TOEFL®

Monograph Series

MS - 3
SEPTEMBER 1996

A Review of Psychometric and Consequential Issues Related to Performance Assessment

Patricia A. Carey

Educational
Testing Service

A Review of Psychometric and Consequential Issues Related to Performance Assessment

Patricia A. Carey

Educational Testing Service
Princeton, New Jersey
RM-96-3

Foreword

The TOEFL® Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language program development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions were invited to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) test and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. The revised TSE, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The TOEFL CBT will take advantage of the new forms of assessments and improved services made possible by computer-based testing while also moving the program toward its longer-range goals, which include

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Abstract

In the context of TOEFL 2000, performance assessment tasks may be used to evaluate foreign students' English language proficiency in a manner that closely resembles the tasks these students would be required to perform in an academic setting. The implications of using alternative assessments in other high-stakes testing environments similar to TOEFL® have been discussed in the recent literature. This paper summarizes the psychometric and consequential issues involved in the use of performance assessments that are of relevance to the TOEFL 2000 project. Based on this review, several findings are of note: (1) On the whole, results from performance assessments show a high degree of task-specific variance, resulting in lower levels of score reliability than are found in traditional assessments. (2) With careful design of scoring rubrics and training of raters, the magnitude of variance due to raters or interactions of raters with examinees can be kept a level substantially smaller than other sources of error variance, the most notable of which is topic or task specificity. (3) Because performance assessment tasks are typically complex, thereby limiting the number of tasks that can be given in a fixed testing time, and because the method and content of task-based measurement can have a large effect on test scores, performance assessments are particularly context-bound and of limited generalizability.

Acknowledgment

The author wishes to thank Daniel R. Eignor and Carol Myford for reviewing earlier drafts of this manuscript and for providing valuable comments.

Table of Contents

Introduction

The movement toward performance assessment is motivated in large part by a desire to measure important behaviors that are not easily measured by multiple-choice tests. As Mehrens (1992) states, "Multiple-choice tests cannot assess all the important domains of educational goals/objectives. Some types of procedural knowledge are not amenable to multiple-choice types of assessments. Measurement specialists have agreed that objective tests cannot adequately cover all objectives." The movement toward alternative assessment is also motivated, in part, by the belief that current standardized achievement measures are too narrow in scope and have a negative impact on education (Miller & Legg, 1993). The perceived negative consequences of existing high-stakes testing programs that rely on multiple-choice items are critical to the arguments made by proponents of a system of assessments that relies on performance-based assessment tasks that are believed to provide better instructional targets (Linn, 1993a). Although the goal of performance assessment in the context of TOEFL 2000 is not to provide better instructional outcomes, the goal is to evaluate foreign students' English language skills, in a high-stakes environment, in a manner that more closely resembles the tasks they would be required to perform in an academic setting, while also improving washback.

There is much discussion in the literature about how to define exactly what is meant by performance assessment. According to Mehrens (1992), performance tests are "techniques that try to establish what a person can do as distinct from what he knows," and they involve "heavy reliance on observation and professional judgment in the evaluation of the response." Fitzpatrick and Morrison (1971) define a performance test as "one in which some criterion situation is simulated with more fidelity and comprehensiveness than in the usual paper-and-pencil test." In the context of TOEFL 2000, performance assessments can be "valuable tools for measuring communication skills such as reading, writing, speaking, and listening" (Stiggins, 1987).

Because the method and content of the measurement can have a strong effect on test scores that represent internal mental processes, alternative assessments may appear to be context-bound and not generalizable (Miller & Legg, 1993). This is particularly likely to be a potential problem in the context of a test of language skills, where students will be asked to respond in writing to prompts on topics on which they may or may not have prior knowledge. The specificity of the task and performance criteria will also affect the generalizability of the construct. According to Stiggins (1987), one of the most critical problems in designing performance-based assessments is the explicit definition of the performance criteria. Defining precise criteria that can be reliably measured may be very difficult unless the task is limited and standardized; but the more limited and standardized the task is, to ensure, among other things, comparability of scores, the less likely the task will represent a realistic, complex problem (Miller & Legg, 1993).

Performance assessments can be seen as a continuum between multiple-choice responses and original student productions. As Fitzpatrick and Morrison (1971) pointed out, "there is no absolute distinction between performance tests and other classes of tests." The distinction is the degree to which the criterion situation is simulated. Hence, performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments (Messick, 1994). Messick goes on to caution that "such basic assessment issues as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made."

