NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Issues in Large-Scale Performance Assessment

Edited by
Gary W. Phillips
Associate Commissioner
National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement
NCES 96-802

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Jeanne E. Griffith
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208–5574

April 1996

Contact:
Arnold A. Goldstein
202–219–1741

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop SSOP, Washington, DC 20402-9328
ISBN 0-16 -048827-0

Foreword

Gary W. Phillips
Associate Commissioner
National Center for Education Statistics

Recently there has been a significant expansion in the use of performance assessment in large scale testing programs. This proliferation has given rise to a range of testing formats such as constructed-responses, essays, experiments, portfolios, exhibitions, interviews, and direct observations. Support for this effort comes primarily from curriculum reformers and policymakers who feel it only makes good sense to test what we teach, and test the way we teach. Many school districts, state testing programs, and national and international assessments have incorporated performance assessments into their programs. Efforts are continuing to use such large scale assessments to shed light on the thinking and learning processes of students, and to encourage teachers to focus their teaching based on the content and skills of the test.

Although, there has been significant support from curriculum and policy stakeholders, the technical feasibility of large scale performance assessments has remained a question. This report is intended to contribute to the debate by reviewing some of the technical issues that must be dealt with by any developer of large scale performance assessments. The report is also intended to surface issues, articulate problems, and where possible, give advice on how to proceed. The report is not intended to be a technical or users manual on how to solve technical problems. The report is being written at a time in which many of the technical problems of large scale performance assessments are just beginning to surface. As these problems are recognized and solved, the state-of-the-art is expected to change rapidly.

The report is divided into five chapters: Validity, Generalizability, Comparability, Performance Standards and Fairness, Equity, and Bias. Each represents a major technical topic that developers of large scale performance assessments should expect to encounter. The following is a brief summary of some of the main points in each chapter.

Validity. This chapter provides a comprehensive view of validity. Validity is defined as not only the evaluative or evidential information about score inferences but also information about actual as well as potential consequences of score interpretations. It is argued that validity is not a property of the test but instead is a property of inferences or interpretations we make from test scores. Messick argues that validity is an essential concept for all types of qualitative as well as quantitative summaries.

The authors argue that since "performance assessments promise potential benefits for teaching and learning, it is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal." The primary adverse consequence that should be investigated is the potential negative impact on individuals or groups derived from sources of invalidity such as construct underrepresentation or construct-irrelevant variance. In the former case, individuals may be scoring low because the assessment is missing something that best represents the construct. In the latter case, individuals may score low because the measurement process contains something irrelevant that interferes with the student's ability to demonstrate proficiency.

Generalizability. This chapter provides an overview of generalizability theory and integrates the literature on reliability of performance assessment with the conceptual framework of generalizability theory. Generalizability theory is viewed as the product of the marriage between classical test theory and analysis of variance methodology.

The methodological development in this chapter is quite Comprehensive and easy to understand. It should be required reading for everyone involved in the development of large scale performance assessments. In addition, the author makes several points that represent lessons learned from the literature. Some of these lessons are:

- The number of raters has very little effect on the error variance found in most generalizability studies. This is probably because the scoring rubrics are generally well defined and the raters are well trained. The main conclusion from this is that increasing the number of raters does not increase test reliability.
- The number of tasks is inversely related to the error variance in most generalizability studies. The fewer the tasks the larger the error variance. One approach to reducing error variance might be to narrow the domain of tasks so that each is a slight modification of the others. The other approach is to increase the number of tasks (say to 10 or more) to reduce the error. Although this may be a costly alternative, it often makes more sense than restricting the domain.

Comparability. In order to provide indicators of trends in academic achievement, large scale performance assessments must be comparable across administrations. With multiple-choice testing this is a relatively easy matter because testing conditions are highly standardized, and large numbers of unidimensional items are scored by computer. Performance assessments on the other hand tend to be less standardized, involve fewer tasks, are more multidimensional, and are scored by humans, not computers. These differences make it more difficult to ensure comparability from one administration to the next.

The authors make the point that the degree of comparability required in performance assessments depends on the kind of decision being made and the importance of the consequences attached to those decisions. As the stakes get higher, the requirements for comparability get higher. The chapter concludes with the observation that strict forms of comparability, such as equating and calibration, may not be possible with many large scale performance assessments. However, weaker forms of comparability, such as statistical and social moderation, are attainable. Finally, the authors observe that many of the problems leading to lack of comparability would be mitigated if more precise content specifications for performance exercises were available.

Performance Standards. This chapter deals with the setting of performance standards on large scale performance assessments. There is much in the literature on setting performance standards on multiple-choice-based large assessments but there is very little on setting standards for large scale performance-based assessments. The authors cite four generalizations from two decades of research on performance standards.

- In almost all cases performance standards are arbitrarily (although not capriciously) set on a performance continuum.
- Performance standards are method dependent.

- Those who set performance standards can't objectively evaluate the quality of their standards.
- Widely used performance standing setting methods presume an underlying unidimensional scale.

The last generalization is particularly out of sync with almost all performance assessments that are often explicitly multidimensional. In addition, the authors argue that new methods of setting performance standards are needed because most performance assessments are often based on only a few tasks, each potentially requiring a separate performance standard.

The chapter describes two new approaches to setting performance standards that do not make the above unidimensionality assumptions. Each of the approaches have been evaluated within the context of the National Board for Professional Teaching Standards (NBPTS). The two new approaches are:

- **Iterative, Judgmental Policy Capturing Procedure**: In this method panelists respond independently to graphic profiles of performance for hypothetical students. The panelists make judgements as to whether the overall performance (or profile) should be considered deficient (1), competent (2), accomplished (3), or highly accomplished (4). Various analytic model fitting methods were used to assign weights to each dimension of the profile. The essence of the Iterative, Judgmental Policy Capturing procedure is that the panelists standard setting policies are inferred from their reactions to the profiles of candidate performance presented to them.
- **Multi-Stage Dominant Profile Procedure:** In this method a variety of integrative procedures are used to get the panelists to formulate explicitly their standard setting policies. This involves more up front group discussion and reflection. The procedure is different from the previous one in that the panelists' standards are generated directly through discussion rather than inferred from panelist ratings.

The chapter concludes with several unresolved technical issues that need to be addressed in setting performance standards.

- Since performance standards usually involve an artificial dichotomization of a performance criteria, how do you minimize misclassification near the cut-score?
- Since performance standards are method dependent, how do you assess this source of error in your procedures?
- How large should the standard setting panel be?
- Who should compose the standard-setting panel (e.g., experts or stakeholder groups)?
- In order to ground criterion-referenced performance standards in reality, how do you incorporate the use of normative information?
- How much training should standard setting panelists receive?
- How do you report the sources of error and any adjustments of the standard setting recommendations?

Fairness, Equity, and Bias. The authors define fairness as essentially the same thing as differential validity. However, they go beyond the narrow psychometric concern of differential validity (bias) to include concerns about the educational and social policy that forms the context

for the assessment (equity). It is possible for an assessment to be considered unbiased in a technical sense, yet be used in the service of a policy that does not promote equity.

The point is made that bias can creep into an assessment at various stages throughout the development, data collection and scoring of an assessment. In order to minimize bias, test developers need to: (1) make sure there is diversity among the developers of the content framework, test administrators and the scoring panels; (2) require a sound sensitivity review on all assessment materials for sexist, ethnically insensitive, or stereotypic assessment stimuli; and (3) conduct statistical differential item functioning studies on all items or performance tasks.

The authors conclude their chapter by discussing the various ways high-stakes large scale assessments may have unintended negative consequences for poor and minority students. The major culprit is that when the stakes are high, educators tend to focus resources on what is tested. This often leads to a narrowing of the curriculum but rise in test scores. Unfortunately, a rise in test scores does not necessarily mean an improvement in the overall quality of education for the general population. When the stakes are high, educators will do the same thing with performance-type assessments that they used to do with multiple-choice testing. For example, they might exclude more students with disabilities or limited English proficiency so they will not count in the aggregate summaries, target instruction to students near the cut-scores and ignore those at the bottom and top, or encourage low achieving students to drop out. Such practices tend to corrupt performance assessments as an indicator and disproportionately impact poor and minority students.

Acknowledgments

This report would not have been possible without the contributions of a great many people. Arnold A. Goldstein provided many helpful comments on the text and organization of the report, and coordinated its production. Maureen Treaty was instrumental in initiating the report and coordinating between the various authors.

Several people provided careful and insightful reviews of the text. We would like to thank Pat de Vito (Rhode Island Department of Education), Emerson J. Elliott (previous Commissioner of Education Statistics), Thomas Fisher (Florida Department of Education), Wayne Martin (Colorado Department of Education), Andrew Porter (University of Wisconsin), and Lynn Winters (University of California at Los Angeles), as well as Susan Ahmed, Robert Burton, Andrew Kolstad, Mary Naifeh, and Mary Rollefson, all members of the National Center for Education Statistics staff.

The final report was prepared by the Media and Information Services Publications Staff. Simone Miranda provided careful editing, Donna DiToto typeset the document, and Phil Carr designed the cover and the overall look of the report.

Contents

Validity of Performance Assessments

Samuel Messick
Educational Testing Service

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the *persons* responding as well as the *context* of the assessment. In particular, what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question. This is the main reason that validity is an evolving property and validation a continuing process.

Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean. In other words, validation is basically a matter of constructing a network of evidence supporting (or challenging) the intended purpose of the testing. This chapter addresses the forms of evidence that this network should reasonably encompass and highlights the need for persuasive arguments or rationales whenever pertinent evidence is foregone.

Introductory Groundwork

The principles of validity apply not just to interpretive and action inferences derived from test scores as ordinarily conceived, but also to inferences based on any means of observing or documenting consistent behaviors or attributes. Thus, the term "score" is used generically here in its broadest sense to mean any coding or summarization of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment device such as work samples, portfolios, and realistic problem simulations.

The Value of Validity

This general usage subsumes qualitative as well as quantitative summaries. It applies, for example, to behavior protocols, to clinical appraisals, to computerized verbal score reports, and to behavioral or performance judgments or ratings. Hence, the principles of validity apply to all assessments. These include performance assessments which, because of their promise of positive consequences for teaching and learning, are becoming increasingly popular as purported instruments of standards-based education reform. Indeed, it is precisely because of these politically salient potential consequences that the validity of performance assessment needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability,

and fairness. As applied to performance assessment and standard-setting, these issues taken together constitute the main concern of the present report.

These issues are critical for performance assessment because validity, reliability, comparability, and fairness are not just measurement principles, they are *social values* that have meaning and force outside of measurement whenever evaluative judgments and decisions are made. As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion (i.e., classical criterion validity) or by expert judgments that test content is relevant to the proposed test use (i.e., traditional content validity).

Indeed, broadly speaking, validity is nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Fundamentally, then, score validation is empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use.

Conceptions of Performance Assessment

We next attempt to clarify the meaning of performance assessment, because different conceptions have distinctly different implications for validation. In essence, a performance assessment requires the student to execute a task or process and bring it to completion (Wiggins, 1993). That is, the student performs, creates, or produces something over a sufficient duration of time to permit evaluation of either the process or the product, or both. This is in contradistinction to the impoverished trace or storable record resulting when one merely marks a correct or preferred option on an answer sheet as in a multiple-choice test, which does not reflect the amount or kind of thinking or effort that may underlie the choice of option. The choice of an answer may reflect recognition or recall, to be sure, but also a worked-through solution or guessing.

Indeed, with respect to task processing, the boundary between multiple-choice (MC) tests and performance assessments is a fuzzy one because some students on many MC items and most students on difficult MC items execute the solution process as a means of selecting the appropriate option (Traub, 1993). A more critical distinction is that the selected option can only be appraised for correctness or goodness with respect to a single criterion. There is no record, as in the typical performance assessment, of an extended process or product that can be scored for multiple aspects of quality.

A further complication is that the contrast between MC items and open-ended performance tasks is not a dichotomy, but a continuum representing different degrees of structure versus openness in the allowable responses. This continuum is variously described as ranging from multiple-choice to student-constructed products or presentations (Bennett, 1993), for example, or from multiple-choice to demonstrations and portfolios (Snow, 1993). Successive intervening stages include items requiring reordering or rearranging, substitution or correction, simple

completion or cloze procedures, short essays or complex completions, problem exercises or proofs, teach-back procedures, and long essays.

Apart from multiple-choice, the remainder of the continuum is referred to as involving "student-constructed responses." However, not all student-constructed responses-notably those involving rearranging, substitution, and simple completion-e properly considered to be performance assessments because they do not yield a storable record of an extended process or product.

Prototypical performance assessments occur more toward the unstructured end of the response continuum and include such exemplars as portfolios of student work over time, exhibits or displays of knowledge and skill, open-ended tasks with no single correct approach or answer, and hands-on experimentation. The openness with respect to response possibilities enables students to exhibit skills that are difficult to tap within the predefined structures of multiple-choice, such as shaping or restructuring a problem, defining and operationalizing variables, manipulating conditions, and developing alternative problem approaches.

Evaluations of student achievement on such open-ended tasks usually rely on the professional judgment of the assessor, and some proponents view such subjectivity of scoring to be the hallmark of performance assessment (e.g., Frederiksen & Collins, 1989; Stiggins, 1991). However, this view appears too restrictive because some performance tasks can be objectively scored and some scoring judgments are amenable to expert-system computer algorithms (e.g., Bejar, 1991; Sebrechts, Bennett, & Rock, 1991).

A more likely hallmark of educational performance assessments is their nearly universal focus on higher-order thinking and problem-solving skills. According to Baker, O'Neil, and Linn (1993), "virtually all proponents of performance-based assessment intend it to measure aspects of higher-order thinking processes" (p. 1211). Indeed, performance assessments in education frequently attempt to tap the complex structuring of multiple skills and knowledge, including basic as well as higher-order skills, embedded in realistic or otherwise rich problem contexts that require extended or demanding forms of reasoning and judgment. In this regard, Wiggins (1993) views "authentic" performance assessments as tapping understanding or the application of good judgment in adapting knowledge to fashion performances effectively and creatively.

This mention of authentic assessments broaches a further distinction. Just as performance assessments are a more open-ended subset of student-constructed responses, so-called authentic assessments are a more realistic subset of performance assessments. In particular, authentic assessments pose engaging and worthy problems (usually involving multistage tasks) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world. The assessment challenge of complex performance tasks in general and authentic tasks in particular revolves around issues of scoring, interpretation, and generalizable import of key aspects of the complex performance, especially if the task is not completed successfully.

In performance assessment, one might start by clarifying the nature of the higher-order competencies or other constructs to be assessed and then select or construct tasks that would optimally reveal them. Or, contrariwise, one might start with an important task that is worthy of mastery in its own right and ask what competencies or other constructs this task reveals. This

contrast embodies a tension in performance assessment between construct-centered and task-centered approaches (Messick, 1994). However, what is critical in performance assessment is not what is operative in the task performance but what is captured in the test score and interpretation. Hence, the validity of the construct interpretation needs to be addressed sooner or later in either approach, as does the nature of convergent and discriminant evidence needed to sustain that validity.

Construct-Driven Versus Task-Driven Performance Assessment

The task-centered approach to performance assessment begins by identifying a worthy task and then determining what constructs can be scored and how. Often the mastery of such a worthy task functions as the target of the assessment in its own right, as opposed to serving as a vehicle for the assessment of knowledge, skills, or other constructs. This might occur, for example, in an arts contest or an Olympic figure-skating competition or a science fair. In such cases, replicability and generalizability are not at issue. All that counts is the quality of the performance or product submitted for evaluation, and the validation focus is on the judgment of quality. But note that in this usage of performance assessment as target, inferences are not to be made about the competencies or other attributes of the performers, that is, inferences from observed behavior to constructs such as knowledge and skill underlying that behavior.

Large-scale educational projects such as dissertations are often treated as targets in this manner, by crediting the complex accomplishment as meeting established standards with no requirement of predictiveness or domain generalizability (Baker et al., 1993). However, action implications of such complex assessments usually presume, with little or no specific evidence, that there is a global prediction of future success, that the knowledge and skills exhibited in the assessment will enable the student to accomplish a range of similar or related tasks in broader settings.

In contrast, such presumptions should be confronted by empirical evidence in the performance assessment of competencies or other constructs—that is, where the performance is the vehicle not the target of assessment. A major form of this evidence bears on generalizability and transfer which, as we shall see, represent critical aspects of construct validity. In effect, the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to.

The task-centered approach to performance assessment is in danger of tailoring scoring criteria and rubrics to properties of the task and of representing any educed constructs in task-dependent ways that might limit generalizability. In contrast, the nature of the constructs in the construct-centered approach guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. Focussing on constructs also alerts one to the possibility of construct-irrelevant variance that might distort either the task performance or its scoring, or both (Messick, 1994). The task-centered approach is not completely devoid of constructs, of course, because task selection is often influenced by implicit construct notions or informal theories of learning and performance. The key issue is the extent to which the constructs guide scoring and interpretation and are explicitly linked to evidence supporting that interpretation as well as discounting plausible rival interpretations.

Sources of Invalidity

Construct-irrelevant variance is one of the two major threats to validity, the other being construct underrepresentation. A fundamental feature of construct validity is *construct representation*, whereby one attempts to identify through cognitive-process analysis the theoretical mechanisms underlying task performance, primarily by decomposing the task into requisite component processes and assembling them into a functional model (Embretson, 1983). Relying heavily on the cognitive psychology of information processing, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge (including metacognitive or self-knowledge) that are implicated in task performance.

In the threat to validity known as "construct underrepresentation," the assessment is too narrow and fails to include important dimensions or facets of the construct. In the threat to validity known as "construct-irrelevant variance," the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct. Both threats are operative in all assessment. Hence a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance.

There are two basic kinds of construct-irrelevant variance. In the language of ability and achievement testing, these might be called "construct-irrelevant difficulty" and "construct-irrelevant easiness." In the former, aspects of the task that are extraneous to the focal construct make the task irrelevantly difficult for some individuals or groups. An example is the intrusion of undue reading-comprehension requirements in a test of subject-matter knowledge. In general, construct-irrelevant difficulty leads to construct scores that are invalidly low for those individuals adversely affected (e.g., knowledge scores of poor readers or of examinees with limited English proficiency).

In contrast, construct-irrelevant easiness occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately in ways irrelevant to the construct being assessed. Another instance occurs when the specific test material is highly familiar to some respondents, as when the text of a reading comprehension passage is well-known to some readers or the musical score for a sight-reading exercise invokes a well-drilled rendition from some performers. Construct-irrelevant easiness leads to scores that are invalidly high for the affected individuals as reflections of the construct under scrutiny.

The concept of construct-irrelevant variance is important in all educational and psychological measurement, including performance assessments. This is especially true of richly contextualized assessments and authentic simulations of real-world tasks. This is the case because, "paradoxically, the complexity of context is made manageable by contextual clues" (Wiggins, 1993, p. 208). And it matters whether the contextual clues that are responded to are construct-relevant or represent construct-irrelevant difficulty or easiness.

However, what constitutes construct-irrelevant variance is a tricky and contentious issue (Messick, 1994). This is especially true of performance assessments, which typically invoke constructs that are higher-order and complex in the sense of subsuming or organizing multiple processes. For example, skill in communicating mathematical ideas might well be considered irrelevant variance in the assessment of mathematical knowledge (although not necessarily vice versa). But both communication skill and mathematical knowledge are considered relevant parts

of the higher-order construct of mathematical power according to the content standards developed by the National Council of Teachers of Mathematics. It all depends on how compelling the evidence and arguments are that the particular source of variance is a relevant part of the focal construct as opposed to affording a plausible rival hypothesis to account for the observed performance regularities and relationships with other variables.

Authenticity and Directness As Validity Standards

Two terms that appear frequently, and usually in tandem, in the literature of performance assessment are "authentic" and "direct" assessment. They are most often used in connection with assessments involving realistic simulations or criterion samples. If authenticity and directness are important to consider when evaluating the consequences of assessment for student achievement, they constitute tacit validity standards, so we need to address what the labels "authentic" and "direct" might mean in validity terms.

The major measurement concern of authenticity is that nothing important has been left out of the assessment of the focal construct. This is tantamount to the familiar validity standard of minimal construct underrepresentation (Messick, 1994). However, although authenticity implies minimal construct underrepresentation, the obverse does not hold. This is the case because minimal construct underrepresentation does not necessarily imply the close simulation of real-world problems and resources typically associated with authenticity in the current literature on performance assessment. In any event, convergent and discriminant evidence is needed to appraise the extent to which the ostensibly authentic tasks represent (or underrepresent) the constructs they are interpreted to assess.

The major measurement concern of directness is that nothing irrelevant has been added that distorts or interferes with construct assessment. This is tantamount to the familiar validity standard of minimal construct-irrelevant variance (Messick, 1994). Incidentally, the term "direct assessment" is a misnomer because it always promises too much. In education and psychology, "all measurements are indirect in one sense or another" (Guilford, 1936, p. 3). Measurement always involves, even if only tacitly, intervening processes of judgment, comparison, or inference. The key issue, then, is not directness per se but the minimizing of construct-irrelevant variance in performance assessment scores.

Aspects of Construct Validity

The validity issues of score meaning, relevance, utility, and social consequences are many faceted and intertwined. They are difficult if not impossible to disentangle empirically, which is why validity has come to be viewed as a unified concept (APA, 1985; Messick, 1989). For example, social consequences provide evidence contributing to score meaning, and utility is both validity evidence and a value consequence. The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the integrative power derives from empirically grounded score interpretation.

However, to speak of validity as a unified concept does not imply that validity cannot be usefully differentiated conceptually into distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied use. The intent of these distinctions is to

provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences.

In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects conjointly function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989). However, these six aspects must not be viewed as separate and substitutable validity types-as the erstwhile trinity of content, criterion, and construct validities often were—but rather as interdependent and complementary forms of validity evidence. As general validity criteria, they can be specialized for apt application to performance assessments—as discussed selectively, for example, by Linn, Baker, and Dunbar (1991) and by Moss (1992)—but none should be ignored.

A brief characterization of these six aspects is presented next, followed by six sections discussing the validity issues and sources of evidence bearing on each aspect:

- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989).
- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
- The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957).
- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test-criterion relationships (Hunter, Schmidt, & Jackson, 1982).
- The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965).
- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989).

