few tasks or even a single task, test takers can easily remember the precise content of the assessment and communicate it to those who have not yet taken the assessment. Those test takers who know the assessment content in advance will have an unfair advantage over those who do not. To avoid this situation, it may be necessary to administer the assessment to all test takers at the same time. Some programs may also wish to consider prepublishing a complete list of all tasks so that all test takers have the same opportunity to prepare (that is, no one has an unfair advantage by having heard what the specific task will be on a given administration of the assessment).

3. *Specifications for timing and assessment conditions need to be followed at each site where the assessment is given.* Assessment administrators need clear, written instructions as well as any necessary training. Those instructions include information about the kinds of advice or materials, if any, that they may offer to the test takers. At the end of the administration, programs should confirm from the assessment administrators that the specifications were followed. Standardized testing conditions help ensure reliability (so that scores can be compared across administrations) and fairness (so that all test takers perform comparable tasks under similar conditions).

Using Statistics to Evaluate the Assessment and the Scoring

The evaluation of any performance assessment should include statistical analyses of the scores on individual tasks and on the assessment as a whole, and these analyses should be performed under the direction of a statistician experienced in performance assessment. Because performance assessments tend to vary widely in format, task type, and scoring methods, no standard set of analyses will be appropriate for all assessment programs. However, the analyses should provide all of the following types of statistical information:

1. *Statistics describing the reliability of the scoring process for individual tasks.* These statistics require that each response (or a representative sample of responses) be scored by at least two different scorers, working independently. If the responses are not routinely double-scored, it may be necessary to double-score a sample of the responses to estimate the reliability of the scoring process. Even if the operational scoring consists of ratings by a single human scorer and an automated scoring program, a valid estimate of the reliability of the scoring process requires a second human scoring of the papers.

   There is more than one way to describe the agreement between the ratings assigned on two scorings of the same responses. The absolute agreement is indicated by the number (or the percentage) of responses that received ratings that agreed exactly, that differed by one point, by two points, etc. (This distribution of differences can be summarized in a single number by computing the mean size of the differences.) The relative agreement—the extent to which the same responses tended to receive above-average, average, and below-average ratings on two independent scorings—is indicated by the correlation between the two sets of ratings. The methods used to estimate reliability should be appropriate for the characteristics of the assessment and the intended use(s) of the results.

2. *Statistics describing the reliability of the scoring process for the assessment as a whole.* These statistics require data from two independent scorings of the responses to each task—if not all the responses, then at least a sample of the responses to each task. If all the responses are double-scored, or if there is a sample of the test takers who have all their responses double-scored, it is possible to compute, for each test taker, a score based only on the first rating of each response and a score based only on the second rating of each response. These data can be used to estimate the reliability of the scoring process that was used operationally. If the double-scored responses are from a different sample of test takers for each task, the estimation procedure is more complicated.

3. *Statistics describing the reliability of the full measurement process, including both the selection of tasks and the scoring of the responses.* A test taker's score on a constructed-response assessment should generalize across both tasks and scorers. The bottom line for reliability is the extent to which test takers' scores would replicate if the same test takers were tested with another set of tasks (constructed to the same specifications) and their responses were scored by another team of scorers (selected by the same criteria and trained by the same procedure).

   The statistics commonly used to describe the reliability of the measurement process are the reliability coefficient and the standard error of measurement. These statistics can be estimated directly from operational data if the assessment consists of several tasks and each of a test taker's responses is scored by a different scorer (or pair of scorers). If the assessment consists entirely of a single task, these statistics can be estimated only by having a sample of test takers take two forms of the assessment. If the assessment consists of a small number of tasks, it may be possible to estimate reliability statistics from operational data, but if each task measures a different type of knowledge or skill, the estimates will show the assessment scores as being less reliable than they actually are.

4. *Statistics describing the performance of the test takers.* These statistics may be computed for individual tasks as well as for the assessment as a whole; they may be computed for subgroups of test takers as well as for the total group. Typically, the statistics will include score distributions and summary statistics: means, standard deviations, and selected percentiles (the 10th, 25th, 50th, 75th, and 90th). They may also include correlations between scores on different portions of the assessment.

   The score distributions and summary statistics indicate the difficulty of the tasks for the group of test takers. They also reflect the extent to which specified subgroups of test takers perform differently from other groups of test takers. The correlations indicate the extent to which the same test takers tend to perform well on different parts of the assessment. (These correlations may be quite weak if the different portions of the assessment assess different knowledge or skills.)

5. *Statistics describing the performance of the individual scorers.* For responses that are double-scored, these statistics should include a comparison of each scorer's assigned ratings with the ratings assigned to the same responses by other scorers. If the responses are not double-scored, the statistics should compare each scorer's ratings on a task with the full group of all ratings of the responses to that same task. In this case (single scoring), it is important to acknowledge the possibility that a scorer might have scored an

atypical sample of responses, particularly if the scorer evaluated a fairly small number of responses. Programs that use monitor (or prescored) papers will have separate statistics that are also extremely useful for evaluating scorers. All such statistics are valuable for identifying scorers who need additional training and for deciding which scorers should be asked to continue as scorers and which should not.

6. *Statistics describing the performance of individual tasks.* These statistics indicate the difficulty of each individual task for the test takers and the extent to which the test takers who do well on the test as a whole tend to do well on any individual task. The difficulty of the task for the test takers is indicated by the distribution of ratings. The relationship between performance on the task and on the test as a whole can best be described by a series of response curves showing, at each point in the test score range, the proportion of the test takers likely to attain each possible rating (i.e., to earn either that rating or a higher rating).

In addition to the analyses listed above, there are other types of analysis that can be useful if the necessary data can be obtained.

DIF statistics. These statistics indicate the extent to which any individual tasks are particularly difficult (or particularly easy) for a focal group of the test takers. A DIF analysis requires a matching criterion so that the members of the focal group are compared with other test takers of similar ability in the skills the constructed-response assessment is intended to measure. It is important that the DIF analysis not compare constructed-response tasks with other types of tasks, such as multiple-choice items. There is substantial evidence that the differences between groups of test takers (e.g., male and female) in their performance on constructed-response tasks often do not parallel the differences in their performance on multiple-choice items.

Trend statistics. These statistics indicate the extent to which the population of test takers is changing, over time, in the skills measured by the assessment. It is important that any scores used as the basis for these statistics be comparable. Scores based on different tasks, computed with no adjustment for possible differences in the difficulty of the tasks, are likely not to be comparable. A common way to evaluate comparability is to have the same responses scored in different years. However, with this approach, a high correlation is no guarantee of comparability; correlations do not reflect differences in the overall level or spread of the scores. Even a high level of absolute agreement does not guarantee comparability, unless the disagreements are evenly divided (i.e., a response is as likely to be rated higher in one year as it is to be rated higher in any other year).

NOTES

NOTES

NOTES