The implications of using alternative assessments for accountability in a high-stakes testing environment have been discussed in the literature. Miller and Legg (1993) examine two aspects of the use and interpretation of standardized alternative assessment procedures: first, their psychometric properties, and second, the consequences of their use. Messick (1989) referred to the psychometric properties aspect of test validation as the evidential basis for test use and interpretation because it is based on the accumulation of psychometric evidence about the test. The second aspect of test use and interpretation is the consequential basis (Messick, 1989), which refers to the reactions of the participants to the testing program. These reactions ultimately affect the uses and interpretations, or the validity, of the test scores (Miller & Legg, 1993).

Although information is limited, some evidence of the psychometric properties of alternative assessments is available. Additionally, the current literature on the consequences of high-stakes testing, which generalizes across traditional assessment programs, can be extrapolated to alternative assessments. This paper will attempt to summarize the psychometric and consequential issues of performance assessments that may be relevant to the TOEFL 2000 project.

Psychometric Properties of Performance Assessments

Validity

Linn, Baker, and Dunbar (1991) suggested eight criteria that need to be studied for the "serious validation" of alternative assessments: intended and unintended consequences of test use, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. Considering validity in terms of consequences forces our attention on aspects of the assessment process that may not be intended or anticipated by the designers of the instruments. We know from experience that results from standardized tests can be corrupted. It should not be assumed that new forms of assessment will be immune to such influences. Directness and transparency, or clarity of scoring, are thought to be important characteristics of an assessment because of the presumed effects they have on teaching and learning. It may be argued that directness is important because focusing on indirect indicator measures may distort instruction. Similarly, transparency is considered important because understanding the basis on which performance will be judged facilitates the improvement of performance (Linn et al., 1991).

Scoring methods are central to the valid interpretation of the results of alternative assessments. While everyone involved must understand how the tasks will be scored, the more precisely the scoring procedures are defined, the more likely that students will be able to memorize acceptable responses (Miller & Legg, 1993). Test tasks based on extended communicative contexts, as are being investigated for TOEFL 2000, appear to be more memorable than stimuli and tasks with little context, such as discrete grammar items. It may be that in achievement testing, one could argue that an examinee who has memorized, or "mastered," the test tasks has demonstrated sufficient knowledge of the content for appropriate decision making. However, for a language proficiency test, it may be possible for an examinee to "perform" at an acceptable level on a large subset of previously used test tasks and yet not have the necessary language competence to pursue an academic program of study in English.

Even if skills can be measured and scored independent of content, method, and context, the primary forms of validity currently reported for alternative assessments derive from their face validity and instructional fidelity (Linn et al., 1991). Parke and Lane (1993) found that variations in the "openness" of the task, including the specific directions to the student, representation of the task problem, and context of the task, can influence the way the student interprets the task and, consequently, the way the student responds to the task. Shavelson, Carey, and Webb (1990) found that, for science assessments, the form in which information was presented in a problem had little effect on the accuracy of students' responses; what did matter was the required form of the response. Face validity, or the appearance of the assessment instrument, however, should never be accepted as sufficient evidence to justify test use and interpretation. Messick (1989) also cautions that content validity is not a sufficient condition to justify test use without understanding the underlying traits being measured (Mehrens, 1992). The need to obtain evidence about consequences is especially compelling for performance-based assessments because particular intended consequences are an explicit part of the assessment system's rationale. The assessments are expected to be used by and have an impact on schools, colleges, and employers. Assessments are expected to have an impact on what and how teachers teach. And assessments are expected to motivate students to put substantially greater effort into their schoolwork (Linn, 1993a).

Linn et al. (1991) caution that an analysis of the cognitive complexity of the tasks and the nature of the responses that they engender should be included among the criteria used in the validation process. Additionally, in view of the limited sampling that is likely to occur with performance-based measures, the tasks selected to measure a given content domain should themselves be worthy of the time and efforts of students and raters. While process sampling will take precedence over traditional content sampling, breadth of coverage should not be overlooked. There may be a trade-off between breadth of content coverage and some of the other criteria. Finally, one of the rationales for more contextualized assessments is that they get students to deal with meaningful problems that provide worthwhile educational experiences (Linn et al., 1991). Magone, Wang, Cai, and Lane (1993) used logical analyses to assess the cognitive complexity of a small sample of mathematics assessment tasks. Their results indicate the tasks assess many of the specified cognitive processes while measuring changes over time in students' mathematical thinking.