Content Relevance and Representativeness

A key issue for the content aspect of construct validity is the specification of the boundaries of the construct domain to be assessed-that is, determining the knowledge, skills, and other attributes to be revealed by the assessment tasks. The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and domain theory (Messick, 1989). If concern is with the application of domain processes in

real-world settings, the techniques of job and task analysis should prove useful both in determining domain structure and in selecting relevant realistic assessment tasks. If concern is with the learning of domain processes, analyses of curricula and instruction should prove useful for determining the construct domain and for selecting assessment tasks attuned to the level of developing expertise of the learners. Such considerations are especially germane to authentic assessment because they bear on the issue of "authentic to what?"

Both job and curriculum analysis contribute to the development of domain theory, as does scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed.

It is also important to make explicit the relationship between the construct domain and the assessment specifications, by formulating what amounts to a test blueprint indicating whether the assessment is to include all components of the construct domain or only part of them. This is important because score inferences should be limited to what can be sustained by the assessment and not casually generalized to a broader construct domain.

Moreover, the description of the construct domain, as well as the assessment specifications, should distinguish them from other similar or related construct domains. Ideally, the descriptions should be clear enough so that test developers or other experts can judge whether a task refers to one construct domain or the other. In any event, as we shall see in the sections on the substantive and external aspects of construct validity, discriminant evidence needs to be produced showing that the focal construct is operative in task performance as opposed to similar or related constructs, that is, evidence to discount plausible rival interpretations.

However, it is not sufficient to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to insure that all important parts of the construct domain are covered (or at least those subsets included in the assessment specifications). This is usually described as selecting tasks that sample domain processes in terms of their functional importance, or what Brunswik (1956) called ecological sampling. Functional importance can be considered in terms of what people actually do in the performance domain, as in job analyses, but also in terms of what characterizes and differentiates expertise in the domain, which would usually emphasize different tasks and processes.

Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity. However, as we shall see in the next section on the substantive aspect, such expert judgment is not sufficient because it is not just domain content that needs to be represented in assessment tasks but domain processes.

In standards-based education reform, two types of assessment standards have been distinguished. One type is called "content standards," which refers to the kinds of things a student should know and be able to do in a subject area. The other type is called "performance standards," which refers to the level of competence a student should attain at key stages of

developing expertise in the knowledge and skills specified by the content standards. Performance standards also circumscribe, either explicitly or tacitly, the form or forms of performance that are appropriate to be evaluated against the standards.

From the discussion thus far, it should be clear that not only the assessment tasks but also the content standards themselves should be relevant and representative of the construct domain. That is, the content standards should be consistent with domain theory and be reflective of the structure of the construct domain. This is the issue of the construct validity of content standards. There is also a related issue of the construct validity of performance standards. That is, increasing achievement levels or performance standards (as well as the tasks that benchmark these levels) should reflect increases in complexity of the construct under scrutiny and not increasing sources of construct-irrelevant difficulty (Messick, 1996). More extensive coverage of these and other issues related to standards-based assessment will appear in the subsequent chapter on standard-setting.

Substantive Theories, Process Models, and Process Engagement

The substantive aspect of construct validity emphasizes the role of substantive theories and process modeling in identifying the domain processes to be revealed in assessment tasks (Embretson, 1983; Messick, 1989). Two important points are involved: One is the need for tasks providing appropriate sampling of domain processes in addition to traditional coverage of domain content; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance.

Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Loevinger, 1957). Such evidence may derive from a variety of sources, for example, from "think-aloud" protocols or eye-movement records during task performance, from correlation patterns among part scores, from consistencies in response times for task segments, or from mathematical or computer modeling of task processes (Messick, 1989, pp. 53-55; Snow & Lohman, 1989). In sum, the issue of domain coverage refers not just to the content representativeness of the construct measure but also to the process representation of the construct and the degree to which these processes are reflected in construct measurement.

Another issue is the extent to which the assessed task processes correspond to domain processes, as opposed to being distorted by sources of irrelevant method variance. This depends on whether the assessment tasks mimic or simulate the domain conditions with sufficient comprehensiveness and fidelity to engage the domain processes with minimal distortion, which is a primary aim of authentic assessment. The point here is that empirical evidence is needed to assure that the higher-order thinking processes that authentic assessments aspire to address are actually operative in task performance. For example, for some individuals the task performance might not reflect problem solving, but rather a memorized solution. As another instance, a verbal reasoning task might be failed by some respondents because of inadequate verbal knowledge rather than poor inductive inference. The test user is in the best position to evaluate the meaning of individual scores under the specific applied circumstances, that is, to appraise the extent to which the intended score meaning might have been eroded by contaminating variables operating locally.

The core concept bridging the content and substantive aspects of construct validity is representativeness. This becomes clear once one recognizes that the term "representative" has two distinct meanings, both of which are applicable to performance assessment. One is in the cognitive psychologist's sense of representation or modeling (Suppes, Pavel, & Falmagne, 1994); the other is in the Brunswikian sense of ecological sampling (Brunswik, 1956; Snow, 1974). The choice of tasks or contexts in assessment is a representative sampling issue, which is central to the content aspect of construct validity. The comprehensiveness and fidelity of simulating the construct's realistic engagement in performance is a representation issue, which is central to the substantive aspect. Both issues are important in performance assessment; they are critical to the very meaning of authentic assessment.

Scoring Models As Reflective of Task and Domain Structure

According to the structural aspect of construct validity, scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). That is, the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria and rubrics.

Ideally, the manner in which behavioral instances are combined to produce a score should rest on knowledge of how the processes underlying those behaviors combine dynamically to produce effects. Thus, the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989). This property of construct-based rational scoring models is called "structural fidelity" (Loevinger, 1957).

To the extent that different assessments (i.e., those involving different tasks or different settings or both) are geared to the same construct domain, using the same scoring model as well as scoring criteria and rubrics, then the resultant scores are likely to be comparable or can be rendered comparable using equating procedures. To the degree that the different assessments do not adhere to the same specifications, then score comparability is jeopardized but can be variously approximated using calibration, projection, and moderation procedures (Mislevy, 1992).

Score comparability is clearly important for normative or accountability purposes whenever individuals or groups are being ranked. However, score comparability is also important even when individuals are not being directly compared, but are held to a common standard. Score comparability of some type is needed to sustain the claim that two individual performances in some sense meet the same local, regional, national, or international standard. These issues are addressed more fully in the subsequent chapter on comparability.

Generalizability and the Boundaries of Score Meaning

The concern that a performance assessment should provide representative coverage of the content and processes of the construct domain is meant to insure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. For example, how

well one can generalize from a sample of writing about a particular topic in a particular genre to skill in writing about other topics in the same or different genre depends on the pattern of correlations among different topic and genre scores. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. This conflict is usually addressed by means of a variety of trade-offs. For example, one suggestion is to increase the number of performance assessments for each student or to increase the number of tasks in each assessment. Here the trade-off is between breadth of coverage and nonassessment instructional activities that might instead have filled the extended testing time.

Another suggestion is to use a matrix-sampling design with different performance tasks administered to different samples of students. Here the gain in breadth of coverage comes at the expense of individual student scores or, at least, of comparable individual scores. Nonetheless, matrix sampling is especially useful when the accountability concern focuses on some aggregate level of inference such as the school, district, state, or nation. Another approach is to develop assessments that represent a mix of efficient structured exercises broadly tapping multiple aspects of the construct and time-intensive open-ended tasks tapping integral aspects in depth (Messick, 1994), which involves a trade-off between the number of performance tasks and the number of brief structured exercises. The internal structure of interrelations among the briefer exercises and performance tasks bears on the substantive and especially the structural aspect of construct validity. Such structures undergird and guide decisions as to how responses should be aggregated into composite or multiple scores to represent the construct.

This conflict between depth and breadth of coverage is often viewed as entailing a trade-off between validity and reliability (or generalizability). It might better be depicted as a trade-off between the valid description of the specifics of a complex task and the power of construct interpretation. In any event, as Wiggins (1993) stipulates, such a conflict signals a design problem that needs to be carefully negotiated in performance assessment.

In addition to generalizability across tasks, the limits of score meaning are also affected by the degree of generalizability across time or occasions and across observers or raters of the task performance. Such sources of measurement error associated with the sampling of tasks, occasions, and scorers underlie traditional reliability concerns; they are examined in more detail in the subsequent chapter on generalizability.

Convergent and Discriminant Correlations with External Variables

The external aspect of construct validity refers to the extent to which the assessment scores' relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations.

The external aspect emphasizes two intertwined sets of relationships for the assessment scores: one between the task scores and different methods for measuring both the same and distinct constructs, and the other between measures of the focal construct and exemplars of different constructs predicted to be variously related to it on theoretical grounds. Theoretically, expected empirical consistencies in the first set include both convergent and discriminant correlation patterns, the convergent pattern indicating a correspondence between measures of the same construct and the discriminant pattern indicating a distinctness from measures of other constructs. These patterns are often displayed in what is called a multitrait-multimethod matrix (Campbell & Fiske, 1959).

Convergent evidence signifies that the measure in question is coherently related to other measures of the same construct as well as to other variables that it should relate to on theoretical grounds. Discriminant evidence signifies that the measure is not unduly related to exemplars of other distinct constructs. Discriminant evidence is particularly critical for discounting plausible rival alternatives to the focal construct interpretation. Both convergent and discriminant evidence are basic to construct validation.

Theoretically, expected consistencies in the second set of relationships mentioned above indicate a lawful relatedness between measures of different constructs. This lawful relatedness has been referred to as "nomological validity" by Campbell (1960) and as "nomothetic span" by Embretson (1983). The basic notion of nomological validity is that the theory of the construct being addressed provides a rational basis for deriving empirically testable links between the assessment task scores and measures of other constructs. Corroborative evidence helps to validate both the assessment and the construct theory. The assessment gains credence to the extent that the score correlations reflect theoretical implications of the construct, while the construct theory gains credence to the extent that score data jibe with its predictions.

Among the relationships falling within the purview of nomological validity or nomothetic span are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.

The issue of utility is evaluated in terms of the benefits or desired outcomes of the assessment relative to its costs (Cronbach & Gleser, 1965; Messick, 1989). Thus, although the cost of performance assessments in terms of time and resources is an important consideration, the choice among alternative assessment approaches should not be determined solely by cost or efficiency. Rather, such decisions should weigh both the costs and the benefits of the assessment, that is, its utility for the applied purpose.

Consequences as Validity Evidence

Because performance assessments promise potential benefits for teaching and learning, it is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal. In this connection, the consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term. Particularly prominent is the

evaluation of any adverse consequences for individuals and groups, especially gender and racial/ethnic groups, that are associated with bias in scoring and interpretation or with unfairness in test use. However, this form of evidence should not be viewed in isolation as a separate type of validity, say, of "consequential validity." Rather, because the values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1964, 1975, 1980).

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989). That is, low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected students to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected students' demonstration of competence.

However, reducing adverse impact associated with sources of test invalidity does not mean that there would necessarily be less adverse impact associated with the valid description of existing group differences. For example, a possible unintended consequence of performance assessment in education is increased adverse impact for gender and racial/ethnic groups because of short-term misalignments in their educational experiences vis-à-vis authentic testing and teaching. If found, one should monitor the situation to see how short-term it is likely to be and what resources are needed to redress the new imbalance. Positive and negative consequences of assessment, whether intended or unintended, are discussed in more depth in the subsequent chapter on fairness.

Aspects of Validity Specialized for Performance Assessment

Some proponents of performance assessment have proposed specialized validity criteria tailored for performance tasks (Frederiksen & Collins, 1989; Linn et al, 1991). In effect, these specialized criteria emphasize selected issues in some but not all of the six general validity aspects just described as they are applied to performance tasks (Messick, 1994; Moss, 1992). However, a few of these specialized criteria highlight different perspectives that warrant further comment. Especially important because it is at the heart of authentic assessment in education is what Linn and his colleagues (1991) call "meaningfulness" and what Frederiksen and Collins (1989) call "transparency." The concern here is that if the assessment itself is to be a worthwhile educational experience serving to motivate and direct learning, then the problems and tasks posed should be meaningful to the students and communicate clearly what is expected of them. That is, not only should students know what knowledge and skills are being assessed, but the criteria and standards of good performance should be clear to them, in terms of both how the performance is to be scored and what steps might be taken to improve performance. In this sense, the criteria and standards of successful performance are transparent or demystified and hence should be more readily internalized by students as self-directive goals (Baron, 1991; Wiggins, 1993).

Evidence needs to be accrued, of course, that the performance tasks are meaningful and that the performance standards are understood and facilitate learning, because the meaningfulness or transparency of performance assessments cannot be taken for granted. Such evidence is also

pertinent to the substantive and consequential aspects of construct validity. Moreover, there are important instances where transparency may be counterproductive, namely, where novelty occurs in the task performance that is not amenable to the transparent standards of goodness. Indeed, the very salience of the transparent standards might hamper the generation of novelty. In such cases, the challenge is to transform the standards or to develop new ones. That is, transparency and creativity may be in conflict.

A concept closely related to meaningfulness is what Frederiksen and Collins (1989) call "systemic validity." Their point is that as instruments of education reform, performance tests should "induce curricular and instructional changes in educational systems (and learning strategy changes in students) that foster the development of the cognitive traits that the tests are designed to measure" (p. 27). However, because interpretation of such teaching and learning consequences as reflective of test validity (or invalidity) assumes that all other aspects of the educational system are working well or are controlled, the use of the label "systemic *validity*" *is* problematic. It might better be called "systemic facilitation" because in practice the issue is not just the systemic validity of the tests but, rather, the validity of the system as a whole for improving teaching and learning. In any event, the concept of systemic validity is a specialized instance of the consequential aspect of construct validity because it focuses on one type of testing consequence—indeed, on one type of systemic consequence-among many (Messick, 1989, p. 85).

Validity As Integrative Summary

These six aspects of construct validity apply to all educational and psychological measurement, including performance assessments. Taken together, along with aspects of validity specialized for performance assessments such as meaningfulness or transparency, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. They are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly. That is, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and are generalizable across a range of tasks, settings, and occasions. Furthermore, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of presumed or documented relationships with nontest behaviors and anticipated outcomes or consequences.

The challenge in test validation is to link these inferences to convergent evidence supporting them as well as to discriminant evidence discounting plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else provide compelling reasons why not, which is what is meant by validity as a unified concept.

Overview

The traditional conception of validity divided it into three separate and substitutable types—namely, content, criterion, and construct validities. This view is fragmented and

incomplete, especially in failing to take into account evidence of the value implications of score meaning as a basis for action and of the social consequences of score use. The new unified concept of validity interrelates these issues as fundamental aspects of a more comprehensive theory of construct validity addressing both score meaning and social values in both test interpretation and test use. That is, unified validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and theoretically relevant relationships, including those of both an applied and a scientific nature. Six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. These six aspects are not separate and substitutable validity types, as in the traditional validity conception, but rather are interdependent and complementary forms of evidence in the unified view of validity. In effect, these six aspects together function as general validity criteria or standards for all educational and psychological measurement, including performance assessments.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210-1218.

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

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*,522-532.

Bennett, R. E. (1993). On the meaning of constructed response. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*,546-553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies*. San Francisco: Sage.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694 (Monograph Supplement 9).

Messick, S. (1964). Personality measurement and college performance. *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110-129). Princeton, NJ: Educational Testing Service. (Reprinted in A. Anastasi (Ed.). (1966). *Testing problems in perspective* (pp. 557-572). Washington, DC: American Council on Education.)

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

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

Messick, S. (1996). Standards-based score interpretation: Establishing valid grounds for valid inferences. *Proceedings of the joint conference on standard setting for large-scale assessments*, Sponsored by National Assessment Governing Board and The National Center for Education Statistics. Washington, DC: Government Printing Office.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243-299). Hinsdale, IL: Dryden Press.

Sebrechts, M. M., Bennett, R. E., & Rock D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856-862.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-396.

Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research, 44*, 265-291.

Snow, R. E. (1993). Construct validity and constructed response tests. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*, 263-273.

Suppes, P., Pavel, M., & Falmagne, J-C. (1994). Representations and models in psychology. *Annual Review of Psychology, 45*, 517-544.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Erlbaum.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*, 200-214.

Generalizability of Performance Assessments

Robert L. Brennan
University of Iowa

Historically, in psychology and education, reliability issues have been addressed principally using classical test theory, which postulates that an observed score can be decomposed into a "true" score and a single, undifferentiated random error term, E (see Feldt & Brennan, 1989). Generalizability theory liberalizes classical theory by providing models and methods that allow an investigator to disentangle multiple sources of error that contribute to E. This is accomplished in part through the application of certain analysis of variance (ANOVA) methods.

The purposes of this chapter are: (a) to provide an overview of those aspects of the conceptual frame work and methods of generalizability theory that are particularly relevant to performance assessments; (b) to integrate the current literature on the reliability of performance assessments into the framework of generalizability theory; and (c) to aid researchers and practitioners in the assessment of the generalizability of performance assessments. Reliability/generalizability coefficients are considered as well as error variances and standard errors of measurement. In addition, the generalizability of both individual scores and group mean scores is discussed.

Generalizability Theory: Basic Concepts

In a sense, classical test theory and ANOVA can be viewed as the parents of generalizability theory. However, generalizability theory has a unique conceptual framework. Among the concepts in this framework are *universes of admissible observations* and G (Generalizability) *studies*, as well as *universes of generalization* and D (Decision) *studies*.

An extensive, in depth explication of the concepts and methods of generalizability theory is provided by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Brennan (1992a) provides a somewhat less detailed treatment. Overviews of *essential* features of generalizability theory are provided by Feldt and Brennan (1989), and Shavelson and Webb (1991). An introduction is provided by Brennan (1992b). Brennan and Johnson (1995) use generalizability theory to treat some of the issues covered in

this paper. Recently, in their consideration of generalizability analysis for educational assessments, Cronbach, Linn, Brennan, and Haertel (1995) have covered topics that partly overlap those treated in this paper.

In this section, the concepts and methods of generalizability theory are briefly explained and illustrated using an example from the performance testing literature reported by Shavelson, Baxter, and Gao (1993), who state that:

> The California Assessment Program (CAP) conducted a voluntary statewide science assessment in 1989-1990. .. Students were posed five independent tasks. More specifically, students rotated through a series of five self-contained stations at timed intervals (about 15 mins.). At one station, students were asked to complete a problem solving task (determine which of these materials may serve as a conductor). At the next station, students were asked to develop a classification system for leaves and then to explain any adjustments necessary to include a new mystery leaf in the system. At yet another, students were asked to conduct tests with rocks and then use the results to determine the identity of an unknown rock. At the fourth station, students were asked to estimate and measure various characteristics of water (e.g., temperature, volume). And at the fifth station, students were asked to conduct a series of tests on samples of lake water to discover why fish are dying (e.g., is the water too acidic?). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table).
>
> A predetermined scoring rubric developed by teams of teachers in California was used to evaluate the quality of students' written responses (California State Department of Education, 1990) to each of the tasks. Each rubric was used to score performance on a scale from O to 4 (0 = no attempt, 1 = serious flaws, 2 = satisfactory, 3 = competent, 4 = outstanding). All tasks were scored by three raters. (p. 222)

Universe of Admissible Observations and G Study Considerations

For the CAP example, the universe of admissible observations (UAO) consists of two facets: tasks (t) and raters (r). Since, in principle, any task could be evaluated by any rater, these facets are crossed in the UAO, and this crossing is denoted $t \times r$. Persons (p) or students are not viewed as part of the UAO. Rather, they constitute the population.

As reported by Shavelson et al. (1993), the G study design for the CAP example consisted of taking a sample of five tasks from the UAO, administering them to a sample of

persons, and then having three raters evaluate all products/results produced by all persons. This is a verbal description of a fully crossed G Study $p \times t \times r$ design.

Strictly speaking, for the CAP example, the G study is a random effects G study because the authors assumed that the potential set of tasks and raters in the UAO were both indefinitely large, with the actual tasks and raters in the G study viewed as samples from the UAO.

G study variance components. The principal results of a G study are estimated variance components for each of the effects in a G study design. These estimates are obtained using analysis of variance procedures (see Brennan, 1992a, for details). For the CAP example, the estimated variance components are reported in the second column of Table 1. For example, the estimated variance component for persons is $\hat{\sigma}_p^2 = .298$, and the estimated variance component for the interaction of persons and tasks is $\hat{\sigma}_{pt}^2 = .493$.

The variance component for persons can be interpreted in the following manner. Suppose an investigator could obtain each person's mean (or expected value) over all tasks and raters in the UAO. The variance of these scores would be σ_p^2, which is estimated to be $\hat{\sigma}_p^2 = .298$ for the CAP data. Similarly, $\hat{\sigma}_r^2 = .003$ is the estimated variance of rater mean scores, where each mean (or expected value) is over all persons in the population and all tasks in the UAO. The estimated variance of task mean scores in the UAO is $\hat{\sigma}_t^2 = .092$, which suggests that tasks differ somewhat in difficulty.