In the performance assessment of competencies or other constructs — that is, where the performance is the vehicle, not the target, of assessment — replicability and generalizability can not be ignored. This is because the consistency or variability of the performances contributes to score meaning, as does generalizability from the sample of observed tasks to the universe of tasks relevant to the knowledge or skill domain at issue. The distinction between competence and performance is an old one, especially in linguistics (Chomsky, 1957). The major point is that although competence must be inferred from observations of performances or behaviors (or from their outcomes or products), these inferences are not often straightforward, particularly inferences about lack of competence from poor performance (Messick, 1994).

In Messick's (1994) distinction between construct-centered and task-centered approaches to performance assessment construction, a construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, a construct-centered approach would determine what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors. Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. Focusing on constructs also alerts one to the possibility of construct-irrelevant variance that might distort the task performance, its scoring, or both (Messick, 1994). Construct-irrelevant test variance can take the form of either construct-irrelevant difficulty or construct-irrelevant easiness. The former refers to the possibility that the assessment instrument may be irrelevantly more difficult to some groups of students. For example, in assessing students' ability to reason mathematically, potential irrelevant constructs would be reading comprehension, writing ability, and familiarity with the context of the tasks. Construct-irrelevant easiness refers to the potential for clues in task format to allow some students to respond correctly in ways that are irrelevant to the construct domain being measured; this may be brought about, for example, by flaws in task construction or the invoking of particular test-taking strategies (Lane, 1993).

A task-centered approach to performance assessment appears to be particularly congenial to fields where the mode of teaching emphasizes repeated demonstration, practice, and critique. In this task-centered approach, we must first decide the actual performances that we want students to be good at and then design those performances while at the same time designing a fair and thorough method of grading them. By virtue of the focus on tasks, the concept of irrelevant variance in task performance loses all meaning because every skill required, however task-specific, is relevant to task completion. One of the advantages of the construct-centered approach is that the focal constructs can help guide the selection or construction of tasks that would optimally reveal them, as well as guide the rational development of scoring criteria and rubrics. This might prove especially useful in constructing so-called authentic assessments, which are believed to be representative of the ways in which knowledge and skills are used in real-world contexts (Messick, 1994). Baron (1991) suggests that an effective performance exercise must incorporate the broad ideas and essential concepts, principles, and processes in a discipline. Strategies for making tasks meaningful and engaging for students are also suggested. Students are more likely to find problems that are situated in real-world contexts engaging; and these problems are more likely to facilitate transfer by demonstrating to students that their knowledge and skills are useful in solving real problems (Baron, 1991).

In studying the validity of performance assessments, one should think carefully about whether the intended domains are being assessed, whether they are well sampled, whether — even if well sampled — one can infer to the domain, and what diagnostically one can infer if the performance is not acceptably high. Even if the domain is the correct one and well defined, the sample adequate, and generalizability possible, validity problems can still remain. For instance, if the assessment is not secure, students will be taught how to do the particular tasks. This not only makes the inference to the domain inappropriate, it means one may make an incorrect inference about the sample performance (Mehrens, 1992). As discussed earlier, in such a setting, a student may be able to perform well on a practiced language proficiency assessment and yet not have the language skills necessary to succeed in an English-speaking academic setting.

Two questions are raised with respect to the issue of the convergent validity of measurement methods: (1) to what extent do the achievement estimates for individual students depend on the particular tasks and/or methods sampled? and (2) do measurement methods converge in assessing students' achievement? Shavelson, Baxter, and Gao (1993) examined convergent validity for mathematics and science assessments within the context of Kane's (1982) extension of generalizability theory. Their findings indicate that not all methods converge with one another. Rather, the evidence is that certain methods may measure different aspects of achievement; that is, student performance is dependent on the method sampled. Performance assessments take more time to administer than multiple-choice tests do. Therefore, performance assessments should be used to assess those understandings and processes that cannot be tapped adequately by multiple-choice tests (Baron, 1991).

In an examination of the links between two performance measures and relevant subtests of a widely used standardized norm-referenced test, Burger and Burger (1994) found moderate correlations representing criterion validity for the performance measures, suggesting the two types of assessment are measuring something in common, but they are not measuring identical skills.

In an international comparative study of writing achievement, Purves (1992) found that while there will always be task effects on the results of any assessment of student writing, such effects may be exacerbated by some other force or forces — perhaps instruction, perhaps rating style, or perhaps other aspects of rating difference across countries. That correlations between tasks were higher in the Purves study when the raters focused on a more stable element in student writing (such as grammar or style) gives greater credence to the latter interpretation. Evidence suggests that the construct "writing" may be misleading. There appears to be strong independence among the task scores, sufficient independence to prevent summing them into some construct such as "writing performance" or "writing ability." It is apparent that different tasks present different problems, which are treated differently by students and judged differently by raters (Purves, 1992).

Reliability

Of particular concern in alternative assessment is consistency across tasks and raters. Generalizability theory (Shavelson, Webb, & Rowley, 1989) provides a natural framework for investigating the degree to which performance assessment results can be generalized (Linn et al., 1991). A score's usefulness largely depends on the extent to which it allows us to generalize accurately to behavior in some wider set of situations, or universe of generalization. Instead of asking how accurately observed scores reflect their corresponding true scores, generalizability theory (GT) asks how accurately observed scores permit us to generalize about persons' behavior in a defined universe of situations. GT extends classical theory in several important ways. First, it recognizes multiple sources of measurement error (such as occasions, forms, raters, and items), estimates each source separately, and provides a mechanism for optimizing overall reliability. Second, although GT provides a reliability coefficient, called a "generalizability coefficient," the theory focuses on variance components that index the magnitude of each source of error. Third, GT distinguishes between relative decisions, where interest attaches to the dependability of the differences among individuals, and absolute decisions, where scores are themselves interpretable without reference to others' performance. Fourth, GT distinguishes between generalizability (G) and decision (D) studies.

G studies estimate the magnitude of as many potential sources of measurement error as possible. D studies use information from G studies to design measurements that minimize error for a particular purpose, not only by examining the trade-off between the number of conditions or different facets, but also by considering a wide variety of designs, including crossed, partially nested, and completely nested designs. In general, the G study should be as fully crossed as possible, so that all sources of variations in the design can be estimated. From the results of a fully crossed G study, the generalizability for a wide variety of nested D study designs can then be estimated (Shavelson et al., 1989).

On the whole, reviews of published articles on alternative assessments seem to indicate a high degree of task-specific variance (Miller & Legg, 1993). For example, Shavelson, Baxter, and Pine (1991) found high variability in task performance in hands-on science experiments. They concluded that to get an accurate picture of individual student science achievement, the student must perform a substantial number of exercises (Shavelson & Baxter, 1992). Similarly, high task-specific variance has been found for topics within mode for writing assessments. This high within-mode and within-person variance results in lower reliabilities across tasks than are typically found in traditional assessments (Miller & Legg, 1993).

Shavelson, Baxter, and Gao (1993) found task variability to be the major source of measurement error for mathematics and science assessments; and rater-sampling variability did not appreciably increase measurement error or decrease generalizability. They concluded that a relatively large number of tasks would be needed to reach an approximate .80 G coefficient. In an earlier Marine Corps assessment G study, Shavelson, Mayberry, Li, and Webb (1990) also concluded that task sampling is the critical element for producing reliable job-performance measurements. Lane, Liu, Stone, and Ankenmann (1993) also found that the person-by-task interaction accounted for the greatest variance in math assessments, while the variance due to raters was negligible.

Interrater reliability is clearly also necessary in alternative assessments, particularly those involving writing or speaking tasks, because the scoring procedures are often subjective. However, with careful design of scoring rubrics and training of raters, as has been the case with the TOEFL Test of Written English (TWE®), the magnitude of the variance components due to raters and interactions of raters with examinees can be kept at levels substantially smaller than other sources of error variance, the most notable of which is topic or task specificity (Linn, 1993a). In summarizing a number of studies of performance assessments, Dunbar, Koretz, and Hoover (1991) presented clear evidence that score reliability tends to be markedly lower than rater reliability. The contrast between score and rater reliability introduces an inevitable reliability/validity tradeoff into performance assessment. Tasks that are more narrowly defined might solve the problem of low score reliability of individual tasks, but that narrowing of task definitions poses an unattractive choice in terms of validity: If inferences are kept broad enough to be important, the validity of the tasks, by and large, cannot be determined; if inferences are narrowed to maintain validity in the face of restricted definition of the task, the task or tasks become unimportant (Dunbar et al., 1991). Another approach for enhancing domain coverage and for improving generalizability involves a trade-off between extended performance tasks and briefer structured exercises. It must be recognized that the contrast between multiple-choice items and open-ended performance tasks is not a dichotomy but a continuum along which different sorts of response structures, like structured exercises, can also be situated (Messick, 1994).