Interaction variance components are somewhat more difficult to interpret. Consider, for example, $\hat{\sigma}_{pt}^2$ in the CAP example. The fact that $\hat{\sigma}_{pt}^2 = .493$ is considerably greater than zero suggests that there is a considerably different rank ordering of person mean scores for each of the various tasks in the UAO. By contrast, the fact that $\hat{\sigma}_{pr}^2 = 0$ means that the various raters rank order persons similarly. Also, $\hat{\sigma}_{rt}^2 = .002$ suggests that the various raters

Table 1

CAP Generalizability Analyses[a]

	G Study $\hat{\sigma}^2$	D Study Estimated Variance Components		
			$n'_t = 5$, $n'_r = 3$	$n'_t = 10$, $n'_r = 1$
Persons (p)	0.298	$\hat{\sigma}^2_p$	0.298	0.298
Tasks (t)	0.092	$\hat{\sigma}^2_T = \hat{\sigma}^2_t / n'_t$	0.018	0.009
Raters (r)	0.003	$\hat{\sigma}^2_R = \hat{\sigma}^2_r / n'_r$	0.001	0.003
pt	0.493	$\hat{\sigma}^2_{pT} = \hat{\sigma}^2_{pt} / n'_t$	0.099	0.049
pr	0.000	$\hat{\sigma}^2_{pR} = \hat{\sigma}^2_{pr} / n'_r$	0.000	0.000
tr	0.002	$\hat{\sigma}^2_{TR} = \hat{\sigma}^2_{tr} / n'_t n'_r$	0.000	0.000
ptr,e	0.148	$\hat{\sigma}^2_{pTR,e} = \hat{\sigma}^2_{ptr} / n'_t n'_r$	0.010	0.015
		$\hat{\sigma}^2_{\tau} = \hat{\sigma}^2_p$ =	0.30	0.30
		$\hat{\sigma}^2_{\delta} = \hat{\sigma}^2_{pT} + \hat{\sigma}^2_{pR} + \hat{\sigma}^2_{pTR,e}$ =	0.11	0.06
		$\hat{\sigma}^2_{\Delta} = \hat{\sigma}^2_{\delta} + \hat{\sigma}^2_T + \hat{\sigma}^2_R + \hat{\sigma}^2_{TR}$ =	0.13	0.08
		$\hat{\rho}^2 = \hat{\sigma}^2_p / \left[\hat{\sigma}^2_p + \hat{\sigma}^2_{\delta}\right]$ =	0.73	0.82
		$\hat{\Phi} = \hat{\sigma}^2_p / \left[\hat{\sigma}^2_p + \hat{\sigma}^2_{\Delta}\right]$ =	0.70	0.80

[a] G study variance components were provided by Xiaohong Gao.

rank order the difficulty of the tasks similarly. The last variance component, $\hat{\sigma}^2_{ptr,e} = .148$, is a residual variance component that includes the triple-order interaction and all other unexplained sources of variation.

The fact that $\hat{\sigma}^2_r, \hat{\sigma}^2_{pr}$, and $\hat{\sigma}^2_{rt}$ are all close to zero suggests that the rater facet does not contribute much to variability in observed scores. By contrast, $\hat{\sigma}^2_t$ and especially $\hat{\sigma}^2_{pt}$ are quite large suggesting that the task facet contributes greatly to score variability.

The G study variance components provide a decomposition of the variance over p, t, and r of <u>single</u> person-task-rater scores:

$$\sigma^2_{X_{ptr}} = \sigma^2_p + \sigma^2_t + \sigma^2_r + \sigma^2_{pt} + \sigma^2_{pr} + \sigma^2_{tr} + \sigma^2_{ptr,e}, \quad (1)$$

which is usually called "total variance" in the generalizability theory literature, because it is analogous to "total" sums of squares in analysis of variance (see Cronbach et al., 1972). That is, in generalizability theory, the phrase "total variance" refers to the sum of the G study variance components. From the second column of Table 1 it is evident that the largest contributors to total variance are persons and person-task interactions.

<u>Other examples</u>. The CAP assessment results are typical, in a sense, of generalizability results for many programs that involve performance assessments. For Example, Figure 1 reports the percent of total variance accounted for by each of the seven variance components in the $p \times t \times r$ design for CAP and five other performance assessment programs. In examining Figure 1 the reader is cautioned not to attach undue importance to the numerical values for the percents, which are for single person-task-rater scores. In particular, these percents should not be interpreted as percents for average (or total) scores for which decisions might be made. For purposes of this paper, what is important is that there are similarities in the profiles of the percents for the various studies. (Note that the magnitudes of the actual variance components across studies would not be comparable because, among other things, the studies involve different scoring metrics.)

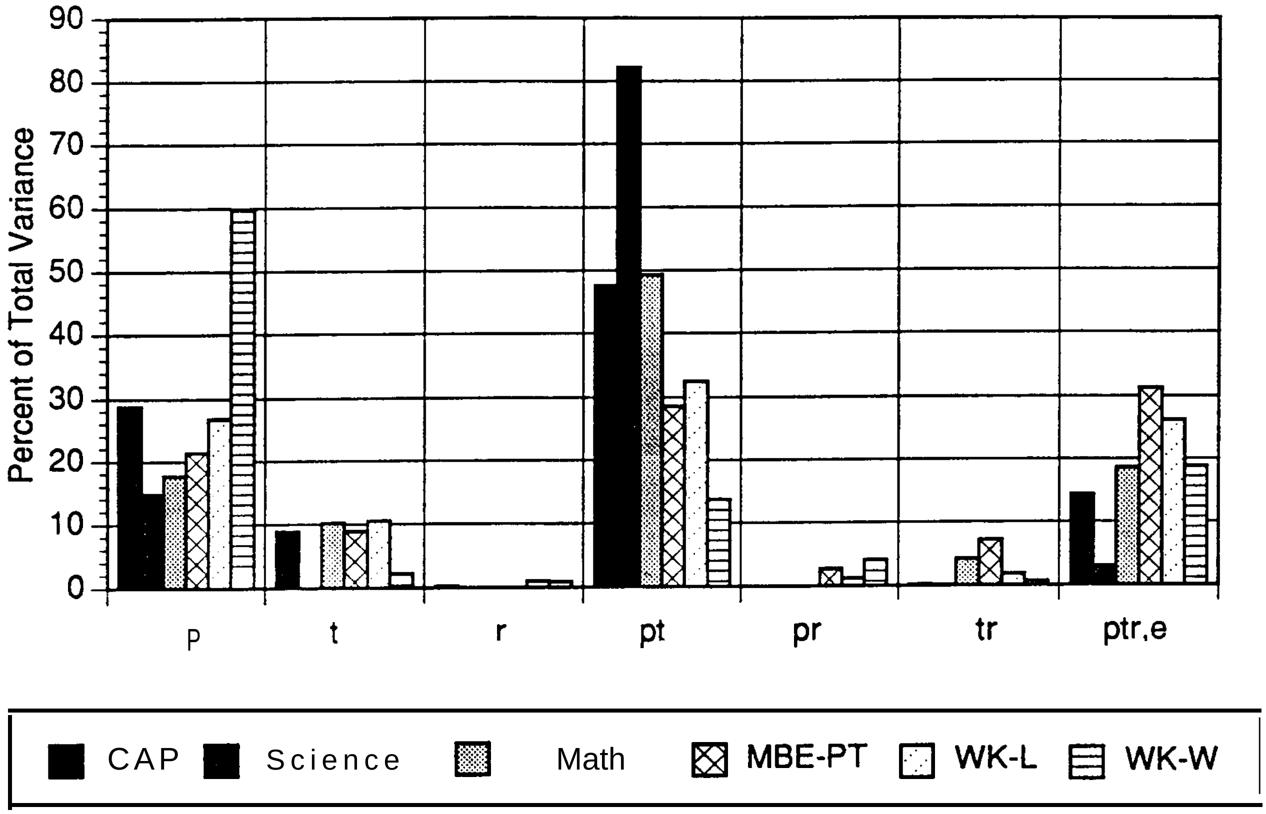


Figure 1. Percents of total variance for variance components for the p x t x r design for six different assessments.

The science and math assessments are discussed by Shavelson et al. (1993) in the same article in which CAP is treated. The MBE-PT results are from a generalizability analysis of performance tasks for possible inclusion in the Multistate Bar Examination (MBE) used by many states to admit candidates to the practice of law (see Gamache & Brennan, 1994). The WK-L and WK-W are for Listening (L) and Writing (W) tests for a new program called Work Keys (WK) being developed by American College Testing (see Brennan, Gao, & Colton, 1995).

For all programs the rater facet contributes relatively little to total variance for single person-task-rater observations, as evidenced by the fact that $\hat{\sigma}_r^2$, $\hat{\sigma}_{pr}^2$, and $\hat{\sigma}_{tr}^2$ are all relatively small. It is also evident that for all programs $\hat{\sigma}_t^2$ is small suggesting that tasks are quite similar in average difficulty.

Perhaps most importantly, however, it is clear from Figure 1 that $\hat{\sigma}_{pt}^2$ is larger than $\hat{\sigma}_p^2$ for all assessments except WK-W and, indeed, $\hat{\sigma}_{pt}^2$ is the largest of all the variance components for CAP, Science, Math, and WK-L. The relatively large magnitude of $\hat{\sigma}_{pt}^2$ suggests that there is only a limited degree of across-task generalizability. A similar conclusion has been reported for other programs by Baker (1992), Dunbar, Koretz, & Hoover (1991), Lane, Stone, Ankenmann, and Liu (1992), Linn (1993), Linn and Burton (1994), van der Vleuton and Swanson (1990), and Welch (1991) among others.[1]

For the $p \times t \times r$ design considered thus far, each rater evaluates all tasks performed by all persons. Sometimes, of course, such a design is not logistically or administratively feasible. A common alternative is a design in which different raters evaluate each task. For such a design, raters are nested within task, the design is denoted $p \times (r{:}t)$, and there are five variance components: σ_p^2, $,$.1, σ_{pt}^2, and $\sigma_{pr:t,e}^2$. For example, this design is employed

operationally in the writing assessment component of the EXPLORE Program (ACT, 1994). For EXPLORE, the average over forms of the estimated variance components is: $\hat{\sigma}^2_p = .47$, $\hat{\sigma}^2_t = .05$, $\hat{\sigma}^2_{r:t} = .00$, $\hat{\sigma}^2_{pt} = .52$, and $\hat{\sigma}^2_{pr:t,e} = .27$. Again $\hat{\sigma}^2_{pt}$ is the largest variance component, and the rater facet contributes little to total variance.

Infinite Universe of Generalization and D Study Considerations

Effectively, Equation 1 states that G study variance components provide a decomposition of the total observed score variance for single person-task-rater scores. Let n_p, n_t, and n_r be the G study sample sizes for persons, tasks, and raters, respectively. Then, for the CAP example, the total observed score variance is the variance of the $n_p n_t n_r$ observed scores, each of which could be 0, 1, 2, 3, or 4. In practice, of course, decisions about persons (the objects of measurement) will not likely be made based on persons' scores for a single task evaluated by a single rater. Rather, decisions will be made based on average scores over multiple tasks and/or raters. Indeed, a typical D (Decision) study consideration is to identify values of n'_t and n'_r (which need not equal n_t and n_r, respectively) that result in acceptably small error variance and/or acceptably large reliability-like coefficients.

Another D study consideration is the specification of a universe of generalization (UG), which is the universe to which a decision maker wants to generalize. In this section it will be assumed that the UG mirrors the UAO in the sense that the task and rater facets are both infinite. Strictly speaking, this means that the UG is a universe of randomly parallel forms of the measurement procedure, where each such form consists of a different sample of tasks and a different sample of raters. (In a subsequent section, a restricted UG is considered that is smaller than the infinite UG.) A person's universe score is his or her mean (or expected) score over all randomly parallel forms of the measurement procedure in the UG. As such, universe score is analogous to true score in classical theory.

An additional D study consideration is the design structure of the D study. In this section, it will be assumed that the D study design mirrors the G study design in the sense

that both are fully crossed, meaning that all persons respond to the same tasks, and the responses/products of all persons to all tasks are evaluated by the same raters.

D study variance components. G study variance components are used to estimate D study variance components for average scores over n'_t tasks and n'_r raters. The process is very simple. Let σ^2_α be the G study variance component for α (e.g., if σ^2_{pt}, $a = pt$), and let $\sigma^2_{\alpha'}$ be the corresponding D study variance component. If α contains t, then $\sigma^2_{\alpha'} = \sigma^2_\alpha / n'_t$; if α contains r, then $\sigma^2_{\alpha'} = \sigma^2_\alpha / n'_r$; and if α contains both t and r, then $\sigma^2_{\alpha'} = \sigma^2_\alpha / n'_t n'_r$. The resulting equations for the estimated D study variance components are provided in the third column of Table 1. Note that D study variance components are designated using upper-case subscripts for *R* and *T* to emphasize that these variance components are for mean scores over n'_r raters and n'_t tasks.

The D study variance component σ^2_p is the variance of persons' universe scores for the infinite UG. As such, it is called universe score variance, which is analogous to true score variance in classical theory.

D study variance components are used to estimate relative and absolute error variances, as well as two reliability-like coefficients called generalizability coefficients and dependability coefficients.

Absolute error variance. Absolute error is simply the difference between a person's observed and universe score. The variance over persons of absolute errors is:

$$\sigma^2_\Delta = \sigma^2_T + \sigma^2_R + \sigma^2_{pT} + \sigma^2_{pR} + \sigma^2_{TR} + \sigma^2_{pTR,e}$$

$$= \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_r}{n'_r} + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{tr}}{n'_t n'_r} + \frac{\sigma^2_{ptr,e}}{n'_t n'_r}. \quad (2)$$

That is, absolute error variance is the sum of all the D study variance components except for universe score variance. For the CAP data with $n'_t = 5$ and $n'_r = 3$, Table 1 reports that $\hat{\sigma}^2_\Delta = .13$. The square root is $\hat{\sigma}_\Delta = .36$, which is the A - type, or absolute, standard error of measurement (SEM). Consequently, adding and subtracting .36 to persons' observed scores

over five tasks and three raters provides approximate 68% confidence intervals for persons' universe scores.

Relative error variance. Relative error is the difference between a person's observed deviation score and his or her universe deviation score. The variance over persons of relative errors is:

$$\sigma^2_\delta = \sigma^2_{pT} + \sigma^2_{pR} + \sigma^2_{pTR,e}$$

$$= \frac{\sigma^2_{pt}}{n'_t} - \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{ptr,e}}{n'_t n'_r} \qquad (3)$$

That is, relative error variance is the sum of the D study variance components that include the index p and at least one other index. The square root of relative error variance is analogous to the SEM in classical test theory. For the CAP data with $n'_t = 5$ and $n'_r = 3$, relative error variance is $\hat{\sigma}^2_\delta = .11$, and the square root is $\hat{\sigma}_\delta = .33$, which is the δ-type, or relative, SEM. Note that $\sigma^2_\delta \le \sigma^2_\Delta$ because σ^2_δ does not contain σ^2_T, σ^2_R , or σ^2_{TR}.

Generalizability coefficient. A generalizability coefficient is defined as:

$$\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\delta}, \qquad (4)$$

where σ^2_τ is a generic notation for universe score variance. That is, a generalizability coefficient is the ratio of universe score variance to itself plus relative error variance. As such, a generalizability coefficient is analogous to a reliability coefficient in classical theory. For the case considered in this section, $\sigma^2_\tau = \sigma^2_p$, and σ^2_δ is given by Equation 3. Therefore,

$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{ptr,e}}{n'_t n'_r}}. \qquad (5)$$

For the CAP data with $n_t' = 5$ and $n_r' = 3, \hat{\rho}^2 = .73$, as indicated in Table 1.

Dependability coefficient. A dependability coefficient is defined as:

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\Delta^2}. \tag{6}$$

That is, a dependability coefficient is the ratio of universe score variance to itself plus absolute error variance. For the case considered in this section, $\sigma_\tau^2 = \sigma_p^2$, and σ_Δ^2 is given by Equation 2 Therefore,

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_t^2}{n_t'} + \frac{\sigma_r^2}{n_r'} + \frac{\sigma_{pt}^2}{n_t'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{tr}^2}{n_t'n_r'} + \frac{\sigma_{ptr,e}^2}{n_t'n_r'}}. \tag{7}$$

The only difference between ρ^2 and@ is that ρ^2 uses σ_δ^2 as error variance, whereas Φ uses σ_Δ^2 as error variance. Since $\sigma_\Delta^2 \geq \sigma_\delta^2$, it follows that $\Phi \leq \rho^2$. From Table 1, for $n_t' = 5$ and $n_r' = 3, \hat{\Phi} = .70$ which is less than $\hat{\rho}^2 = .73$.

Consequences of different sample sizes. Figure 2 provides $\hat{\rho}^2, \hat{\Phi}, \hat{\sigma}_\delta$, and $\hat{\sigma}_\Delta$ for the CAP example with n_t' ranging from 1 to 12 and n_r' ranging from 1 to 3. These results might be employed to examine the consequences of using various numbers of tasks and raters.

Perhaps the most striking result in Figure 2 is that the number of raters has very little influence on the magnitude of SEM's and coefficients. This is a direct result of the previously noted fact that variance components involving the rater facet are quite small, presumably because the scoring rubrics are well defined and the raters are well trained.

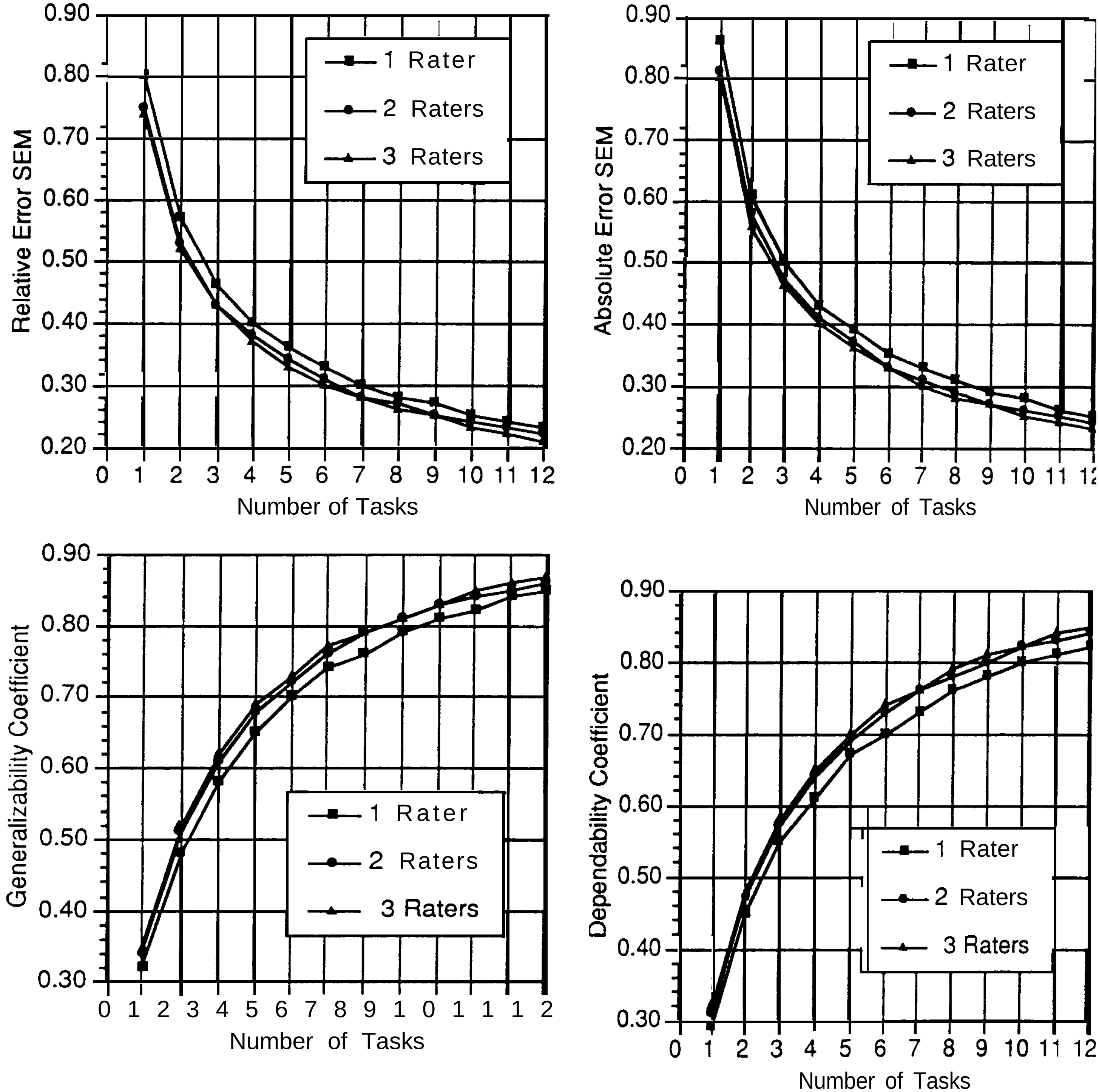


Figure 2. Generalizability Results for CAP

Consequently, there seems to be no compelling psychometric reason to employ more than one rater.

Another obvious fact from Figure 2 is that increasing the number of tasks has a dramatic effect on lowering SEM's and increasing the values of coefficients. With one rater, nine tasks are required for $\hat{\rho}^2 \geq .80$, and ten tasks are needed for $\hat{\Phi} \geq .80$ (see also the last column of Table 1). Developing, administering, and scoring that many tasks obviously would not be a trivial undertaking.

On the other hand, in some circumstances, an investigator might argue that it is not too sensible to choose n'_t on the basis of the resulting magnitude of $\hat{\rho}^2$ or $\hat{\Phi}$, because both depend on universe score variance. Such an investigator might be satisfied with a somewhat low value for a coefficient provided individual persons were measured accurately enough. If so, the investigator would be more interested in the magnitude of SEM's than coefficients. To consider this possibility, recall that CAP scores for a single task range from O to 4, which means that average scores over n'_t tasks have the same range, but with fractional scores frequently occurring. Assuming normally distributed observed scores, if the investigator wanted to be 95% certain that persons' observed and universe scores differed by no more than two points, only three tasks are required with one rater. However, to be 95% certain that observed and universe scores differ by no more than one point, 12 tasks are required with one rater, and about 10 tasks are required with three raters![2]

When $n'_r = 1$, Figure 3 provides 68% and 95% Δ-type confidence intervals for the CAP data for examinees with a universe score of $\tau = 2$, assuming normally distributed observed scores about τ. For example, when $n'_t = 3$, there is a 68% probability that examinees with a true score of 2 will obtain observed scores between 1.5 and 2.5.