Baker (1992) demonstrates that, for purposes of improving generalizability for performance-based history tasks, increasing the number of tasks is more important than increasing the number of trained raters. A similar pattern of more rapidly increasing generalizability as a function of number of tasks rather than as a function of number of raters is reported for open-ended mathematics problems by Lane, Stone, Ankenmann, and Liu (in press).

Mehrens (1992) argued that "the major problems for valid performance assessment relate to the limited sampling and lack of generalizability from the limited sample to any identifiable domain." Low levels of generalizability across tasks limit the validity of inferences about performance for a domain and pose serious problems regarding comparability and fairness to individuals who are judged against performance standards based on a small number, and perhaps a different set, of tasks. This lack of generalizability has implications for TOEFL 2000 when considering the ability to generalize about foreign language competency from performances on a relatively small number of language tasks.

Linn and Burton (1994) suggest, however, that exclusive attention on generalizability coefficients is misguided; it is more informative to focus on the effects of the number of tasks on standard errors and the effects of standard errors of varying sizes on specific decisions or uses of assessment results. In the context of pass/fail decisions, it is more important to know the standard error and what that error implies about the minimum score needed to be confident about a pass decision and the maximum score needed to be confident about a fail decision (Linn & Burton, 1994).

In a study of the influences of item format on the internal characteristics of tests of language skills, Bray and Dunbar (1994) found traditional and integrated item formats to be very similar in average difficulty and discrimination. In this study, traditional items were characterized by tasks involving the identification of errors, with errors separated by type of error and accompanied by minimal context. Integrated items, on the other hand, simulated the revising and editing stages of the writing process. In this format, a number of errors are imbedded in a portion of text. Both formats were able to measure the same skills with essentially the same reliability per item; however the two formats did not offer the same reliability per unit of testing time. For example, the reliability of a 60-minute integrated test was .89; whereas that for a 60-minute traditional format test was .95. Thus, with integrated tasks, it is possible that any given skill may be represented by too few items to allow reliable skill scores to be reported. When extrapolated to a test with lower reliability to begin with, this decrease in reliability, or the increased time required by additional performance tasks, may be unacceptable.

Lane, Liu, Stone, and Ankenmann (1993) suggest that inter-task relationships in science, writing, and mathematics performance assessments have indicated that the generalizability of individual-level scores derived from assessments consisting of relatively small numbers of tasks is questionable. As Messick (1989) has indicated, the context of measurement — such as the assessment procedures — affects the validity of the scores derived from the student responses. Thus, the amount of time allocated for an assessment can affect the validity of the derived scores. An omission of tasks on the part of examinees may be indicative of speededness. Lane et al. (1993) used a likelihood ratio goodness-of-fit statistic to investigate speededness for a mathematics assessment. They found item parameters to be generally invariant across longer and shorter forms of the tests.

Equating

There have been limited studies of the linking of performance assessments to date, and those that have been done have dealt with equating. In a study of equating issues related to a direct writing assessment, Harris, Welch, and Wang (1994) found similar equating results using three methods: linear equating, equipercentile equating, and a Rasch-based equating method. The Rasch-based method utilized the polytomous Rasch partial-credit IRT model (implemented via FACETS; Linacre, 1988) to estimate the item parameter estimates. A bootstrap procedure was used to investigate equating error.

The statistical equivalence of tests over time will need to be examined, and it would seem that equivalent tasks would need to be developed to guarantee equivalent scores. Experience with writing assessment has shown the difficulty of constructing comparable tasks of equal difficulty (Legg, 1987). Given that constructing equivalent tasks can be difficult, some procedure for estimating equivalence of scores and adjusting for lack of equivalence will need to be implemented. However, because performance assessments yield fewer independent pieces of data and because specific assessment tasks should not be reused in large-scale assessment because of security issues, equating problems are formidable. Nevertheless, the scores on different forms of performance assessments will need to be linked so that they represent the same level of achievement regardless of when the performance was assessed, which tasks were given, or which raters scored the performance (Mehrens, 1992).

Linn (1993b) distinguished five types of linking that are applicable to performance assessments by their degree of statistical rigor. Linn referred to these five types of linking as: equating, calibration, statistical moderation, prediction, and social moderation. Equating, the most demanding form of linking, requires forms to measure the same construct with an equal degree of reliability. It is likely to be more difficult to approximate the goal of equating when assessments consist of a relatively small number of tasks than when tests involve a relatively large number of tasks because the relative weight or importance of each task is greater when the number of tasks on an assessment is relatively small and the tasks themselves are likely not to be parallel.

Calibration provides a means of comparing scores or tests that satisfy somewhat less stringent requirements than those for equated tests. As noted by Mislevy and Stocking (see Linn, 1993b), calibration assumes that two tests measure the same thing, but the tests may be designed to measure performance at different levels or with different degrees of reliability. Calibration of tests designed to measure performance at different developmental levels is frequently referred to as vertical equating. For successful calibration, it is important that the two assessments be well matched in terms of content coverage, the cognitive demands that are placed on students, and the conditions under which the assessments are administered.

Statistical moderation is a term that has been used to describe two different situations in which there is a desire to compare results obtained from different sources. In one situation, statistical moderation consists of the use of an external examination, or anchor, to adjust teacher-assigned grades, for example. The process used in some countries to adjust scores on examinations in different subject areas or to compute a total examination score for students taking examinations in different subjects is also referred to as statistical moderation. Statistical moderation of scores to an anchor test can most simply be accomplished by setting the means and standard deviations of the two tests equal. While it is not necessary for the two tests to measure exactly the same thing, this type of moderation may prove to be problematic because scores will be adjusted based on an external examination that is not equally relevant to the content of different forms of the tests. The second type of statistical moderation is used, for example, when comparisons are desired among students who take different combinations of achievement tests (or modules, in a possible TOEFL 2000 scenario). The moderation adjusts scores for differences in means and standard deviations of students taking different tests. Although comparisons are made among students based on their statistically moderated scores on different combinations of tests, the scores cannot be considered equivalent in any rigorous sense.

Prediction is the weakest of the four statistical forms of linking results on one test or set of assessment tasks to another. Predictions can be made as long as there is some relation between the performance on one assessment and the performance on another. The precision of the prediction will depend on the strength of the relations, and the prediction itself is sensitive to context, group, and time. Mislevy and Stocking (see Linn, 1993b) illustrated the group-dependent nature of predictions using an example based on the multiple-choice and essay sections of Advanced Placement (AP®) examinations.

Finally, with social moderation, also called consensus moderation, performances on distinct tasks are rated using a common framework, and are interpreted in terms of a common standard. The primary requirements are concerned with the development of a consensus on the definition of standards and on the performances that meet those standards. In this context, scoring of performances and comparability of scores depends heavily on professional judgements and a system of spotchecks and verification. One could hardly expect that the results would be interchangeable in the strict sense of equated scores. It seems likely, however, that some hybrid approach that supplements social moderation with statistical checks and possibly statistical adjustments will prove to be more acceptable than social moderation alone (Linn, 1993b).

One of the ways in which performance assessments likely differ from traditional multiple-choice tests is with respect to the expected level of local item dependence (LID). LID will clearly have an effect on any IRT-based equating procedure in that IRT models assume that items are locally independent. Because a setting is established for performance assessments, and students can be asked to make multiple responses to directions or questions related to that setting, performance assessments appear likely to produce far greater LID than is produced when using traditional multiple-choice tests. When LID occurs, it can mean that the multiple observations of behavior are not covering as wide a range of behavior as intended. Additionally, item information and reliability will be overstated and measurement error will be underestimated. When positive LID occurs, it increases the strength of the relationship between some items and therefore the relationship between an item and the total test score, producing higher item discrimination for these items. If these LID items are then separated and placed in a variety of subsequent test forms, the predicted discrimination can be inaccurately high (Yen, 1993).