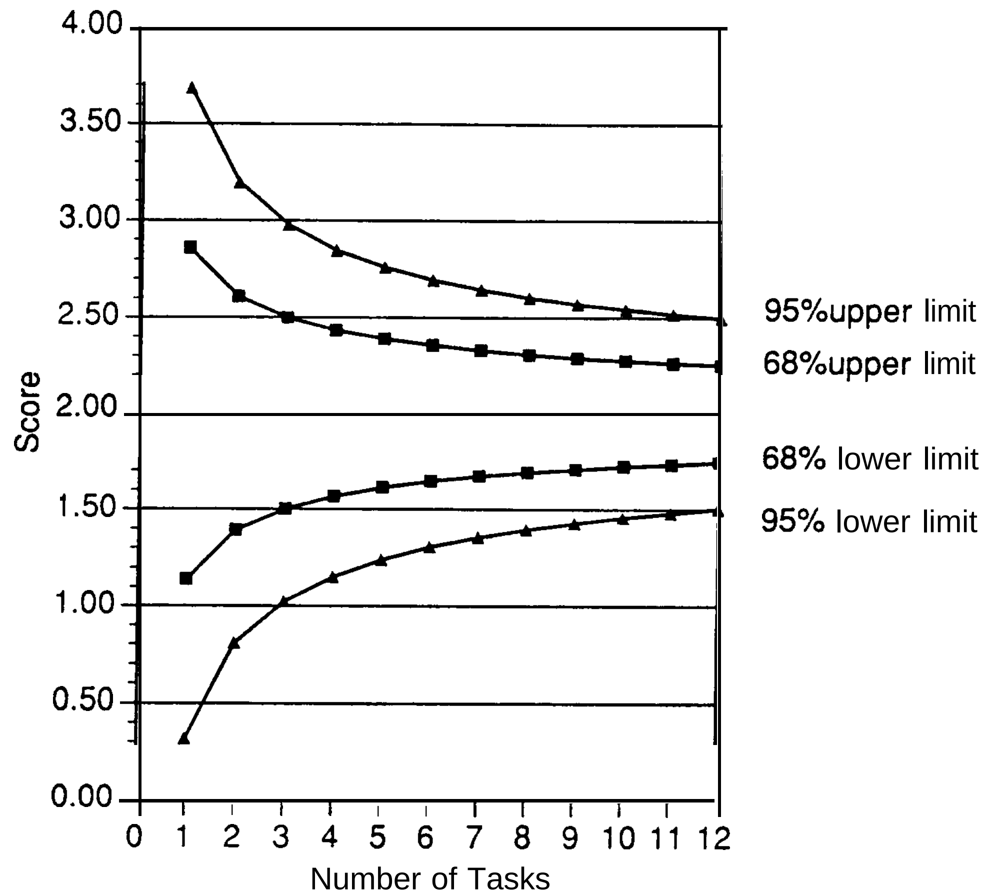


Figure 3. Confidence intervals for CAP data based on absolute error SEM's using a single rater and various numbers of tasks, assuming a universe score of 2.

The limits of the confidence intervals in Figure 3 can also be used as minimum passing and maximum failing scores, as discussed by Linn (1994). For example, from Figure 3, with $n_t' = 8$, a minimum passing score of 2.6 and a maximum failing score of 1.4 are required for 95% confidence that correct decisions will be made when the standard is set at 2.

Contributions of Different Facets

This section gives additional consideration to the influence of rater, task, and other facets on the measurement precision of performance assessments. In doing so, each facet is treated somewhat separately. This is done to isolate issues associated with the various facets, and to facilitate relating traditional results to the types of generalizability results discussed in the previous section. In practice, of course, generalizability analyses involving many facets simultaneously are greatly to be preferred.

Generalizing Over Raters

The examples in the previous section suggest that for current performance assessments the rater facet does not contribute much to variability in observed scores. While these results have been found in a wide range of current performance assessments, it should not be assumed that such results are necessarily inevitable. For example, Linn (1993) reports that the classic studies of Starch and Elliott (1912, 1913) on the grading of high school work in English and mathematics demonstrated an extraordinary range of grades assigned to written essays and extended responses in geometry. Indeed, the Starch and Elliott studies provided considerable support to the increased use of objective testing in the early part of this century. Undoubtedly, lack of trained raters and agreed-upon scoring rubrics contributed to the Starch and Elliott results.

Until recently, for any measurement procedure involving subjective scoring, probably the most frequently discussed measurement issue was inter-rater reliability. Indeed, if inter-rater reliability was high, it was often assumed that there were no other reliability issues of consequence. Although this narrow perspective no longer

predominates, inter-rater reliability is still a very important issue. Indeed, high inter-rater reliability is viewed by most investigators as a necessary, although not sufficient, condition for adopting a performance test.

The phrase "inter-rater reliability" may seem to have a self-evident interpretation. Actually, however, there are at least two general measurement perspectives on inter-rater reliability. One perspective typically involves inter-rater reliability coefficients. The other perspective considers error variances or standard errors. Both perspectives involve observed differences in ratings, but the two perspectives are not isomorphic.

Inter-rater reliability coefficients. In the performance testing literature, two general conclusions about inter-rater reliability seem to predominate. First, when tasks are the same for all students and scoring procedures are well-defined, inter-rater reliability tends to be quite high. Second, when different students respond to different tasks, choose their own tasks (e.g., select their own essay topics), or produce unique products, then inter-rater reliability tends to be relatively low. This appears to be especially true for portfolio assessments (see Gearhart, Herman, Baker, & Whittaker, 1992, and Koretz, Klein, McCaffrey, & Stecher, 1993). Another way to state these conclusions is to say that when tasks are standardized inter-rater reliability tends to be high, and when tasks are not standardized it tends to be low.

These conclusions are to be expected based on the manner in which variance components enter the standardized and non-standardized inter-rater reliability estimates. The standardized estimate is typically obtained by correlating the ratings of two different raters to the responses of a group of persons to the same task.[3] In terms of the variance components introduced in the first section, this correlation is approximately equal to the generalizability coefficient

$$\rho^2 = \frac{\sigma_p^2 + \sigma_{pt}^2}{\sigma_p^2 + \sigma_{pt}^2 + \sigma_{pr}^2 + \sigma_{ptr,e}^2}. \quad (8)$$

The denominator of Equation 8 is identical to the denominator of ρ^2 in Equation 5 when $n'_t = n'_r = 1$. For the case considered in this section, $n'_t = 1$ because only one task is involved in the correlation, and $n'_r = 1$ because a correlation between two raters gives an estimate of reliability for a single rater.

The numerators of the generalizability coefficients in Equations 5 and 8 differ in that Equation 8 includes not only σ^2_p but also σ^2_{pt}. When all persons respond to the same single task, effectively the single task is hidden and fixed for all persons. Consequently, the universe of generalization is a restricted universe in which raters constitute the only random facet. Statistically this leads to σ^2_{pt} being included with σ^2_p in the numerator of Equation 8. Because σ^2_{pt} is usually quite large in performance testing, the numerator of ρ^2 in Equation 8 is likely to be large resulting in a relatively high value of ρ^2.

Effectively, σ^2_{pt} is part of universe score variance in Equation 8, whereas almost always σ^2_{pt} is more properly viewed as part of error variance (see Equations 2 and 3). For this reason (and another reason considered later), almost always inter-rater reliability coefficients for standardized situations are too big relative to more appropriate estimates of generalizability for the reported scores on a performance test (see Equation 5).

The non-standardized estimate of inter-rater reliability is typically obtained by correlating two ratings of a different task or product for each person. In such cases, the design has tasks or products nested within persons and crossed with raters, $[(t{:}p) \times r]$, and the correlation is approximately equal to the generalizability coefficient

$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{t:p} + \sigma^2_{pr} + \sigma^2_{tr:p,e}} \ , \qquad (9)$$

where $\sigma^2_{t:p} = \sigma^2_t + \sigma^2_{pt}$ and $\sigma^2_{tr:p,e} = \sigma^2_{tr} + \sigma^2_{ptr,e}$. The non-standardized coefficient in Equation 9 will be less than the standardized coefficient in Equation 8 for two reasons: (a) the numerator of the non-standardized coefficient is smaller than the numerator of the

standardized coefficient by σ^2_{pt}; and (b) the denominator of the non-standardized coefficient is larger than the denominator of the standardized coefficient by $\sigma^2_t + \sigma^2_{tr}$, which contributes to the relative error variance σ^2_δ for the non-standardized coefficient.

The variance component σ^2_{pt} does not appear in the numerator of the non-standardized coefficient in Equation 9 because tasks vary over persons, and therefore person-task interaction does not contribute to universe score variance. Rather σ^2_{pt} is included in $\sigma^2_{t:p}$ and contributes to error variance. The different role that σ^2_{pt} plays in Equations 8 and 9 is often the principal reason that the non-standardized inter-rater reliability coefficient is smaller than the standardized coefficient. In addition, Equation 9 will tend to be smaller than Equation 8 if tasks are unequal in average difficulty (i.e., σ^2_t is relatively large), and/or raters are differentially capable of rating different tasks (i.e., σ^2_{tr} is relatively large). Either or both of these conditions are probably more likely to characterize portfolio assessments than other types of performance assessments.

The generalizability coefficients in Equations 8 and 9 approximate the inter-rater reliability coefficients most frequently reported in the literature. Note, however, that they are for making decisions based on only <u>one</u> rating of only <u>one</u> task. This is evident from the fact that the variance components in Equations 8 and 9 containing r and/or t *are* all divided by one. Frequently, inter-rater reliability coefficients are incorrectly interpreted as estimates of reliability when persons' scores are the sum or average of two ratings. Such estimates are easily obtained from Equations 8 and 9 by halving variance components that contain r. The resulting estimates of generalizability for two ratings are necessarily larger than those for a single rating.

<u>Standard errors and error variances</u>. In the performance testing literature, issues of rater reliability are most frequently discussed in terms of inter-rater reliability coefficients. Sometimes, however, such issues are treated (and perhaps more appropriately treated) from the perspective of differences in the actual ratings (e.g., a plot of two ratings for each person). It is intuitively clear that such differences reflect error in

some sense. Indeed, it can be shown that, when data are available for two ratings and decisions will be based on the mean of $k' = 1$ or 2 ratings, the A-type standard error for a given person is

$$\hat{\sigma}_{\Delta_p} = |X_{p1} - X_{p2}| / \sqrt{2k'}. \tag{10}$$

This standard error is the absolute value of the difference between the two ratings for the person divided by $\sqrt{2}$ if $k' = 1$, or divided by 2 if $k' = 2$. This equation is appropriate whether or not each person responds to the same task(s) -- i.e., whether or not tasks are standardized. The average over persons of individual error variances is the overall A-type error variance:

$$\hat{\sigma}^2_{\Delta} = \sum_p \hat{\sigma}^2_{\Delta_p} / n_p. \tag{11}$$

If the two ratings are based on the same task, then task is fixed and generalization is over the rater facet, only. If the two ratings are based on different tasks, then generalization is over both tasks and raters. In general, of course, standard errors for generalizing over raters, only, are likely to be smaller than standard errors for generalizing over both tasks and raters. For example, Gamache and Brennan (1994) report that $\hat{\sigma}_{\Delta_p}$ for generalizing over raters, only, was on average about 60% as large as $\hat{\sigma}_{\Delta_p}$ for generalizing over both raters and tasks. Their results are for experimental performance tasks for a law examination.

Adjudication and other issues. In performance testing, it is relatively common practice to obtain a third rating if two ratings differ by more than one rating scale point. There are various ways in which this third rating might be used, but very often the net effect is that the final two ratings for a person differ by no more than one point. Under these circumstances, for $k' = 2$, $\hat{\sigma}_{\Delta_p}$ is either .5 or 0. Furthermore, if p_1 is the

proportion of persons with a difference score of 1, the overall value of $\hat{\sigma}_\Delta$ will be $.5\sqrt{p_1}$, which is likely to be quite small. For example, if $p_1 = .36$ then $\hat{\sigma}_\Delta = .30$ of a rating scale point.

In short, whenever some process is used to adjudicate ratings, standard errors of measurement are likely to be relatively small, and in this sense the ratings will appear quite reliable. Hence, an adjudication process is beneficial only if it does not distort the intended construct being rated. Such distortion could occur, for example, if the adjudicators were systematically influenced by the original ratings, or if the adjudicators were experts untrained in the scoring rubric.

Carefully constructed scoring rubrics, an intensive training session for raters, and an adjudication process usually produce ratings with small error variance. However, small error variance does not guarantee that inter-rater reliability coefficients will be high. This follows from the fact that universe score variance, σ_p^2, is in the numerator of any inter-rater reliability coefficient, but is absent from error variance. Consequently, if σ_p^2 is small relative to error variance, then an inter-rater reliability coefficient could be small even if error variance is small. This "thought experiment" illustrates that an inter-rater reliability coefficient encapsulates information about the magnitude of true differences among person (universe score variance) relative to errors.

If true differences among persons are of no consequence -- as might be the case in a fully criterion-referenced context -- then inter-rater reliability coefficients may be of little value. Real world testing contexts are seldom so clear cut, however. In short, error variances and inter-rater reliability coefficients capture overlapping, but still different, types of information. Often, therefore, it is sensible to report both. In any case, it is always advisable to report estimated variance components (see AERA, APA, NCME, 1985, p.19).

Rater vs. Task Reliability

The use of well-constructed scoring rubrics with well-trained raters can substantially reduce errors attributable to raters. However, virtually all currently available research on performance testing suggests that generalizing over tasks is an error-prone activity, no matter how well the tasks are designed, primarily because σ^2_{pt} tends to be relatively large, and secondarily because tasks tend to be somewhat different in difficulty (σ^2_t is greater than 0).

Dunbar et al. (1991) reviewed a number of studies of direct assessments of performance, primarily in the area of writing. To compare the influence of raters and tasks on reliability they computed average reliability due to raters and what they called "score reliability."[4] In the notation of this chapter, the Dunbar et al. (1991) reliability due to raters is approximated by Equation 8 (provided all persons are rated by the same raters), and

$$\text{score reliability} \cong \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pt}}. \qquad (12)$$

Since Equation 12 involves σ^2_{pt}, in the terminology of this chapter it seems more sensible to refer to this coefficient as "task reliability ." It is equivalent to the generalizability coefficient in Equation 5 for a single task and an infinite number of raters. Another interpretation of Equation 12 is that it is a generalizability coefficient under the assumption that raters are perfectly consistent in all respects, and the only source of error is person-task interactions.

Although comparing Equations 8 and 12 is awkward, these equations do give some sense of the relative influence of raters and tasks on generalizability. Note also that multiplying the right-hand sides of Equations 8 and 12 gives the generalizability coefficient in Equation 5 for a single task and a single rater. (See Kane, 1982, pp. 145-146, for a discussion of a similar result in terms of the reliability-validity paradox.) Table

2, with slight modifications, is from Dunbar et al. (1991). This table summarizes results from the studies they examined.

Several observations are evident from Table 2. Most importantly, for all studies task reliability is relatively small suggesting that person-task interactions are a considerable source of error. It is also evident from Table 2 that there is considerable variability in rater reliability. In particular, rater reliability tends to be lower for the older studies. Commenting on this, Dunbar et al. (1991) state:

> Lindquist (1927), for example, provided very general scoring rubrics for a 10-point holistic scale yielding a mean rater reliability of .33, whereas Hieronymus and Hoover (1987) developed very specific rubrics for a 4-point scale and obtained a coefficient of .91 from ratings made by a group of professional readers. The data in Hildebrand (1991) were collected by the same procedures with the same instrumentation used by Hieronymus and Hoover, but the setting was that of a field experiment rather than a controlled standardization of a set of scoring protocols. This contrast, a controlled standardization versus a field experiment, demonstrates the effect that administrative conditions can have on the reliability of raters. (p. 293).

It appears, then, that often methods can be found to increase rater reliability. However, comparable methods do not exist for increasing task reliability. Of course, tasks should be developed as carefully as possible, and if this is not done then it is likely that reliability will be adversely affected. However, well-constructed tasks do not ensure high reliability, primarily because there is considerable variability in examinee performance on different tasks -- even for tasks in the same domain. That is, σ^2_{pt} tends to be relatively large.

Of course, in principal, σ^2_{pt} can be reduced by narrowing the domain of tasks. So, for example, an investigator could define the tasks in a domain in such a way that each of them is simply a slight modification of the others. Doing so may well decrease σ^2_{pt} and, therefore, increase task reliability. However, restricting the domain of tasks in this way leads to a narrowing of the universe of generalization and, in this sense, a decrease in

Table 2[a]

Reliability Studies of Direct Assessments of Performance

Data source	Measurement context	Average reliability	
		Raters	Tasks
Lindquist (1927)	Assess effectiveness of a laboratory method for instruction in writing for college students	.33	.26
Coffman (1966)	Determine the validity of objective tests for predicting composite essay scores	.39	.26
Swartz, Patience, & Whitney (1985)	Develop an assessment of writing skill for awarding high school equivalency diplomas	.74	.60
Applebee, Langer, & Mullis (1986)	Characterize national trends in writing skill among 9-, 13- & 17-year olds in NAEP	.78	NR
Breland, Camp, Jones, Morris, & Rock (1987)	Evaluate the use of essay tests to predict performance in college-level writing courses	.59	.41
Hieronymus & Hoover (1987)	Develop performance-based measure of writing skill in grades 3 through 8	.91	.46
Hildebrand (1991)	Measure the effects of revision strategies on scores in direct writing assessment	.67	NR
Purves (1992)	Develop writing tasks and scoring protocols for international comparisons of achievement	NR	.42
Welch (1991)	Assess generalizability of essay scores on test for second-year college students	.76	.44[b]

Note. The reliability estimates reported here are simple averages of all the coefficients reported in the original. They are adjusted to reflect an assessment based on 1 reader and 1 sample of performance via the Spearman-Brown formula. NR means "not reported."

[a]From Dunbar, Koretz, and Hoover (1991), with minor modifications.

[b]Computed using Equation 12, which gives a different result from that reported by Dunbar, Koretz, and Hoover (1991).

validity (see Kane, 1982). This is an example of the so-called "reliability-validity paradox." It is generally not advisable to take steps to increase reliability that lead to a decrease in validity.

One of the most important considerations in the development of a performance test is a careful specification of the task facet in the universe of generalization. At a minimum, an investigator should be able to defend the set of tasks in a performance test as a reasonable representation of the domain of tasks that might have been used. Otherwise, there is little basis for claiming that performance on the particular tasks in a performance test can be generalized to a larger universe of tasks. (See Shavelson et al., 1993, p. 216, for an example of a specification of a domain of tasks.)

The importance of accurate specification of a subject matter domain for performance assessments has been illustrated by Shavelson, Gao, and Baxter (1996). For the domain of elementary science they demonstrated that an inappropriately broad specification of the domain leads to overestimating task variability and underestimating generalizability. Conversely, an inappropriately narrow specification of the domain will likely lead to underestimating task variability and overestimating generalizability. Note, however, that generalizability theory per se does not tell an investigator how narrow or wide a domain should be. It is the investigator's responsibility to clearly specify the domain and defend that specification.

<u>Other Facets</u>

Shavelson et al. (1993) provide the following perspective on relevant facets for performance assessments:

> . . . we view a performance assessment as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. We view the task facet to be representative of the content in a subject-matter domain. The occasion facet includes all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment. We view the rater facet as including all possible individuals who could be trained to score performance reliably. These

three facets are, traditionally, thought of as sources of unreliability in a measurement.

In addition, we incorporate a method facet into our definition of the universe of generalization. This formulation moves us beyond reliability into a sampling theory of validity (cf. Kane, 1982). Specifically, we view the method facet to be all possible methods (e.g., short answer, computer simulation) that a decision maker would be equally willing to interpret as bearing on student achievement.

Occasion as a facet. From a classical perspective, sampling variability due to occasions most closely corresponds to the notion of test-retest reliability. Also, variability due to occasions is incorporated in traditional notions of intra-rater reliability, which reflects variability in ratings for the same raters on two occasions. Ideally, from the perspective of minimizing error variance, an investigator would like examinee performance-test products to be minimally changed over occasions during which no instruction occurred. Similarly, an investigator would like ratings for the same raters to be stable over occasions.

For at least two reasons, there are very few studies in the performance testing literature that incorporate more than one occasion. First, doing so is logistically difficult and quite costly. Second, in operational settings collecting data on two occasions is usually not an intended part of the testing process. Even so, it is highly desirable that at least small-scale G studies be conducted that involve occasion as a facet in order to examine the extent to which an investigator can legitimately claim that scores obtained on one occasion are generalizable to scores that might be obtained on different, but similar, occasions.

Ruiz-Primo, Baxter, and Shavelson (1993) and Shavelson et al. (1993) examined the stability of several elementary science performance assessments. Their results suggest that variance attributable to the interaction of persons, tasks, and occasions (σ^2_{pto}) was very large -- indeed, many times larger than universe score variance and also larger than σ^2_{pt}. However, variance attributable to persons and occasions (σ^2_{po}) was quite small. This means that, over an infinite number of tasks, there was little person-occasion interaction in their data, but for any single task persons were rank ordered

differently on different occasions. The obvious remedy for large σ^2_{pto} is to use a large number of tasks and/or occasions in making decisions about examinees, but this may not be feasible in practice.

Another study of the stability of science performance assessments has been reported by Tamir (1974), who found that reliability was on the order of .35 for a design that involved equivalent problems, different raters, and two occasions. Also, Carey (1991) and Mayberry and Hiatt (1990) studied the stability of military job performance, and found retest reliabilities of about .70.

Usually, when occasion is considered as a facet in the performance testing literature, it is associated with the time when examinee performance occurs or products are created. Actually, however, there is a second occasion facet that could influence the generalizability of scores on performance assessments. This facet involves the occasion(s) on which the ratings are obtained. This second occasion facet would be important if judges' rating were not stable over time. For example, Wainer (1993) comments that:

> During the course of the 1988 NAEP writing assessment, some 1984 essays were included in the mix for the purpose of assessing change. The difference in the scores of these essays, from one assessment to the next, was so large that it was deemed wise to determine change through the very expensive restoring of a large sample of 1984 essays by the 1988 judges (Johnson & Zwick, 1988). No mere statistical adjustment was apparently sufficient (p.15).

Actually, this is a very complicated situation because change in occasion is confounded with change in judges (and perhaps subtle changes in rubrics or training procedures).

It would be imprudent to use the results of the studies referenced in the previous paragraphs as a basis for sweeping conclusions about the extent to which scores on performance tasks are invariant over occasions. However, given these results and results with other modes of testing, it certainly would be surprising if scores on performance tasks did not exhibit at least some variability over occasions of testing and/or rating.

Method as a facet. Cronbach et al. (1972), Brennan (1992a), and Kane (1982) have all observed that generalizability theory often blurs distinctions between reliability and validity. This is indeed the case if method or mode of testing (e.g., performance tasks, multiple-choice items, short-answer questions, etc.) is incorporated as a facet in the universe of generalization.

If results are invariant over mode of testing, then there is evidence of convergent validity, and a supportable argument can be made that different modes of testing provide exchangeable information. If results are not invariant over mode of testing, then different modes provide different types of information about student performance.

Shavelson et al. (1993) examined mode of testing for two science performance tasks and four methods (observations of actual performance, notebook reports of steps employed, computer administration of tasks, and short-answer questions). They concluded that not all methods converge, and "certain methods may measure different aspects of achievement (p. 229). " Although this is only one study, it seems plausible that their conclusions will generalize to other settings. This does not mean, however, that one mode is preferable to another. Such a conclusion can be drawn only through a joint consideration of psychometric properties, content, context, logistical, and cost considerations.

Scoring rubrics/procedures. Wainer (1993, p.15) suggests that performance assessments yield acceptable levels of accuracy only when scoring rubrics are rigidly defined. This may be one reason why most empirical analyses of performance assessments effectively consider scoring rubrics as fixed in the sense that only one rubric is used. In principal, however, there may be many rubrics that could be used. This is another example of blurred distinctions between reliability and validity. If two or more rubrics are in principal equally acceptable, then the issue is primarily in the realm of reliability. However, if the acceptable rubrics are not equally preferable, then the matter

is largely one of validity. For example, the "ideal" rubric may be so costly to implement that a simpler rubric is adopted for operational use.

Put in the language of generalizability theory, the issue is the extent to which an investigator can generalize over rubrics, or the extent to which examinee scores are in some sense invariant over rubrics. If scores vary depending on the rubric, then it is not likely that scores can be interpreted meaningfully without a clear understanding of the specific rubric employed. This is one reason why the interpretation of performance test scores is often more demanding than for traditional modes of testing. Decision makers must understand not only what is being tested but also the standards and procedures used to assign scores.

Generalizability of Group Mean Scores

In many contexts performance assessments are used to make decisions about groups (e.g., classes, schools, or districts) rather than individuals. In such cases, the objects of measurement are groups rather than persons, and the scores of interest are group mean scores (see Kane & Brennan, 1977).

Complete Desire

Suppose pupils (p) are nested within groups (g), and all pupils respond to the same tasks (t). This is a description of the $(p{:}g) \times t$ design. Of course, the pupils' products must be scored by raters, but the rater facet will be suppressed in this section to simplify discussion. As noted previously, the literature suggests that doing so is not too problematic if raters are well trained to use carefully constructed rubrics.

For the $(p{:}g) \times t$ design there are five variance components: σ^2_g, $\sigma^2_{p:g}$, σ^2_t, σ^2_{gt}, and $\sigma^2_{pt:g,e}$. Since the objects of measurement are groups, the facets in the universe are pupils and tasks, both of which are assumed to be random, here. Under these circumstances, universe score variance is

$$\sigma^2_\tau = \sigma^2_g. \qquad (13)$$

That is, universe score variance is variance attributable to group means.

Letting $n'_{p:g}$ be the D study number of pupils within each group, and letting n'_t be the D study number of tasks, relative error variance is '

$$\sigma^2_\delta = \frac{\sigma^2_{p:g}}{n'_{p:g}} + \frac{\sigma^2_{gt}}{n'_t} + \frac{\sigma^2_{pt:g,e}}{n'_{p:g}n'_t}, \tag{14}$$

and absolute error variance is

$$\sigma^2_\Delta = \sigma^2_\delta + \frac{\sigma^2_t}{n'_t}. \tag{15}$$

Note, in particular, that when group means are the objects of measurement, variability due to persons is part of both relative and absolute error variance. Consequently, both error variances decrease not only when the number of tasks increases, but also when there is an increase in the number of pupils within each group.

Using Equations 13 and 14 a generalizability coefficient is still given by Equation 4, and using Equations 13 and 15 a dependability coefficient is still given by Equation 6.

Gao, Brennan, and Shavelson (1994) report the following estimated variance components for the CAP science performance assessments described in the first section of this chapter:

$$\hat{\sigma}^2_g = .09,\ \hat{\sigma}^2_{p:g} = .21,\ \hat{\sigma}^2_t = .06,\ \hat{\sigma}^2_{gt} = .07 \text{ and } \hat{\sigma}^2_{pt:g,e} = .52\,. \tag{16}$$

The G study that led to these estimates was based on a random sample of 15 students from each of 40 California public schools.

The pupil-by-task variance component is the largest, by far, which is consistent with other research cited earlier. (In this case, however, note that pupil-by-task variance is confounded with residual error.) The second largest variance component is for pupils

within schools, which contributes to error variance. Large variation among pupils within schools leads to uncertainty about school universe scores. Note, in particular, that $\hat{\sigma}^2_{p:g} = .21$ is much larger than universe score variance among schools, $\hat{\sigma}^2_g = .09$.

Variation attributable to tasks is relatively small suggesting that tasks differ somewhat in difficulty. Variation attributable to school-by-task interaction is of the same magnitude suggesting that task difficulty varies somewhat by school.

Figure 4 provides estimates of the absolute SEM and the generalizability coefficient for 1 to 12 tasks and 20, 40, and 60 pupils within a school. From Figure 4 it is clear that both n'_t and $n'_{p:g}$ are influential in determining $\hat{\sigma}_\Delta$ and $\hat{\rho}^2$. Also, there are trade-offs between increasing $n'_{p:g}$ and n'_t. For example, when $n'_{p:g} = 20$, doubling the number of tasks from 6 to 12 leads to about as much improvement in the generalizability coefficient as doubling the number of persons from 20 to 40 when $n'_t = 6$. Note, also, that it appears that measurement precision does not improve much by using more than about 50 persons within a school.

Percent passing intervals. Another use that can be made of the variance components for a $(p:g) \times t$ design is to estimate confidence internals for the percent of examinees from a school who will exceed a particular score. A procedure for doing so has been described by Linn and Burton (1994, pp. 7-8).[5] Using the Gao et al. (1994) data, Figure 5 provides the Linn-Burton approximate 68% confidence intervals for a school with a universe score of 2.25, given a passing score of 2. Under these circumstances, the true pass rate is about 71%, but observed pass rates may exhibit great variability y, even for relativel y large values of $n'_{p:g}$ and n'_t.

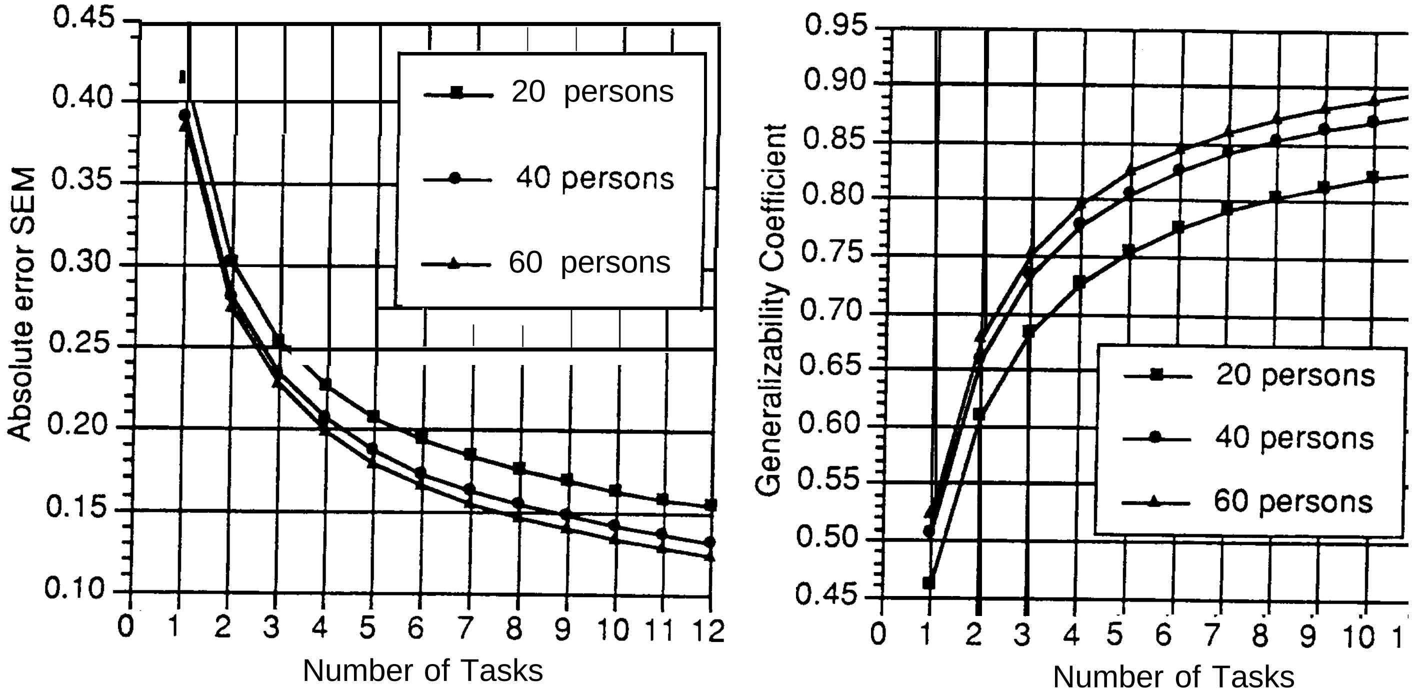


Figure 4. Results for CAP data from Gao, Brennan, and Shavelson (1994).

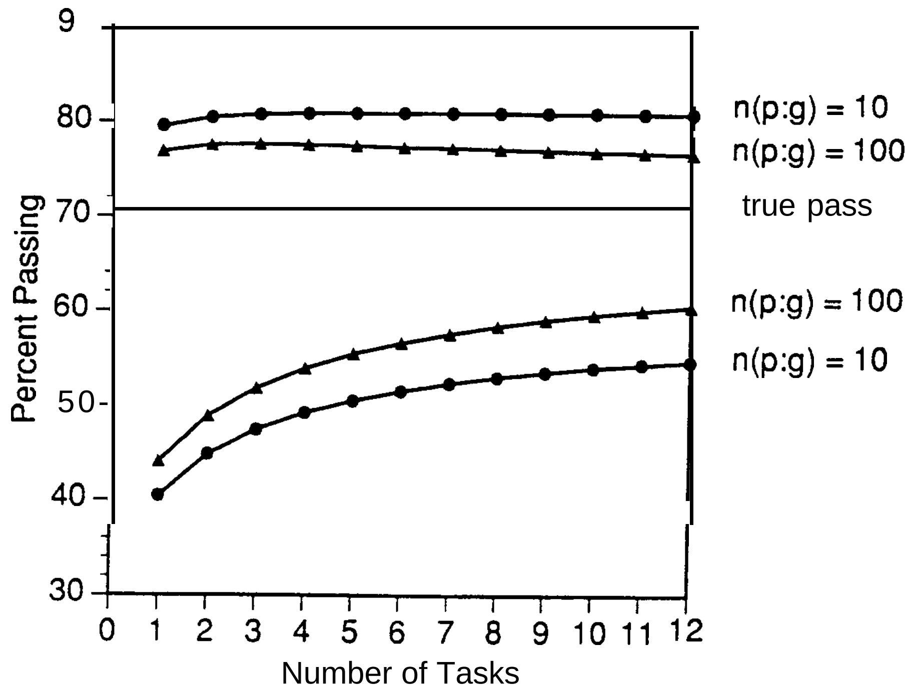


Figure 5. 68% confidence intervals for percent of examinees who will pass in a school with a universe score of 2.25 assuming a passing score of 2. Results are based on application of the Linn-Burton (1994) procedure using variance components in Gao, Brennan, and Shavelson (1994).

For example, when $n_t' = 5$, a 68% confidence interval for percent passing covers a 23 percentage point range from 55% to 78%. This means that, over replications (e.g., different times of testing) 68% of the time, the observed pass rate (given a passing score of 2) for a school with a universe score of 2.25 would likely range from 55% to 78%. Not only is this a very broad range, but also the interval is not symmetric about the true passing rate of 71%. For these data, it appears that very large numbers of tasks and persons would be required to obtain narrow intervals.

Violations of "conventional wisdom." Not infrequently, investigators assume that reliability for groups is necessarily greater than reliability for persons, and/or error variance for groups is necessarily less than error variance for persons. Using generalizability theory, Brennan (1995) has shown that this "conventional wisdom" is not necessarily true. In particular, violations of this conventional wisdom with respect to generalizability coefficients are quite likely, especially when variability due to persons within groups is relatively large, and/or the number of persons within groups is relatively small.

Brennan (1995) shows that, when pupils (over groups) are the objects of measurement, a generalizability coefficient is

$$\rho^2 = \frac{\sigma_g^2 + \sigma_{p:g}^2}{\sigma_g^2 + \sigma_{p:g}^2 + \dfrac{\sigma_{gt}^2}{n_t'} + \dfrac{\sigma_{pt:g,e}^2}{n_t'}}. \qquad (17)$$

In this case, universe score variance involves not only variability due to groups but also variability due to persons within groups.

Consider, again, the Gao et al. (1994) estimated variance components in Equation 16. Using these estimates in Equation 17 with $n_t' = 5$ gives $\hat{\rho}^2 = .72$. Using Equations 13, 14, and 4, it can be shown that when schools are the objects of measurement $\hat{\rho}^2 < .72$ if $n_{p:g}' < 15$. In other words, when $n_t' = 5$ and $n_{p:g}' < 15$, school reliability is less than pupil reliability, which violates the conventional wisdom.

Suppose $n'_t = 5$ and $n'_{p:g} = 10$. In this case, as noted above, school reliability is less than pupil reliability. However, school error variance (both relative and absolute) is less than pupil error variance, in accordance with the conventional wisdom. This paradox is easily resolved by noting that when scores are aggregated both universe score variance and error variance are likely to decrease, but not necessarily at the same rate. Hence, for at least some purposes it can be misleading to consider either generalizability coefficients or error variances in isolation. This is an important consideration given the current trend towards reporting aggregated scores on performance assessments.

Matrix-Sampling Designs

Historically, matrix sampling designs have been used primarily to estimate moments (especially the mean) for a distribution of examinee scores. More recently, such designs have been used in NAEP to estimate entire distributions of examinee scores.

Gao, Shavelson, and Baxter (in press) and Gao et al. (1994) have described how matrix sampling designs can also be used in performance testing to estimate error variances and coefficients when schools (or other groups of examinees) are the objects of measurement. In this approach, each sample of pupils within a school is randomly split into k sub-samples with each sub-sample taking a different set of tasks. The principal advantage of this design is that a pupil needs to take only a small number of tasks, but data are collected for a large number of tasks. This can lead to substantial reductions in test-taking time without negatively affecting measurement precision, provided relatively large numbers of pupils are available. For example, Gao et al. (in press) found that a matrix sampling design with one hour of testing time per pupil gave a level of generalizability equal to that of a $(p{:}g) \times t$ complete design with 2.5 hours of testing time.

The disadvantages of this design are that large numbers of pupils are required, and many tasks need to be developed. In many circumstances, however, these disadvantages may be an acceptable price to pay for the substantial reduction in per-pupil testing time and the increased content coverage inherent in having many tasks administered.

Concluding Comment

As this chapter illustrates, from at least some perspectives, scores on performance assessments are less generalizable than scores on more traditional tests. At the same time, however, the apparent realism (what some call the "authentic" nature) of performance assessments is intensely appealing to many people. This appeal has led some researchers and practitioners to down-play the importance of reliability/generalizability considerations in the evaluation of performance assessments. That is unfortunate from a technical *viewpoint and often unnecessary from a practical* perspective, as long as excessive claims are not made for performance assessments.

It is undeniably clear that, in most cases, the realism of performance assessments, as currently conceptualized, is purchased at the price of some limitations on generalizability. That does not render such assessments undesirable per se. It does suggest, however, that decision makers need to be cognizant of reasonable restrictions imposed by budget limitations, student time, and rater availability -- restrictions that directly or indirectly limit generalizability.

How might the dependability of scores on performance assessments be increased? When scores are reported for groups only, multiple matrix sampling procedures may help. Also, for either group-level or individual-level scores, it may be advisable to supplement performance assessments with more traditional modes of testing. Another possibility may be create "small" performance assessments that require less administration time, thereby permitting a student to respond to a larger number of assessments. When only a few time-intensive, "large" performance assessments are used, it is probable that in most circumstances student-level scores will not be very dependable.

Footnotes

[1]Strictly speaking, the magnitude of σ^2_{pt} is influenced by the occasion (o) on which the data were collected. That is σ^2_{pt} reflects not only pt interactions but also pto interactions. A similar type of "occasion confounding" influences the other variance components containing p, at least theoretically. By the same line of reasoning, variance components containing r may be influenced by the occasion on which ratings were obtained.

[2]Assuming normally distributed observed scores, a 95% confidence interval covers a width of approximately four SEM's. Consequently, for the difference between the upper and lower limits to be two points, $\hat{\sigma}_\Delta = 2/4 = .50$. For the difference to be one point, $\hat{\sigma}_\Delta = 1/4 = .25$. These values can be used in Figure 1 to obtain the required numbers of tasks and raters.

[3]Throughout this section it is assumed that the same two raters are used for all persons. If different raters were used for each person the reported equations would be different, but the basic conclusions would be unaffected. Also, the discussion could be couched in terms of any number of raters. Two raters are assumed here because that is the most common circumstance.

[4]Dunbar et al. (1991) employed a procedure discussed by Gulliksen (1950, pp. 212-214) to estimate reliability due to scores -- what Gulliksen called "content reliability." Gulliksen (1950) described his coefficient as the "reliability of an essay test corrected for attenuation due to the inaccuracy of reading (p. 214)." It can be shown that Gulliksen's content reliability is $\sigma^2_p / [\sigma^2_p + \sigma^2_{pt}]$, which is called score reliability by Dunbar et al. (1991).

[5]The Linn and Burton (1994) procedure makes heavy use of normality assumptions. For this and other technical reasons, the results should be interpreted cautiously. Still, such results are likely to aid decision makers in qualifying any statements about percents of passing examinees.

<u>References</u>

American College Testing. (1994). *EXPLORE technical manual.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). *Writing: Trends across the decade, 1974-84* (National Assessment of Educational Progress Rep. No. 15-W-01). Princeton, NJ: Educational Testing Service.

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

Breland, H. M., Camp, R. Jones, R. J., Morris, M. M. & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examinations Board.

Brennan, R. L. (1992a). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R L. (1992b). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 14*,385-396.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing tests. *Educational and Psychological Measurement., 55*, 157-176.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14* (4), 9-12.

Carey, N. B. (1991). Setting standards and diagnosing training needs with surrogate job performance measures. *Military Psychology, 3*(3), 135-150.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*, 151-156.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New* York: Wiley.

Cronbach, L.J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995, summer). <u>Generalizability analysis for educational assessments</u> (Evaluation Comment). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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*, 289-303.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Gamache, L. M., & Brennan, R. L. (1994, April). *Issues of generalizability: Tasks, raters, and contexts for the NCBE-PT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Gao, X., Brennan, R. L., & Shavelson, R. J. (1994, April). *Generalizability of group means for performance assessments under a matrix sampling design.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Gearhurt, M., Herman, J. L., Baker, E. V., & Whittaker, A K. (1992). *Writing portfolio at the elementary level: A study of methods for writing assessment.* (CSE Technical Report 337). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. [Reprinted by Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.]

Hieronymus A. N., & Hoover, H. D. (1987). *Iowa tests of basic skills: Writing supplement teacher's guide.* Chicago: Riverside.

Hildebrand, M. R. (1991). *Procedural facilitation of young writers' revisions.* Unpublished doctoral dissertation, The University of Iowa.

Johnson, E. G., & Zwick, R. J. (1988). *The NAEP technical report.* Princeton, NJ: Educational Testing Service.

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

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 27*, 267-292.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). *Interim report: The reliability of Vermont portfolio scores in the 1992-93 school year.* (CSE Technical Report 370). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1992, April). *Empirical evidence for the reliability and validity of performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Lindquist, E. F. (1927). *The laboratory method in freshman English.* Unpublished doctoral dissertation, The University of Iowa.

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

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-15.

Mayberry, P., & Hiatt, C. (1990). *Incremental validity of new tests in prediction of infantry performance* (CRM Report No. 90-110). Alexandria, VA: Center for Naval Analysis.

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

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41-53.

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

Shavelson, R. J., Gao, X., & Baxter, G. (1996). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes, and prior knowledge* (pp. 131-141). Boston, MA: Kluwer Academic Publishers.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Starch, D., & Elliot, E. C. (1912). Reliability of grading high school work in English. *School Review, 20*, 442-457.

Starch, D., & Elliot, E. C. (1913). Reliability of grading high school work in mathematics. *School Review, 21*, 254-259.

S wartz, R., Patience, W., & Whitney, D. R. (1985). *Adding an essay to the GED writing skills test: Reliability and validity issues* (GED Testing Service Research Studies No. 7). Washington, DC: American Council on Education.

Tamir, P. (1974). An inquiry oriented laboratory examination. *Journal of Educational Measurement, 11*, 25-33.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1-21.

Welch, C. (1991, April). *Estimating the reliability of a direct measure of writing through generalizability theory*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Comparability

Edward H. Haertel
Stanford University

Robert L. Linn
University of Colorado

Large-scale assessment programs are moving away from objective, multiple-choice tests to include more "authentic", "direct", or "performance-based" assessments. Using these newly rediscovered assessment formats, students demonstrate what they know or can do by engaging in tasks that should be of interest to them. Scores might even be based on records of actual classroom performance, in the form of student portfolios, projects, or exhibitions. This shift in testing method and rationale raises significant technical and normative issues, several of which center on the problems of assuring that measurements taken at different times, in different places, or using different performance exercises can be validly compared. In this chapter, we attempt to sort out some threats to the comparability of measurements, with special attention to issues that arise with performance assessment. Research and anecdotal evidence are presented when available to inform the likely magnitude of potential problems or to suggest ways to respond, but much of this discussion is necessarily theoretical. Practice is evolving rapidly and to date there are more questions than answers.

With multiple-choice tests, comparability is less problematical. Administration conditions are highly standardized; test stimuli are limited to self-contained, written materials; students work in isolation from one another; and their mode of responding is tightly controlled. Scoring is an almost perfectly objective, mechanical process. Even with these tests, of course, comparability has been questioned. Are scores from a district's first year with a new test comparable to scores the following year? (see, for example, Linn, Graue & Sanders, 1990). Are scores for language-minority students comparable in meaning to scores from native speakers of standard English? Are scores on different forms of the test comparable? How about scores obtained under high-stakes versus low-stakes testing conditions (Madaus, 1988)? How comparable are scores with the same name from tests by different publishers?