Yen (1993) proposed six procedures as possible ways of reducing LID, or of analyzing data so that the LID has minimal negative effects on measurement characteristics. These include:

(1) Construct multiple independent tasks varying along relevant dimensions; or create multiple independent items within a task.
(2) Administer the test under appropriate conditions, avoiding LID due to speededness, fatigue, or undesired interference or assistance.
(3) Combine the grading of LID items; although it may be difficult to determine a priori in which responses are locally dependent.
(4) Review items empirically using an appropriate goodness-of-fit index.
(5) Construct separate scales for LID items. Alternately, construct a testlet or testlets that contain the LID items.
(6) Use testlets because they provide a more accurate way of relating the performance on that set of items to the other items in the test. The loss of information by the use of testlets can be minimized if only those items that actually are locally dependent are included in each testlet (Yen, 1993).

Procedures for modeling integrated tasks, in which multiple skills are measured within a single prompt, are in the early stages of development (see Mislevy, 1994). These models clearly have potential for use with the types of tasks being pursued by the TOEFL 2000 project.

Consequences of Using Alternative Assessments

Fairness

Fairness clearly is a major consideration in judgments regarding the appropriateness of the uses and interpretations of an assessment. It would be a serious mistake to assume that performance-based assessments are somehow immune to problems of bias or adverse impact (Linn, 1993a). The factors that influence fairness include the relevance of the task for different racial/ethnic and gender subpopulations, the objectivity in scoring, and the degree to which students of all achievement levels have had the opportunity to practice the skills as they will be presented in the assessment (Miller & Legg, 1993).

One important reason for favoring rich contextualization of problems or tasks is to engage student interest and thereby improve motivation and effort. Contextualization seems to imply that the task and its setting are invested with sufficient content so that the problem situation is meaningful in terms of the student's experience, and specific or concrete exemplars can be used as the basis for problem representation and solution. One approach to coping equitably with differential student responsiveness to context is to develop an aggregate measure of the construct across a variety of item contexts in a effort to balance the effects of different student backgrounds and interests (Messick, 1994).

Teaching to the Test

Many alternative assessment procedures, especially those that involve writing, provide a more holistic approach to assessment, which would have a positive effect on those who teach to the test. Such an approach to assessment might also result in a more holistic approach to instruction, in which individual skills are tied together in the instruction of more complex constructs. Alternative assessments should be created in such a way that, if educational institutions teach to the test, they will be teaching what we believe students should know (Shavelson et al., 1990). Alternative assessment also opens more avenues for understanding student learning and diagnosing student skills because it is typically multifaceted and provides a wider profile of student strengths and weaknesses (Miller & Legg, 1993). Substantial changes in instructional strategy and resource allocation are required to give students adequate preparation for complex, time-consuming, open-ended assessments. Validly teaching for success on these assessments is a challenge in itself (Linn et al., 1991).

Measurement-driven instruction may create problems for designing instruction that promotes learning without initial planning of clear instructional strategies from clear objectives. On the other hand, clear instructional strategies will be more difficult to specify without reducing the tested skills to a level that can be learned through drill and practice. An important property of alternative assessment is that learners are expected to apply higher-order thinking skills to new and unfamiliar situations. Or, in the context of TOEFL 2000, examinees will likely be asked to exhibit performances involving highly complex interactions of context and language. Thus, the generalizability of the measured skills across situations and content areas becomes crucial for valid interpretation of test scores (Miller & Legg, 1993).

Test Preparation and Testwiseness

Many preparation programs exist for traditional forms of assessment and are frequently used with high-stakes tests. Test-preparation programs teach test-taking skills, provide practice on skills matched to standardized tests, and frequently present questions in the same format as the test items. Proponents of authentic assessment argue that test-preparation activities would be less costly because the link between instruction and assessment would be closer. Test-preparation activities for authentic assessments would result in learning the underlying skills and concepts, not just how to take the test. However, if objectives were written too specifically or scoring rubrics were known, test preparation and testwiseness might become the same problem as seen with traditional assessments (Miller & Legg, 1993). The extensive TOEFL test-preparation industry, which bases much of its training on disclosed tests, may be able to enhance test performance to the extent that scores on a performance assessment of language proficiency may become distorted. The cost and time involved in generating a sufficiently large pool of assessment tasks in an effort to counteract this problem may prove to be prohibitive.

Test Security and Cheating

The increased complexity of testing procedures with alternative assessment will broaden the range of methods for artificially increasing test scores. Shepard and Dougherty (1991) found that teacher ratings of other teachers revealed several questionable test administration procedures. Once performance assessments have been used in a high-stakes environment, they cannot be reused to test the same higher order thinking process. As discussed earlier, the more precisely the scoring procedures are defined, the more likely that students will be able to memorize acceptable responses (Miller & Legg, 1993). Since performance assessments consist of fewer tasks, memorizing exposed tasks would allow students to recall a larger proportion of the complete testing instrument than is possible with a longer, multiple-choice test. Additionally, the richly contextualized authentic tasks, if compromised, could provide some students the opportunity to prepare for the test by studying the topic(s). Although the tasks in a test of language proficiency may not require prior knowledge of a topic for successful performance, prior knowledge does have a strong influence on the ability of students to comprehend what they read, and such exposure would give an advantage to students who had access over those who had none.

Legal Issues

Courts have ruled that states are forbidden from using competency examinations to grant diplomas unless the tested content had been taught in the schools long enough to provide the opportunity to learn and master the material. Legal challenges might be expected with respect to performance assessments on four issues: magnitude of subpopulation differences; intent; notice; and validity (Miller & Legg, 1993). As Linn, Baker, and Dunbar (1991) state, "Gaps in performance among groups exist because of difference in familiarity, exposure, and motivation on the tasks of interest." Because of the cultural differences inherent in the test-taking population of a test of language proficiency such as the TOEFL test, the typical concerns about racial or gender biases may not be pertinent. However, the cultural differences, themselves, may introduce biases with contextualized tasks that may be differentially interpreted. Differential prior knowledge about the topics may also be exacerbated when the assessment contains fewer tasks, and therefore fewer topics. The changes in instructional strategy required to give students the opportunity to learn the skills necessary for complex assessments will take time to implement, especially in an international setting. Finally, as with traditional assessments, teaching to the test and test-preparation techniques, methods that result in an increase in scores without an increase in skills on the underlying trait measured by the test, have the potential to invalidate the scores as measures of complex language proficiency.

References

Baker, E. L. (1992). The role of domain specifications in improving the technical quality of performance assessment (Tech. Rep.). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

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

Bray, G. B., & Dunbar, S. B. (1994, April). Influence of item format on the internal characteristics of alternate forms of tests of language skills. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice, 13*(1), 9-15.

Chomsky, N. (1957). *Syntactic structures.* The Hague, Netherlands: Mouton.

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

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 237-270). Washington, DC: American Council on Education.

Harris, D. J., Welch, C. J., & Wang, T. (1994, April). Issues in equating performance assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6*(2), 125-160.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice, 12*(2), 16-23.

Lane, S., Liu, M., Stone, C. A., & Ankenmann, R. D. (1993, April). Validity evidence for QUASAR's mathematics performance assessment. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (in press). Reliability and validity of a mathematics performance assessment. *International Journal of Educational Research.*

Legg, S. M. (1987, April). Understanding topic difficulty. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Linacre, J. M. (1988). *FACETS: A computer program for many-facet Rasch measurement.* Chicago: MESA Press.

Linn, R. L. (1993a). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.

Linn, R. L. (1993b). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83-102.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-8.

Magone, M. E., Wang, N., Cai, J., & Lane, S. (1993, April). An analysis of the cognitive complexity of QUASAR's performance assessment tasks and their sensitivity to measuring changes in students' thinking. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (3rd ed.).* New York: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

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

Mislevy, R. J. (1994). Test theory and language learning assessment. Paper presented at the Language Aptitude Invitational Symposium, Center for the Advancement of Language Learning, Arlington, VA.

Parke, C., & Lane, S. (1993, April). Designing performance assessments: An examination of changes in task structure on student performance. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English, 26*(1), 108-122.

Shavelson, R. J., & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49*(8), 20-25.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessments in science. *Applied Measurement in Education, 4*(4), 347-362.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*(9), 692-697.

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine corps rifleman. *Military Psychology, 2*(3), 129-144.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory: New developments and novel applications. *American Psychologist, 44*(6), 922-932.

Shepard, L. A., & Dougherty, K. C. (1991, April). Effects of high-stakes testing on instruction and achievement. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

50% RECYCLED PAPER
10% Post Consumer Waste

Cover Printed on Recycled Paper

58701-13954 • Y96M.5 • 253702 • Printed in U.S.A.