With performance testing, similar questions and some new ones arise; and finding satisfactory answers poses greater technical challenges than with more objective tests. This is mainly because most performance tests are less standardized than multiple-choice tests and because they almost always contain substantially fewer items. Standardization is weaker because administration and scoring are more complicated and more difficult to control. Many aspects of testing and scoring may be left to the on-the-spot judgment of administrators, scorers, or even the examinees themselves. Indeed, incorporation of student choice or provision for local adaptation of performance assessments may be touted as a virtue.

The number of independent items on performance tests is usually smaller because each item is more complex and takes longer to complete. Also, the separate questions students respond to within a given performance task are often interdependent. Again, this is seen as an advantage of these tests relative to multiple-choice tests that cover a large number of independent, decontextualized facts or ideas. But, as Yen (1993) has noted, this interdependence can result in

an exaggerated impression of the reliability of the measurement. Furthermore, the use of more items makes scores more comparable because the effects of students' individual reactions to specific items tend to average out when more items are used. One student may have special knowledge that makes this item easy, or another may have a particular misconception that makes that item difficult. These sorts of random influences matter less when averaged over more items. Similarly, a larger number of items increases comparability across test forms, because the particular features of each item matter less when more items are put together. With performance testing, therefore, because scores are likely to be based on fewer independent items, the idiosyncrasies of items and examinees have greater influence on the overall score.

The first major section of this chapter addresses the comparability of scores obtained using a single performance task. This is followed by a section on comparability across performance tasks, focusing especially on tasks intended to be interchangeable or to measure the same areas of knowledge and skill. The third major section takes up comparability at the level of tests including more than one performance exercise. In each section, different threats to comparability are discussed.

A question that must surely arise is what degree of comparability is necessary for operational use of performance assessments. Of course, no simple answer is possible. Different aspects of comparability will be more or less relevant in a given situation. As with any psychometric desiderata, the stringency of comparability requirements will depend on the kind of decision being made (e.g., "absolute" decisions about status with respect to a cutting score versus "relative" decisions about the rank ordering of students or schools); the importance of the consequences attached to those decisions; the level of aggregation at which scores will be reported and used (individuals versus aggregates like classrooms, schools, or states); the relative costs of mistakenly passing versus mistakenly failing an individual; the quality of other relevant, available information and how it is combined with performance test information; and the ease with which faulty decisions can be detected and revised. Still other factors include the cost of additional testing; the size, importance, and duration of the testing program; and the time and money available for further research and development. The psychometric quality of similar performance assessments elsewhere will also inform expectations. Finally, it is difficult to escape the impression that the level of reliability typical of multiple-choice objective tests represents some kind of standard to which tests using other formats may be held.

Comparability of Scores Obtained Using a Single Performance Task

In this section we address the comparability of scores obtained using just one performance exercise with different students or with the same student on more than one occasion. The section begins with comparability of administration conditions, then turns to comparability in scoring, and finally takes up comparability across student populations, examining some student characteristics that may influence the meaning of scores.

Administration Conditions. Imagine a classroom where students are taking a multiple-choice test. Each works alone, silently attending to his or her own paper. If space permits, students may be seated at every other desk. All have received identical instructions, read from a script provided for the test administrator. They work from identical sets of printed questions, recording their responses on identical answer sheets. Rules about what student questions the teacher may answer (and how they are to be answered), whether calculators may be used, and similar matters

are clearly specified. The test is accurately timed. All these administration conditions can be replicated any time the test is administered.

It is more difficult to predict what the same classroom might look like during a performance assessment. With some performance tasks, the testing session might look much the same. For others, the scene would be entirely different. Students might be working in groups; might be using nontext equipment or manipulables; might be free to consult whatever reference materials happened to be available in the classroom; might be free to ask the teacher questions the task designers never anticipated. The logistics of equipment setup and cleanup might compromise the accuracy of timing. The number of students in the class, size of the room, and configuration of desks or other facilities (e.g., sink, electrical outlets) might all affect performance and therefore compromise comparability over classrooms. Even if scripts are provided, the demands made on the test administrator to maintain order and provide logistical support may be considerably greater than with written examinations. Consequently, comparability across test administrators may be diminished.

The argument has been made that if performance assessments are to drive classroom instruction, and if group activity is a desirable feature of classroom instruction, then performance assessments ought to involve group activity. Moreover, the argument goes, ability to work in groups is itself a valued learning outcome and should be assessed. Regardless of the rationale, group activity complicates the comparability of performance assessments. At the individual student level, it may be difficult to disentangle the contributions of each student to a common product. Even if each student turns in his/her own work, the scores of those placed with more or less able peers may not be comparable. Comparability is also compromised at the aggregate level. A teacher's decision to distribute the most able students across all the groups formed or to place them in a single group will affect average group-level performance. Most important, the competencies of individuals cannot be inferred from group-level performances without making strong and probably untenable assumptions about the nature of group processes (Webb, 1995).

With portfolio-based assessments, administration problems multiply. The portfolio usually consists of some required and some optional entries representing the student's best work, culled from up to a year or more of classroom instruction. In this context, rules about appropriate versus inappropriate collaboration or coaching are hard to specify and harder to enforce. A major determinant of the quality of portfolios from a given classroom is likely to be the amount of time and effort the teacher devotes to portfolio-relevant assignments. In addition, the conditions under which students create their portfolios may vary substantially from one classroom to another. Research papers written by students with access to well-stocked school libraries versus an incomplete set of encyclopedias are clearly not comparable unless the conditions under which they were created can somehow be taken into account—a problem for which there is as yet no solution.

The context in which a given performance task is administered also encompasses the perceived consequences of success or failure. In some state testing programs, for example, improvement or decline relative to the previous year's test scores is rewarded or penalized. A similar issue may arise for the National Assessment of Educational Progress (NAEP) if it should ever change from a low-stakes national testing program to an aggregate of potentially high-stakes state-level testing programs.

A final requirement for comparability across administrations relates to test security. In most large-scale assessment programs, comparability over time is of paramount importance. Trend lines are always more informative than isolated results for a single year. Most designs for maintaining comparability over time call for keeping at least some items secure so that they can be readministered under comparable (secure) conditions. With performance exercises, maintaining security may be considerably more difficult than with the simpler, briefer, more numerous items used in the past. Just by virtue of their novelty, performance exercises may be more memorable. Beyond that, the use of a smaller number of such tasks, requirements for special equipment, and even the virtue of greater student interest and involvement may all work against test security, encouraging students to remember what they have done and talk about it with their peers. For low-stakes testing applications in which the focus is on overall trends as opposed to measurement of individuals, one solution might be to administer an exercise to different, randomly equivalent, groups of examinees in successive years.

A counter argument has been advanced (e.g., Frederiksen & Collins, 1989) that performance assessments ought to be public, that students and teachers should know exactly what standards of excellence are, and what is expected. In performance assessments like diving or a skater's compulsory figures, the argument is compelling. But in schools, performing an assessment task is usually interpreted less as an end in itself than as an indicator of some broader capability with respect to a construct domain the task is intended to represent. When the intended inference is to some broader domain, there is a serious risk that teaching directly to a specific task will change the skills it requires and undermine its validity (Madaus, 1988). Nonetheless, at least in theory, a solution to the problem of noncomparability between secure and compromised administrations might be to publicize a large number of performance exercises, from which those on the test are to be selected, well in advance of the first administration.

Comparability in Scoring. In large-scale testing programs, scoring usually follows a model that has evolved over the past decade or more out of experience with performance assessments of students' writing. A scoring rubric is developed and anchor papers are chosen to exemplify each performance level, or sometimes the boundaries between successive levels. Raters are trained in the use of the rubric and must attain a criterion level of accuracy in scoring a set of benchmark papers. During operational scoring, raters are organized into small groups ("tables"), each with a more experienced "table leader" who can answer specific questions as they arise. Some previously scored papers are seeded throughout those being scored for the first time so that each rater's accuracy can be monitored continually. This helps to keep raters' standards from drifting over the course of a scoring session. If an individual rater appears to be performing below standard, the table leader may "read behind" that rater, monitoring his or her performance more closely until the problem is corrected. By these methods, with some care and effort, relatively high levels of scorer consistency can usually be attained (see, for example, Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993). Even with rigorous training and frequent calibration checks, however, scorers still contribute a source of uncertainty and thereby increase the overall measurement error. Moreover, open-ended responses are more complex and time consuming to score than selected responses; interrater reliability must be examined for each new assessment.

Performance assessment scoring may be viewed as a social process in which a group of individuals negotiates meanings and comes to consensus about the interpretation of scoring rubrics. Rules for handling unusual responses may be formulated on-the-fly and communicated

verbally to members of the group. If the entire scoring process is replicated later, the emergent consensus may not be quite the same. This especially threatens the comparability of scores from year to year. If it is possible to seed some papers from the previous year along with new papers to be scored, any drift in standards may be estimated. But if writing prompts or other aspects of assessments have changed from year to year, it may not be possible to have raters make equivalent judgments of responses to different tasks, nor to conceal the year each response represents.

Photocopying of students' papers, sometimes entailed by the logistics of rescoring, may also affect comparability. A study of score reliability for student performance data from the Kentucky Instructional Results Information System (KIRIS) found a significant effect of scoring photocopies versus originals at one but not another grade level (Richard Hill, personal communication, January 14, 1994).

Finally, perhaps making a virtue of necessity, most state-level performance assessment systems rely on classroom teachers for much of the work of scoring and in some designs, teachers may score their own students' work. It is reasonable to assume that in some cases, teachers' scorings of their own students' work may not be comparable to scores assigned by other raters, especially if teachers are aware of the identity of the students.[1]

*Comparability Across Student **Populations.*** Every test is designed with some target population in mind. Decisions about content, format, layout, timing, and instructions are all conditioned by the age, language, and culture of the intended examinees. This is in part because the curriculum's intended learning outcomes differ according to students' ages and other characteristics, but also because any assessment task requires skills beyond those it is intended to measure. Successful performance typically depends on examinees' understanding that they should show their best work; their willingness to do so; their ability to understand the task requirements; and their mastery of the communication skills necessary to produce storable responses. These and other attitudes or capabilities not explicitly part of what is to be measured but nonetheless necessary for successful test performance may be referred to as the test's ancillary or enabling skill requirements. If some examinees are deficient in a test's *ancillary* abilities, then it is biased against them. They will not score as well as other examinees equally proficient with respect to the knowledge or skills the test was designed to assess. Variation in scores of otherwise equally capable examinees due to differences in their ancillary abilities gives rise to *construct-irrelevant variance* in test scores.

The ancillary skill requirements of performance assessments are likely to exceed those of more conventional tests, although, as with any test, they may be minimized through careful test design. The materials and instructions provided are more complex and varied and the required modes of responding are more demanding. Perhaps the most obvious threat to the validity of most performance assessments is their dependence on reading and writing. Scores for native speakers of standard English may not be comparable to scores of students for whom standard English is a second language or dialect. Consider a performance assessment in mathematics, say, that calls for students to explain their approach to a problem. Even if communicating about mathematics is explicitly included in the knowledge and skills to be assessed, scores are likely to be interpreted as reflecting primarily mathematics proficiency. Clearly, when the examinees are language-minority students, that interpretation must be made with considerable caution. At the

very least, the relative contributions of different abilities to the construct measured will vary from one language group to another.

In addition to language, comparability across student groups may be limited by differences in motivation (or understanding of and compliance with the demand characteristics of the testing situation). The rhetoric of performance assessment sometime's suggests that students will almost inevitably find these new forms of testing activities stimulating and engaging. There is an often-repeated story of a fifth grader in California asking, after a performance test in science, "Can we take the test again tomorrow?" But children do not always take to the activities adults think they should, and other anecdotes suggest that students' favorite assessments may be those requiring the least effort, especially the least writing. Low motivation may distort the scores of *all* students and may be especially problematical for students who must struggle with writing.

Developers of performance assessments must also remain alert to the impact of a range of disabilities that might be irrelevant in more traditional testing situations. Difficulties with fine motor coordination, physical handicaps, or color blindness may threaten the comparability of scores for individual students.

Finally, a major determinant of performance test scores may be students' previous exposure to similar tasks, given either as performance assessments or in the course of regular classroom instruction. (Students are less likely to be familiar with novel testing formats.) If performance assessments become routine in the business of teaching and testing, then differences in exposure to novel task formats may no longer threaten score comparability. But at least for the next several years, students' uneven experience with performance testing may be a consideration. Initial dramatic year-to-year improvements in scores may have more to do with students' acquisition of ancillary skills than of target skills.

Comparability Across Performance Tasks

The previous section described threats to comparability that may arise even when the assessment task itself is held constant. In this section, we add a layer of complexity, focusing on the comparability of scores from different performance tasks. Following this discussion, the third major section of the chapter examines comparability of scores from entire tests. It will be seen there that aggregating across a number of performance exercises can mitigate noncomparabilities at the level of individual tasks, although careful test design, empirical studies, and usually, statistical adjustments will still be required to assure adequate comparability in large-scale testing programs.

The comparability of entire tests depends strongly on the comparability of single items, but item-level comparability is also of interest in its own right. Consider the case of a state-level testing program that administers different sets of items to different students in order to improve school-level achievement estimates, but which also produces individual-level scores. Unless students' scores are based solely on items administered to all of them in common, some degree of comparability must be assumed across the items given to different students. (There is a dilemma here. The more comparable the matrix-sampled items are, the less matrix sampling improves content coverage.) Item-level comparability is also implied when benchmark proficiency levels are illustrated with items drawn from an exercise pool. The interpretability of proficiency scales is only enhanced to the extent that the illustrative exercises are truly representative of all those at the corresponding proficiency level.

Measurement intents, ancillary requirements, and error variance. In analyzing the comparability of single performance tasks, it is helpful to think of scores as having three distinct components. First is the *intent* of the measurement-the construct-relevant knowledge, skills, or dispositions the task is designed to measure[2]. Second is the set of ancillary requirements-additional construct-irrelevant knowledge, skills, or dispositions required for task success, including but not limited to familiarity with task instructions, motivation to show one's best work, and test: wiseness. Language proficiency and other communication skills may fall within a measurement's intent or within its ancillary requirements, as discussed in the previous section. Finally, scores inevitably include a component of *error—A* complex mixture of both random and idiosyncratic influences on scores. The comparability of performance tasks is a function of the similarities and differences among their intents, their ancillary requirements, and their error structures. Thus, it is useful to distinguish these components even though it may not be possible to separate them or to determine in which of these three respects two assessment: tasks actually differ.

This section next turns to comparability when all three of these components are the same and then when error structures, ancillary abilities, or both diverge. The comparability of performance tasks with different intents is then briefly considered. The section concludes with a brief treatment of additional issues raised by assessment designs permitting student or teacher choice.

Comparability among tasks with the same intent, the same ancillary requirements, and the same error structure. If tasks require the same ancillary and intended abilities and if the scores they yield are equally accurate for students at any given level of those abilities, then the scores they yield should be comparable. Such tasks satisfy the requirements for equating, the strongest form of linkage identified by Mislevy (1992a) and Linn (1993). In principle, given any two such tasks, say X and Y, it should be possible to find a single equating function that could be used to transform scores on task X to equivalent Y-scores or conversely. Note that even in this ideal case, it would not necessarily be appropriate to use the tasks interchangeably without applying some statistical adjustment. One task's scoring rubric might be more stringent than the other's, for example, in which case the raw scores from the two tasks would not be on the same scale. Note also that statistical equating methods might be difficult to apply if scores on the two tasks took on only a few discrete values (e.g. 1 - 6).

The most important question about this ideal case of common intents, ancillary requirements, and error structures is the degree to which it can be approximated in practice. An ongoing NSF-sponsored project being carried out by the RAND Corporation in collaboration with several other institutions is examining the comparability of science performance assessments designed to have varying degrees of parallelism. In one comparison, two tenth-grade science performance assessments were constructed, called *Rate of Cooling* (ROC) and *Radiation* (RAD). In ROC, students compare three fabrics to see which is the best insulator; and in RAD they compare three paint colors to see which absorbs the least radiant heat energy. Each experiment included three trials, one for each material, in which temperature was measured at specified time intervals. Temperature changes were recorded and graphed, and parallel or identical questions were posed about temperature change, heat energy, and the concept of rate of change. Despite the formal similarity of the two tasks, observations during pilot studies suggested significant differences in the accuracy of students' measurement, the time required for the two tasks, the difficulty of manipulating the equipment, and the kinds of errors students made. This study

highlights the importance of attention to ancillary requirements of tasks as well as to the abilities they are intended to measure.

Shavelson, Gao, and Baxter (in press) report another study that shows how different levels of task comparability affect generalizability. The study underscores the importance of clearly specifying the domain of generalization. Three experimental tasks involving the choice behavior of sow bugs were administered to a sample of fifth and sixth grade students. Task 1 required students to design an experiment to determine if the bugs would choose a light or dark environment. Task 2 called for an experiment to determine if they would choose a wet or dry environment. Task 3 asked students to determine what combination of conditions from the first two experiments (wet and dark, wet and light, dry and dark, or dry and light) the bugs would choose.

The generalizability coefficient for a single experiment was .51 when the data from all three tasks were analyzed together. When the third task was excluded from the analysis, however, the generalizability coefficient increased to .62, showing that the first two tasks were more nearly equivalent to each other than either was to the third experiment. The third experiment involved the crossing of two factors rather than the consideration of a single factor, and as Shavelson, Gao, and Baxter (in press) noted, such designs were not part of the elementary science curriculum. Here, the first two tasks more nearly approximated the ideal of common intents, ancillary requirements, and error structures than did the set of all three tasks.

Few generalizability studies have clearly distinguished between tasks intended to be interchangeable and those intended to measure different aspects of the subject-matter domain. Results such as those presented by Shavelson, Gao, and Baxter illustrate the importance of this distinction for the design of performance-based assessments and for the evaluation of their comparability. They also suggest that greater use might be made of multivariate generalizability models in analyzing performance test data.

Comparability among tasks with the same intent, the same ancillary requirements, but different error structures. If two tasks measure the same abilities but with different degrees of precision, they satisfy the requirements for *calibration,* Mislevy's (1992a) and Linn's (1993) second strongest form of linking. Large-scale testing programs rarely examine differences in precision at the level of single tasks, but it is clear from a consideration of the calibration model that such differences could be important. Consider, for example, a writing assessment in which responses to two different writing prompts, X and Y, are scored using prompt-specific rubrics derived from the same "generic" scoring rubric. Suppose the correlation between students' scores on the two tasks is .64. It would be typical to treat the two tasks as parallel (in particular, equally reliable) and to assume that each correlates .80 with the same underlying true score. But suppose prompt X is a little more ambiguous than Y, or its rubric is not as clear, or that anchor papers for X are not as well chosen. The correlation of .64 between the two tasks could equally well imply that Tasks X and Y correlate .750 and .853 with the true score, respectively. Following the logic of Linn's (1993) discussion of calibration, if scoring rubrics are constructed so that comparable proportions of students are classified at each score level, the most able students would then have a better chance of receiving the highest classification using Task Y and weaker students would have a better chance using Task X. Likewise, schools with the highest-achieving students would have a relative advantage on Task Y and conversely.

Evidence about differences in error variances across tasks could be obtained in several ways. In the example above, differences between X and Y due to the anchor papers or rubrics used in scoring could be detected by comparing the interrater reliabilities for the two tasks. Differences due to the prompts themselves could be estimated from the correlations of Tasks X and Y with those on a third prompt, say Z. Further evidence might be obtained by comparing patterns of correlations of Task X versus Y with other variables, although any differences might reflect discrepancies in the tasks' ability requirements as well as their error structures.

Comparability among tasks with the same intent, different ancillary requirements, and the same error structures. If assessment tasks differ in their ancillary requirements, then scores on those tasks will be comparable only if each is given to students in full command of the ancillary abilities it requires. Conversely, any single task that relies on ancillary abilities that one group of students possesses and another does not will be biased in favor of the first group. In particular, tasks with different ancillary requirements may be needed for comparisons across language groups or groups of students with different instructional histories. An example may be helpful. Consider the case of cross-national assessments where tests must be administered in several languages. The designers of such tests try to avoid knowledge or conventions that may be unfamiliar in some countries, unless that knowledge is specifically what an item is intended to measure. Items are translated and back-translated to assure that the meanings of questions posed in different languages are as near to identical as possible. The goal of these efforts is a set of tests differing only in certain of their ancillary requirements. Ideally, all of the students taking a given test would have full command of the language in which it was written, so that apart from random error, score differences would reflect only what the tests were intended to measure. If a mathematics test written in French, say, were given to students in France and the United States, it would be biased against U.S. students who lacked (ancillary) language abilities. Likewise, a test written in English would be biased against the French students. Matching the ancillary requirements of the test forms to those of the student populations greatly improves comparability, although the absolute level of comparability attained is difficult to ascertain.

An analogous problem arises in trying to measure complex reasoning abilities among students who have studied different topics in a given content area. Thinking must be about something. If the intent is to measure complex thinking processes, the content knowledge to which those processes are applied may be regarded as ancillary to the intent of the measurement. Take as an example the idea of a food chain. In one classroom, children who have studied the ecology of a meadow might construct a food chain with foxes, mice, and grass seeds. In another classroom, children who have studied the ocean might construct a food chain with baleen whales, krill, and phytoplankton. Some biology students may be able to discuss the relation of structure and function in the crayfish and others in the frog. Some students may be able to discuss number patterns in Diophantine equations and others in continued fractions. In each case, even if the reasoning processes are the same, those processes are applied to different knowledge structures. An assessment tied to any particular curriculum will be biased against students who have studied some other curriculum.

One response would be to attempt to construct a "curriculum-fair" test, with items sampled from a lot of different curricula, but that is not likely to be practical, especially not with performance tests. Another approach would be to build items equally unfamiliar to all students, but that is likely to change the nature of what is measured, resulting in a test that depends more on 'G,' or general mental ability. Moreover, the logic and rhetoric of performance testing call for

closer ties to the curriculum. Even if irrelevant, decontextualized items were fair to everyone, they would be rejected on other grounds. The most popular solution. is to offer students and teachers some degree of choice among alternative performance tasks intended to measure the same thing, but as discussed below, such student or teacher choice brings its own complications.

As difficult as it is to approach comparability among tests written in different languages, that problem seems almost trivial compared to building comparable tests of higher-order thinking that reflect different curricula. At present, the best that can be hoped for appears to be linkage at the level of Mislevy's (1992a) or Linn's (1993) moderation. Either human judges will have to reach consensus that alternative assessments reflect the same abstract performance standards ("social moderation") or adjustments will have to be imposed based on the correlations of different assessments with some common anchor test ("statistical moderation"). As Linn and Mislevy both discuss, linkages established in this way will require frequent reexamination. They are unlikely to remain stable over time or to be consistent across different groups of examinees.

Comparability among tasks with the same intent, different ancillary requirements, and different error structures. Even tasks intended to be interchangeable are at best only approximately parallel. It will be some time before our cumulative experience with performance assessment is sufficient to develop "rules of thumb" about the functional exchangeability of tasks constructed in different ways. Work like that of Shavelson, Gao, and Baxter, discussed earlier, is illustrative of the kinds of studies that will be required.

A more extreme case of assessments with the same intent but different ancillary requirements and different reliabilities is a comparison of performance assessments, simulations, and written examinations by Shavelson, Baxter, and Pine (1992). In this study, direct observations of students performing hands-on science tasks were compared with a scoring of their notebooks, with their responses to computer simulations, with short-answer questions, and with multiple-choice questions, all designed to measure the same content. The authors found that in some cases, computer simulations were fairly good surrogates for hands-on tasks, but that paper-and-pencil tests did not appear to measure the same capabilities. Correlations between hands-on and written task versions were only moderate. Hands-on performance was better for students who had received more innovative science instruction than for those who had not, whereas paper-and-pencil test scores did not differ for the two instructional groups. However, paper-and-pencil tests correlated more highly with standardized achievement tests than did the hands-on tests. Taken together, these results suggest that hands-on versus paper-and-pencil tests measure different things. In particular, hands-on tasks measured some abilities developed especially through innovative science instruction (presumably part of the intent of the measurement), whereas paper-and-pencil test scores were more influenced by (presumably ancillary) ability requirements that they had in common with the standardized achievement tests.

Comparability among tasks with different measurement intents. If comparability is viewed as a purely technical problem, then the best that can be hoped for by way of linkage among tasks with different intents is projection or moderation (Linn, 1993; Mislevy, 1992a). Comparability among single performance exercises designed to measure different abilities is problematical at best, which may be one reason why generalizability studies of performance exercises tend to show such large person-by-task interactions[3]. As noted earlier, ongoing research at the RAND Corporation and elsewhere should help to inform the levels of generalizability that can be achieved with tighter control over performance exercise content. One strong implication of this

analysis is that complex assessment tasks incorporating student or teacher choice should be designed to assure that whatever specific materials or questions are selected, some common set of target skills is engaged.

An alternative perspective on comparability among tasks with different measurement intents would focus not on the adequacy of generalization to a common universe, but rather on the value placed on the disparate performances represented by the separate tasks themselves. The performance of student A on task X and of student B on task Y might be considered comparable if they represent distinct, nonexchangeable, even incommensurable, attainments that are nonetheless regarded in some sense as being of equal worth. This normative component is rarely made explicit (Wiley & Haertel, in press). Even though the rhetoric of educational reform often stresses individualization, the theory and practice of educational measurement is almost exclusively concerned with the comparison of different students' attainments with respect to a common construct domain or intended learning outcome. In particular, some testing programs (e.g., portfolio-based assessments) offer teachers or students significant latitude in determining what assessment tasks they will undertake or what work they will submit for evaluation. When common scoring rubrics are applied to the resulting range of different performances, they effectively assign scores representing the value placed on a student's attainment, as opposed to describing the attainment itself (i.e., the score alone tells how good the student's work was, but not what the student did). There is a need for further development of both technical and philosophical bases for making such judgments explicit and for reaching consensus about the meaning of such scores.

Student or teacher choice. In classroom testing, it has long been popular to include sets of questions among which students choose a specified number to answer, omitting the remainder. This practice is generally discouraged by measurement specialists because it muddies the generalization from the sample of tasks administered to the domain they represent. If students are free to choose, they will probably avoid questions to which they do not know the answer. Thus, one cannot infer that a student who answers 80 percent of the attempted questions correctly would have about an 80 percent chance on another question randomly chosen from the domain. Choice also reduces comparability because the scorer must judge the relative quality of students' answers to different questions and because students will differ in their ability to choose the subset of questions they can answer to best advantage. Thus, test wiseness looms larger among the constructs actually measured. Despite these disadvantages, as explained in the last section, comparability could in principle be improved if students or teachers could select tasks requiring ancillary skills that matched their own capabilities from among a set of tasks having the same measurement intent. Thus, for example, some students might choose a food-chain assessment based on the ecology of a meadow, and others a food-chain assessment based on the ecology of the ocean.

Performance assessments in some large-scale testing programs have given students or teachers choices among alternative texts. Students may be offered a choice of what to read, to increase their motivation or to provide a context for studying their ability to offer reasoned justifications for such choices. Teachers might be offered choices in the materials used with younger children so that they can make allowances for differences in their interests or levels of reading ability.

The NAEP began to experiment with student choice in the 1992 Reading Assessment. Students at grades 8 and 12 were given paperback anthologies containing seven stories of about 1,000 words each, selected from various sources of authentic young-adult publications, by different authors and featuring a range of cultures. During the 50-minute testing period, students were to select and read one story and then answer twelve questions about why they chose the story they did, how they liked it, and about plot, theme, character, setting, and any personal relevance the story might have had to themselves. Clearly, allowing students to choose a story reduces comparability because the stories might not be equally difficult or the questions asked might not apply equally well to all of them. Most problematical, student choice in this context complicates generalization about the entire eighth or twelfth grade population's performance on any specific story. Further research in subsequent NAEP Reading Assessments should help to gauge the magnitude of these potential effects. NAEP Readers are to be given to nationally representative samples under a condition in which a specific story is assigned as well as under the choice condition. It should thus be possible to learn how students perform on a story they have chosen versus one randomly assigned.

In the matrix sampling situation where students respond to different matrixed tasks, it is reasonable to assume that each task would be answered similarly if administered to a different, randomly equivalent sample. Hence, the data may be treated statistically as resembling a set of responses to a long test (formed from all of the alternate test forms together), with data missing at random. Where students choose which task to complete, the problem is complicated because the missing data are no longer random. Consequently, one may no longer assume that on average, each task would be answered in the same way by those examinees who did not respond to it as by those who did. As discussed by Allen, Holland, and Thayer (1993) and Wainer, Wang, and Thissen (1994), equating tests where examinees choose which problems to attempt depends on strong, usually untestable assumptions. Furthermore, recent experimental evidence presented by Wang, Wainer, and Thissen (1995) raises serious doubts about the viability of those assumptions.

The 1991-92 administration of the Standard Assessment Tasks (SATS) to 7-year-old children in England illustrates some potential problems when teachers are offered choices. A task was included at Level 2 that asked children to read aloud from one of 20 specified books (Gipps, 1993). The teacher could select any of the 20 books for any given child. As a result, some children were asked to read from a book that was new to them while others were asked to read from a book with which they were quite familiar. As noted by Gipps, "obviously for some children the task was much easier since they already knew the story and had practiced reading it" (p.11). Reading aloud from an unfamiliar book and reading aloud from a book after having practiced the reading may both be skills worthy of assessment, but tasks calling for one or the other should not be considered interchangeable. Such variability in assessment conditions seriously undermines the comparability of children's performances.

In current student portfolio systems, even "required" entries offer considerable latitude for student choice. One required entry from the 1991-92 Kentucky Instructional Results Information System (Kill IS) writing portfolio, for example, was "A personal response to a cultural event, public exhibit, sports event, media presentation, or to a book, current issue, math problem, or scientific phenomenon". Such broad specifications reflect the tension between instructional and accountability purposes for portfolios. Considerable value is placed on giving students the opportunity to choose what they consider their best work, but comparability is necessarily

compromised. Less standardized entries are more difficult to compare and opportunities for savvy students or teachers to make strategic choices are increased.

Comparability of Tests Including Performance Exercises

This last major section moves from the level of individual exercises to entire tests or exercise pools. Even though current assessment programs still rely primarily upon objective written questions, the discussion focuses on a hypothetical assessment comprising performance assessments only. The section begins by discussing the importance of clear content specifications for any assessment, and then argues that in order to assure comparability over time, assessments using performance exercises may require more detailed and more fine-grained content specifications than are typical of current assessment programs.

The importance of content specifications in large-scale assessments. A major goal of most large-scale assessment programs is to provide scores that are comparable over time. In principle, such comparability requires that the content assessed remain the same. Mislevy (1992b, p.200) for example, acknowledges that despite the appropriate and successful use of unidimensional statistical models to scale NAEP data, the NAEP exercise pools are not really unidimensional. He likens the exercise pool to the standard market basket of goods used to calculate the Consumer Price Index. If the NAEP's "market basket of skills" is changed, then the meaning of NAEP scale scores is changed.

That being said, for large-scale assessments that rely almost exclusively on brief, objective test questions, close attention to the precise content assessed does not seem to be very important. A large number of items are used, so that each specific item has only a small influence on aggregate score distributions. Moreover, most items are retained from one assessment cycle to the next, so that even if items added to the pool are somewhat different from those removed, the overall content mix changes only slowly. Finally, in assessments using item-response theory (IRT) methods, the difficulty of each new item is automatically accounted for as it is added to the pool, and items measuring abilities less highly correlated with the major dimension assessed by the entire pool may be assigned lower discriminations and thereby receive less weight.

IRT equating and gradual replacement of exercises notwithstanding, as ideas about curriculum and instruction have evolved and curriculum frameworks have been revised, there have at times been significant shifts in the kinds of content tested. Turning once again to NAEP, Baldwin (1989) and Pandey (1989) documented substantial shifts over time in the proportions of reading and mathematics exercises classified into different content categories. Even trends over adjacent time intervals have been based on quite different content mixes.

Minor shifts in content from year to year are tolerated in any assessment program, as some exercises are retired due to technical flaws, because they have been published as illustrative items or because some aspect of their content has become dated. The new items created to replace such retired items rarely measure the identical skills. Nonetheless, major shifts in content may require the introduction of a new scale and the start of a new trend line. The shifts in the NAEP content frameworks and assessments, for example, were judged to be too great for the 1990 mathematics assessments and the 1992 reading assessment to continue the same scale and trends. As was stated in the 1992 reading assessment report, "the changes in the 1992 reading framework and assessment activities preclude any comparisons between the results of this report and those for previous NAEP reading assessments" (Mullis, Campbell, & Farstrup, 1993, pp.

2-3). Comparisons to prior assessments are accomplished by "readministering the long-term reading assessment" (p. 3) to special trend samples. Thus, for a period of time after a major shift in the exercise pool, NAEP actually maintains two trend lines reflecting the old and the new frameworks and assessment activities.

Content specifications for assessments based on performance exercises. With performance exercises, careful and detailed specifications become much more important. Not only the measurement intents but the exercise format, ancillary content, and other characteristics may need to be specified, possibly down to the level of individual exercises. This is because relatively few performance exercises can be used and each is generally designed to measure something different from all the others. Detailed content specifications also take on greater importance with performance exercises because on the whole, they are more heterogeneous in format (and therefore ancillary requirements) than objective written test questions. More detailed specifications are called for because there is more to specify. Timing, materials, specific administration instructions and conditions, and other aspects of performance exercises are more variable than for multiple-choice questions.

It is no surprise that in general, the performance exercises included in large-scale assessments each measure a distinct learning outcome. Because of their cost in development, administration time, scoring time, and other resources, relatively few performance tasks can be included in an assessment. At the same time, most assessments must sample broad content domains, defined by ambitious, inclusive documents like the NAEP content frameworks or state curriculum frameworks often covering a year or more of instruction. Every exercise has to contribute as much as possible. The information gained by adding a task covering some new curriculum element is greater than that gained by adding redundant measures of content-and-process categories already sampled.

The facts that performance exercises are fewer in number and more distinctive imply that they must be sampled with greater care than multiple-choice items if the "market basket" of knowledge and skills assessed is to remain stable over time. Rather than treating a successive year's tasks as another random sample from a broad content domain, new exercises could be constructed such that each was closely parallel to the specific exercise it was intended to replace. Such a detailed sampling plan would permit the more fine-grained psychometric analysis described below.

Evacuating comparability of performance-based assessments. Statistical analyses of tests are largely divorced from content analyses. Although factor analysis may be used to establish that a heterogeneous collection of items is "sufficiently unidimensional" to justify IRT scaling, it is common to ignore finer content distinctions when items are scaled. In test design, item selection is often aimed primarily at assuring sufficient precision throughout the measurement range, with content representativeness being treated merely as a constraint, (i.e., a certain number of items must come from each content category). Analytical methods that were appropriate for tests comprising many brief items are proving inadequate for tests comprising a smaller number of more costly items, more carefully chosen. New statistical methods may be invented (e.g., Haertel & Wiley, 1993), but in the meantime, it may be possible to do a better job with current statistical tools.

Consider a test or exercise pool comprising ten performance exercises. All ten might be administered to each examinee to obtain an individual score, or the ten exercises might be

administered to distinct random samples of students in a school in order to obtain a school-level estimate of overall achievement. Under the analysis proposed, detailed specifications would be prepared for each of the ten separate exercises, each specification describing a *stratum* that the corresponding exercise represented. Then another, corresponding set of ten exercises would be written, closely parallel to the first ten. There would now be two exercises in each stratum. The purpose of this finer stratification of the exercise domain and the additional exercise construction would be to unconfound random and task-specific variance from that attributable to the distinct measurement intents of different exercises (each different intent corresponding to its own stratum). It is, of course, an unanswered empirical question what degree of error reduction would be possible. The answer would depend on the relative magnitudes of within-stratum and between-stratum variation. (See Cronbach, Linn, Brennan, & Haertel, 1995, for further discussion).

In a series of ten separate generalizability studies, each pair of parallel exercises would be administered to a sample of examinees to assess within-stratum generalizability, and the results of these analyses would be used to estimate the precision of a composite score treating exercise within stratum as random, but stratum as fixed. Error variances calculated in this way might be considerably less than the those calculated under the assumption that the ten tasks are all randomly sampled from a single, undifferentiated pool.

This model is not without its difficulties. To begin with, most performance exercises appear to be unique. It would be no easy task to define what each was measuring in a way that was sufficiently specific to realize the benefits of tighter control on assessment design and at the same time sufficiently general to permit the creation of one or more parallel forms. And after the specifications were written, the work of creating additional exercises and carrying out all those G studies would still remain. Perhaps most serious, in a typical assessment, the ten carefully specified strata would encompass only a small fraction of the content the assessment was intended to represent. The idea of fielding assessments year after year that amounted to minor variations on the same ten exercises would seem unappealing. If the assessment were high-stakes, instruction would soon focus on those ten strata, and the validity of the assessment as a measure of the original, broader domain would erode.

In fact, there would be no need to limit the evolution of the test through time. Rather, it would be accepted that part of the work of developing a performance exercise was at the same time to specify its "shell," that is, its stratum definition, and, at least for purposes of analysis, to construct a second exercise conforming to the same shell. Comparisons over time would be based only on common shells, but during each assessment cycle some new shells could be introduced. Data could be analyzed both to give precise estimates of status and trends with respect to a particular "market basket of skills" and also to obtain more honest estimates of generalizability to broader universes (see Kane, 1982, for further relevant discussion). Over time, the cumulation of studies focusing on domains in different content areas, defined in different ways and at different levels of specificity, would lead to steady improvements in the technology of performance assessment.

Summary

Comparability of performance exercises is a complex topic, encompassing multiple concerns of greater or lesser importance in any given situation. This chapter's discussion is somewhat

theoretical due to the paucity of empirical studies on performance tasks constructed following specific, replicable procedures. Although most of the specific threats to comparability of performance assessments also pertain to brief, objective, written test questions, their magnitude is greater for performance exercises because fewer such items are used in a typical test and because standardization is often considerably weaker for these types of items. In this chapter, we first examined the comparability of scores obtained using a single assessment task. Even holding the task constant, comparability may be limited due to uncontrolled variation in administration conditions, inclusion of group activities in the assessment, or task definitions that incorporate student or teacher choice concerning the specific tasks attempted or materials used. Also, test security may be more difficult to maintain for performance exercises, threatening comparability if tasks are reused with the same students or with different classes in the same school during several periods of the school day. Additional issues include consistency in scoring and the use of performance exercises with different student populations, especially students from different language groups, with handicapping conditions, or with different instructional histories.

Occasions often arise when scores on two or more different performance exercises must be compared, most notably when students or teachers are offered a degree of choice in the performance exercises selected. The chapter next turned to comparability across performance exercises, considering first the case of two exercises intended to measure the same knowledge, skills, or dispositions (same measurement intent); requiring the same additional, enabling skills (common ancillary requirements) to respond to the performance task itself; and having equal levels of accuracy (common error structures). Two such exercises satisfy the requirements for statistical *equating*, and present minimal problems of comparability.

The next case considered was that of common intents and ancillary requirements but different error structures. Tasks with this degree of similarity satisfy the requirements for statistical *calibration*, a weaker form of linkage than equating. For some purposes, one would want scoring rubrics for two such tasks to classify examinees in such a way that scores on the less reliable task had higher variance, and for other purposes, one would want the variances to be the same. Information on reliability at the level of specific tasks is rarely reported, but would not be especially difficult to obtain, and would be helpful in making decisions about category boundaries in scoring rubrics and other aspects of data analysis.

The case of common intents, different ancillary requirements, and common error structures arises when performance exercises are translated into different languages (as in cross-national assessments) or when higher-order thinking is compared for students studying different curricula by having each group reason about the material they have studied. As a practical matter, it is likely that the highest level of comparability attainable across curriculum-specific assessments will be that of statistical or social *moderation*.

Tasks intended to measure the same learning outcomes but having different ancillary requirements and error structures might include performance tests versus paper-and-pencil "surrogates." Here again, linkage at the level of statistical or social moderation is probably the best that can be attained. The same is true a fortiori of tasks with different intents.

The third major section of the chapter turned to comparability at the level of entire tests. Here, many of the concerns of the previous chapter are somewhat mitigated because the specific qualities of each item count for less in a composite score. It is argued that imprecision could be reduced and that the remaining imprecision could be evaluated more accurately if content

specifications for performance exercises were prepared in greater detail and tests were constructed as purposively weighted collections of tasks sampled from narrow content domains. Some practical implications of such test design strategy were considered. It may be hoped that over time, the cumulation of studies on methods of domain definition and performance task construction will lead to better assessment technology to improve educational practice.

Notes

[1] This may be viewed as a threat to comparability, but Moss, Beck, Ebbs, Matson, Muchmore, Steele, Taylor, and Herter (1992) offer another interpretation. They argue that classroom teachers are privileged observers, able to bring a richer interpretive context to bear in evaluating their own students' work: "The central interpretation will be the classroom teacher's interpretation, and it will be based not only on the portfolios but also on extensive knowledge of the students, their goals, and their instructional opportunities... .This approach . . . acknowledges the singular value of the teacher's knowledge base in making interpretations, which cannot be duplicated by outside readers" (Moss, et al., 1992, p. 19). It might be added by way of rejoinder, however, that privileged observers can also be biased observers.

[2] In using the term "intent, " we do not mean to imply that intentions are sufficient. To disentangle these components and determine the extent to which each is reflected in actual test scores would require supportive evidence from empirical research.

[3] Large person-by-task interactions may also be reported because variance components representing the person-by-task interaction and the person-by-task-by-occasion interaction are confounded. The latter, three-way interaction represents the instability across a single individual's (hypothetical) repeated trials on the same task. The total of these two variance components is often labeled "person-by-task interaction."

References

Allen, N. L., Holland, P. W., & Thayer, D. T. (1993). *The optional essay problem and hypotheses of equal difficulty* (ETS Technical Report No. RR-93-40). Princeton, NJ: Educational Testing Service.

Baldwin, J. (1989). Reading trend data from the National Assessment of Educational Progress: An Evaluation. In National Center for Education Statistics, *Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons* (Report No. CS 89-499, pp. 85-94). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1995). Generalizability analysis for educational assessments. *Evaluation Comment* (Summer 1995, whole issue). Los Angeles, CA.: Center for the Study of Evaluation & National Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles.

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

Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher, 18*(9), 27-32.

Gipps, C. V. (1993, April). *Reliability, validity, and manageability in large-scale performance assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum.

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

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

Linn, R. L., Graue, M. E., &: Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice, 9 (3), 5-14.*

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (Eighty-seventh yearbook of the National Society for the Study of Education, part 1, pp. 83-121). Chicago: University of Chicago Press.

Mislevy, R. J. (1992a). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Mislevy, R. J. (1992b) Scaling procedures. In E.G. Johnson & N.L. Allen (Eds), *The NAEP 1990 Technical Report* (Report No. 21-TR-20, pp 199-213). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), *12-21.*

Mullis, I. V. S., Campbell, J. R., & Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States. Report* No. *23* ST06. Washington, DC: National Center for Education Statistics, U. S. Department of Education.

Pandey, T. (1989). Mathematics trends in NAEP: A comparison with other data sources. In National Center for Education Statistics, *Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons* (Report No. CS 89-499, pp. 95-108). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

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

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Measurement, 21*(4), 22-27.

Shavelson, R. J., Gao, X. & Baxter, G. P. (in press). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum, R Dochy, & D. Van Aalsvoort *(Eds.), European innovations in assessment and evaluation.* Norwell, MA: Kluwer.

Wainer, H. Wang, X-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement, 31,* 183-199.

Wang, X-B, Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied* Measurement in *Education, 8, 211-225.*

Webb, N.M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17,239-261.*

Wiley, D. E., & Haertel, E. H. (in press). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell *(Ed.), Implementing Performance Assessment: Promises,, Problems, and Challenges. Hillsdale,* NJ: Erlbaum.

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

Setting Performance Standards for Performance Assessments: Some Fundamental Issues, Current Practice, and Technical Dilemmas

Richard M. Jaeger
University of North Carolina, Greensboro

Ina V. S. Mullis
Boston College

Mary Lyn Bourque
National Assessment Governing Board

Sharif Shakrani
National Center for Education Statistics

Since the turn of the century, with the pioneering work of Binet, human cognitive abilities have been broadly assessed by administering selected-response, pencil-and-paper tests that purport to measure various achievements, aptitudes, and knowledge that are held to be indicative of success in school and of schooling, prerequisite to effective performance in academic pursuits, essential to success in a wide variety of careers, and predictive of performance on the job. Selected-response tests that require the darkening of small bubbles with number 2 pencils have grown to be a cultural universal in schools, when seeking employment, and when seeking professional licensure and certification.

Just in the past few years, the validity of selected-response, pencil-and-paper measures of school achievement and skill-based knowledge has been widely challenged. The use of selected-response tests of students' subject-matter achievement in every state in the Union has exposed the "Lake Wobegon Effect," wherein the average achievement of every state, and of most large school systems, was found to be above the national average (Cannell,1987). Studies of the consequential validity of standardized tests of student achievement (cf., Smith,1989 and Shepard,1990), have shown that such tests drive curricula and instruction in the nation's schools. Lorrie Shepard (1990), Lauren Resnick (1990) and others have demonstrated that currently-used standardized achievement tests are grounded in the cognitive psychology of the 1970's, where it was assumed (falsely) that hierarchical learning of basic skills was prerequisite to students' acquisition of the higher-order analytic skills that are the real objective of formal education.

A new day has arrived in the testing of students. At the national level, through the New Standards Project (Resnick & Resnick, 1992) and the National Assessment of Educational Progress (NAEP; cf., Applebee, Langer, Mullis, Latham & Gentile, 1993), and in a number of states (e.g., California, Vermont, Maryland; cf. Aschbacher, 1991), test-like devices that attempt to measure students' skills and abilities to perform various tasks directly are under development or in use. These "performance assessment" measures are changing the landscape of student assessment. Rather than presenting students with printed multiple-choice questions that require

selection of a correct option from among those presented, performance assessments require students to construct responses to a wide range of problems. Performance assessments have been termed ‘‘authentic assessments, ” since they often provide tasks that require problem-solving skills, and present challenges that are thought to model realistic applications that students will face later in life.

Introduction

In the enthusiasm that surrounds the new assessment methodologies, it must be realized that the demands of sound educational and psychological measurement are ever-present. Assessments of students’ abilities must satisfy professional measurement standards, as exemplified in the 1985 *Standards for Educational and Psychological Testing,* regardless of the testing and measurement method used. If the potential of performance assessments of students is to be realized, such assessments must be demonstrated to yield measurements that (1) are sufficiently reliable to support the selection or classification of individuals or the evaluation of aggregates of students at local, state, or national levels, (2) validly support inferences concerning the achievements, aptitudes, and performance capabilities of those assessed, (3) fairly, and in an unbiased way, reflect the abilities of those assessed without regard to gender, race, ethnic group membership, or economic circumstance, and (4) when performance standards must be set, support the classification of examinees into decision-relevant categories (e.g., worthy of admission, certification, licensure, graduation, etc.) or their labeling as "basic," "proficient," "advanced," etc.

Unfortunately, much of the available methodology for assessing the psychometric quality of measurement instruments was developed with selected-response tests in mind. In large part, the applicability of this methodology to performance assessment measures is either obviously limited, questionable, or untested. For example, currently-used methods for establishing performance standards on tests either demand pencil-and-paper, multiple-choice test items (Nedelsky, 1954) or presume the use of selected-response achievement test items (Angoff, 1971; Ebel, 1972; Jaeger, 1982).

This chapter is concerned with one component of the psychometrics of performance assessment: methods for setting performance standards on large-scale performance assessments of students. This topic is essentially unexplored in the educational and psychological testing literature although some practice-based experience has been accumulated in conjunction with NAEP, the teacher assessment program of the National Board for Professional Teaching Standards, and some statewide assessments of students. Even though methodology for setting performance standards on performance assessments is embryonic, we attempt in this chapter to define critical issues, review what is known and being done, and define a research agenda that can guide future inquiry. As the reader of this chapter will learn, the state of the art is far from a state of grace. Much work remains to be done.

Performance standard-setting is a judgmental process. Its results are therefore subjective by definition. Of necessity, the outcomes of any performance standard-setting process are arbitrary[1] but they need not be capricious[2]. Effective standard setting requires that qualified persons make reasoned judgments in response to questions that are understandable and within their ability to judge, following well-structured efforts to inform them about the nature and consequences of the alternatives they face.

Many of the conclusions derived from two decades of research on performance-standard setting for selected-response tests can be expected to generalize to performance assessments. First, performance standards rarely occur naturally (one relatively rare exception is in some performance assessments in the military, where correct completion of a task; e.g., cleaning a rifle correctly, is directly observable and can be scored "pass" or "fail"). Far more often, the boundary between performances judged to warrant classification as a passing performance or eligibility for valued rewards, and those that do not, result from a process of negotiation or arbitration. Performances judged acceptable or worthy will most often differ in magnitude, rather than kind, from performances judged to be unacceptable or inadequate. It is this conclusion, as much as any other, that has led some to label all standard-setting processes as capricious (cf., Glass, 1978).

Second, performance standards are method-dependent. Specification of performance that warrants graduation, certification, classification as proficient, etc. depends to a substantial degree on the method used to elicit judgments from those who set performance standards (for a summary of literature on this point, see Jaeger, 1989).

Third, those who set performance standards cannot be assumed to be trustworthy judges of the quality of the methods they have used. Virtually all surveys of panelists who set performance standards have concluded that such panelists are "somewhat confident" to "very confident" about the quality of their resulting standards, regardless of method they have used and subsequent empirical evidence concerning the coherence and consistency of the performance standards they have set.

Finally, widely-used performance-standard-setting methods presume the existence of an underlying interval scale of performance on the test or assessment for which standards are to be set. This conclusion follows from the first (above), from the typical practice of averaging the standard-setting recommendations of members of standard-setting panels, and from the use of parametric statistical procedures in evaluating the precision, consistency, and coherence of performance standards. As will be noted below, performance assessment exercises are typically multidimensional in their measurement characteristics, are rarely exchangeable, and are, therefore, far less likely to satisfy the local independence assumptions of unidimensional scaling models. Such exercises violate the assumptions that underlie typical performance-standard-setting methods.

The Structure of This Chapter

The balance of this chapter consists of three major sections and a summary. In the next section, under the heading "Some Fundamental Issues in Performance Standard-Setting," we describe typical institutional and individual applications of performance standard-setting so as to delineate the contexts in which new methods of setting performance standards for performance assessments would be used. We also build the case that the performance-standard-setting methodology of the past is not applicable to a wide range of performance assessments in these contexts.

In the third section of the chapter we provide prominent examples of early efforts to establish performance standards for performance assessments. These examples include the National Assessment of Educational Progress, the assessment of highly accomplished teachers by the National Board for Professional Teaching Standards, and a variety of statewide assessments of students.

The fourth section of the chapter, under the heading "Some Technical Dilemmas," addresses such issues as the artificial polytomization of the performance continua to which standard-setting methods must be applied, the method-dependent nature of performance assessments, and a host of other technical issues that must be addressed in setting performance standards on performance assessments.

In our summary and conclusion we indicate important methodological and procedural issues that are yet to be addressed by the known methodology of setting performance standards for large-scale performance assessments, thus providing a rudimentary road map to needed research.

Some Fundamental Issues in Performance Standard-Setting

Among the various reform movements affecting our Nation's schools, two thrusts in particular have major implications for the tests used in student assessments. First, the education community, including teachers and curriculum specialists as well as politicians and educational policy makers are diligently working to develop standards-for discipline-oriented curricula, for assessment, and for the delivery of educational services. Greatest attention to date has focused on the curriculum standards specifying what students must know and be able to do to be considered proficient in various subject-matter areas. The first such standards were released by the National Council of Teachers of Mathematics in 1989. Since that time, at least a half-dozen other professional organizations and collaboratives have made substantial progress toward similar documents for other academic disciplines. Currently, curriculum standards in civics, geography, history, foreign languages, science, social studies, and the arts have been released. However, few of these documents contain formal performance standards, although some provide a few illustrations.

Simultaneously, there has been growing concern that assessments should directly address more complex types of learning than those measured by conventional tests (National Council on Education Standards and Testing, 1992). In addition, assessments should require applications of knowledge and skills to "real-world" situations faced by individuals at work, in their own lives, and as citizens of a community. These more "authentic" forms of assessment require task performance in contexts that reflect the realities of everyday life, rather than simply providing isolated bits of information or knowledge (Wiggins, 1993). Taken together, these two reform thrusts signal the need for assessments that emphasize how well students can perform particular tasks vis-à-vis subject-matter standards. Further, the more that such performance assessments focus on the complexities of problem-solving and reasoning rather than declarative knowledge, comprehension, or routine application skills, the more consistent they will be with the goals set forth in the current curriculum standards.

If performance assessments are to be used in accountability-based strategies for promoting systemic educational reform, the issue of designing methods for setting performance standards on performance assessments must be addressed. A number of national and local assessment enterprises are beginning to examine this issue, notwithstanding the inherent complexities associated with developing curriculum standards, with developing performance assessments that accurately measure proficiency on those standards, and with establishing clear and concrete understandings of what evidence students should provide to show that they have met the standards.

How Performance Standards Are Used

Before tackling the numerous challenging technical and procedural issues implicit in this dramatic change in educational testing approaches, it is important to understand the various purposes of educational testing and potential uses for the information provided by performance standards. For this discussion, we have identified two broad classes of applications: those pertaining to *institutions* and those pertaining to ***individuals***.

The two applications can differ considerably regarding the amount of detail required for an individual. The needs for information about institutional quality generally are more global in nature, because such data typically are used for making relatively long-term as compared to day-to-day decisions. Sampling of individuals can be used in assessing institutions, given that it is not necessary to obtain broad-based information about the performance of each individual. By assessing different aspects of curriculum standards for various representative groups of students, information can be aggregated across the different groups to develop a picture of educational performance for an entire institution or system.

In contrast, information for making decisions about individuals often needs to be comprehensive and timely because it may directly influence an important and relatively-immediate decision about selection or placement of the individual into an educational program or institution.

There are similarities in the ways that institutional and individual performance assessment results can be used. Both types of assessments can be used for formative or summative purposes. Feedback about success with particular aspects of curriculum can be used on an ongoing basis to make adjustments in an institutional program or pedagogical approach, or to develop specific remedial steps for individual students. Beyond such formative uses, assessments or tests also can provide the basis for final decisions about institutions, programs, or individuals. For example, schools or school systems may participate in summative evaluations to receive accreditation and students may need to pass a test to graduate from high school, or become licensed in a profession.

Institutional Applications of Performance Standard-Setting

Description for Public Information. Foremost among the uses of performance standard-setting is the public's right to know about the quality of our educational system and its products. Society as a whole has a vested interest in knowing whether today's graduates have the knowledge and understanding necessary to contribute to an informed citizenry and whether they can be relied upon to set enlightened policy for the future. Employers deserve to know whether students have the knowledge and skills required to meet daily production needs or for our nation to compete in a global economy. Parents want to know that their children are receiving a high-quality education, and students themselves should be interested in whether they are becoming well-prepared for adult life.

Public Accountability. Beyond the basic right to know, those responsible for funding and providing oversight for various aspects of the nation's education system need information to monitor schools' performance. Whether for the nation, a school-district, or a school building, student achievement results can be used to provide guidance on how resources invested in education might be augmented or used differently. Information about students' competencies can

also be used in formulating approaches to improving education. With the publication of the National Commission on Excellence in Education report, *A Nation at Risk*, and general discontent from employers and parents about the effectiveness of America's schools, the call for accountability information has been growing for more than a decade. In fact, this widespread concern has contributed directly to the movement toward a standards-based education system, including assessments to monitor students' progress in attaining performance standards. The role of assessments increasingly includes helping to define and guide instruction to make it more effective. The content of assessments impresses on students, school staff, and parents the importance of tested subject matter and associated expectations for learning.

Program Evaluation. One particular kind of accountability information concerns assessments about special programs designed to address the needs of particular groups of students. For example, special programs have been developed for students with limited proficiency in English, students determined to be at increased risk of academic failure due to disadvantaged socioeconomic backgrounds (e.g., ESEA, Chapter 1), or to provide special opportunities for high-achieving and talented students. Information needs about particular programs relate to decisions about sets of practices or groups of students within the education system, school district, or school.

Individual Applications of Performance Standard-Setting

Selection. One of the better known uses of educational testing in the United States is for selection for college entrance, with many high school students taking either the SAT or ACT as part of the college admission process. However, students may also take tests to be admitted to special programs or schools, such as those designed for students with special aptitudes in science or the performing arts. Although neither the ACT nor the SAT program recommends the use of fixed performance standards for college admission, the practice is, unfortunately, widespread.

Classification. Sometimes related to selection, the major purposes of classification assessments are to match students' performance levels against pre-set criteria to assign students or apportion them to different treatments, curricula, or programs. For example, a student might receive compensatory education, be placed in an advanced mathematics program, or be retained at the same grade level based, in part, on test results. Such assessments can document accurately each student's progress at the end of an extended period of instruction. These data can be used to profile students' strengths and weaknesses in particular areas, serving a diagnostic as well as a placement function.

Certification. Certification serves a public function, enabling individuals to demonstrate to others that a certain standard of performance has been met or that a certain set of skills has been mastered. That is, the individual who has been certified will be known to have completed a particular course of study and to have demonstrated a specified criterion of performance. The public then expects the individual to be able to perform particular tasks or functions. Although the Advanced Placement program involves testing for the receipt of credit for particular first-year college courses and some school districts test students to certify levels of minimal competence as part of receiving a high school diploma, certification is used less as part of the public education system in the United States than in some other countries (e.g., the French Baccalaureate, the Abitur in Germany, and the British Public Examinations). Test-based certification, however, is widely used as a professional endorsement of candidates'

employability in the United States. For example, licensing or certification occurs for nurses, accountants, architects, lawyers, and in many other occupations and professions. Also, the National Board for Professional Teaching Standards is developing procedures for advanced teacher certification. Currently, efforts are underway to develop standards for many more occupations, and employers have suggested certification as part of the public education process.

In *Learning a Living, the* Commission on Achieving Necessary Skills (SCANS) for effective job performance recommended that students should develop a cumulative résumé containing information about courses taken, projects completed, and assessment results. As students met the standards set for specific skills, that mastery would be noted on their résumé. When students had met the standards across courses and SCANS competencies, their résumé would show that they had been awarded the Certificate of Initial Mastery (CIM).

Why New Methods Are Needed for Setting Performance Standards on Performance Assessments

Because standard-setting permeates many aspects of life, including the occupational licensing and certification processes referenced above, the difficulties associated with applying such procedures to educational uses may not be apparent. For example, standards are very much part of athletics (e.g., performance ratings for the Olympics, life-saving badges in swimming, certification for the ski-patrol, or a black-belt in the martial arts), and they are central to guiding aspects of daily life such as transportation (e.g., the safety of aircraft) and regulation of food and drugs (e.g., classifications of over-the-counter drugs, standards for safe fish or meat, and even classifications of the "hotness" of canned chili peppers.)

Although, as is discussed in the next section, experience with performance standard-setting can be gained from these efforts as well as initial attempts in educational contexts, many challenges remain. Generally, problems arise from attempts to make decisions based on very few tasks that often measure very diverse and discrete accomplishments. Consensus has not been reached on how to best aggregate performance information across those tasks to make an overall judgment either for individuals or institutions.

Making Judgments Based on Few Tasks. This issue relates to the breadth of material covered by a content standard and the number of performances required to make confident judgments about students' mastery of the material (i.e., to classify a performance, based on a performance standard). For example, to determine whether students can use addition of whole numbers in appropriate contexts they could be given several problems to solve involving various applications of whole-number addition. Perhaps after successfully completing 10 or so such problems (provided they presented diverse stimuli within the definition of whole-number addition), some confidence could be achieved in generalizing to the domain and judging the students' proficiency in applying whole-number addition to real-world problems.

Unfortunately, however, performance standards eventually will need to address far larger domains, such as geometry, algebra, or even mathematics as a whole. To date, the work in setting curriculum standards has defined each domain as a vast and diverse network of knowledge, skills, and understandings. Also, especially as students grow older, tasks need to be far more complex, encompassing students' abilities to select among a repertoire of approaches, apply an effective strategy, and reach an appropriate solution (such as, actually building a machine that works). These two goals—the need for large numbers of tasks in order to assess

students reliably and the need for broad tasks, so as to sample a domain adequately-are difficult to reconcile, in that a broader array of performance tasks would yield far greater confidence in determining students' success in reaching a performance standard. Yet, the complexities of performance tasks, in terms of resources and time, preclude administering large numbers of them to individual students.

Performance tasks consume more resources to develop, administer, and evaluate, and require more energy and time on the part of the students participating in assessments. While educators may be more receptive to devoting fiscal resources as well as students' and teachers' time for participating in realistic assessments, there are reasonable limits. Performance assessments generally contain relatively few, complex exercises. Educators are often appropriately reticent to make high-stakes decisions about students' educational careers based on a limited number of data points.

Each Performance Typically Entails Degrees of Success. Also, assuming for the moment a curriculum standard and agreement about a performance task measuring it, students will undoubtedly demonstrate differing degrees of success in accomplishing the task. For example, drawing on national, state, and school-district experiences across the past several decades in moving from multiple-choice to direct writing assessment, students' essays generally are not evaluated as "right" or "wrong." Usually, a rating scale is employed whereby each essay is classified as more or less successful according to categories on an established scale. Summarized across students, performance can be described at various levels of the rating scale. Sometimes, there is a cut-point on the scale indicating that minimal competence has been reached, or that performances on the essays demonstrate higher-level writing proficiency, or both. Often, however, performance results are reported in terms of a scale, without making judgments about the performance in relation to a performance standard. Such polytomous results often are complicated to aggregate. First, decisions may need to be made for each task about what constitutes "minimal" or "excellent" work, as well as what level of response is required to meet the standard being measured. It is interesting that work across states, conducted by The New Standards Project, showed considerable agreement in ranking students' work, but far less unanimity in where cut-points fell (Linn, et al., 1991).

Historically, often because of resource limits, writing assessments only contained one task and student classification decisions were therefore difficult to make. More recently, however, research is showing that generalizability among performance tasks is very tenuous, and this underscores the importance of including a number of tasks. As previously stressed, one of the problems with performance assessments is including enough tasks for adequate domain coverage. The aim is to increase the number of tasks so as to increase the likelihood of accurately measuring performance across a domain of interest. This, then, leads to the problem of aggregating performance information across tasks. Because tasks measure different accomplishments, it may be likely that examinees' scores across tasks will be diverse. While a profile can lead to better understanding of student performance for diagnostic purposes, it would complicate making an absolute judgment about students' overall performances.

To be judged proficient in a domain, do students need to be proficient on each performance task? Or perhaps, is it sufficient to be proficient on most but not all of the tasks? Or should the overall judgment be based on an average across tasks? Each of these options requires a different approach to performance standard-setting.

Lack of Score Transferability. Finally, given the complexities of the tasks used in performance assessments and the difficulty of evaluating students' performances on the tasks, it is not surprising that developing equivalence across tasks is a problem. Not only will an individual's performances differ from task to task, but some tasks simply will be more difficult than others. One reason for having performance standards is to attempt to establish comparability among the performances of students in different jurisdictions. That is, theoretically, parents of students in one state and school district could compare their children's performances to students' achievements in another state or country, or employers and colleges could make comparisons across students. Further, the nation and school systems could monitor trends across time in attaining curriculum standards. Yet, a score on one task often means little in predicting the score on a second, seemingly related, but different task, either for individuals or for groups of students.

Developing comparable standards across performance assessments appears to be the most problematic venture of all. There is, first, the problem of designing appropriate performance tasks in terms of content standards, and identifying student work on the tasks that exemplifies success in meeting the standards. Implementing such assessments under consistent conditions and evaluating resulting performances reliably pose enormous operational challenges. Subsequently, difficulties arise in generalizing from performances on a few tasks to performance in an entire domain of content and in aggregating results across tasks in defensible ways.

Prominent Examples of Current Practice

Although the modem metamorphosis of performance assessment has only recently come on the scene, a number of such programs are operated in contexts that require the establishment of performance standards. Here we review a few prominent examples of performance assessment programs and describe methods that have been used to set performance standards. In providing these descriptions, we are not necessarily endorsing the practices they represent. As we have already noted, the problem of setting performance standards for complex performance assessments is challenging, and current solutions may not withstand careful scrutiny. Nonetheless, it is useful to see where we are before setting out on a new journey.

How Performance Standards Are Set for NAEP

The 1992 NAEP national and trial state assessments in reading, mathematics, and writing included a variety of exercise formats. Standard-setting activities were modified to accommodate various scoring procedures used in NAEP, and these depended on item format. Table 1 summarizes exercise format types by subject and grade. Multiple-choice (MC) items as well as most short constructed-response (SCR) items were scored dichotomously, that is, '1 = right' and '0 = wrong'. However, extended constructed-response (ECR) items were scored using rubrics appropriate to the item or prompt, on an ordered scale, generally from '1' to '4' or '1' to '6', with higher scores representing progressively better performance.

In the case of dichotomously-scored exercises NAEP employed a *modified Angoff* (Angoff, 1971) procedure for establishing cut-scores. Since the policy framework for standard setting required three standards for each grade level, *Basic, Proficient, and Advanced*[3], a broadly-representative group of judges was asked to estimate the probability of a correct response at the three levels for each of the dichotomously-scored items. The probability estimates were averaged over items and judges, and then multiplied by 100, to yield an estimated probability in the percent-correct metric for each achievement level.

Table l—Number of Items, by Exercise Format in the 1992 NAEP Assessments by Grade

Grade	Math		Reading		Writing		
	MC	SCR	ECR	MC	SCR	ECR	ECR (prompts)
4	119	59	5	42	35	8	9
8	118	65	6	56	63	16	11
12	115	64	6	59	67	19	11

Sources:— NAEP 1992 Mathematics Report Card for the Nation and the States; NAEP 1992 Reading Report Card for the Nation and the States.

Legend

MC = multiple choice items (dichotomously scored)
SCR = short constructed response items (dichotomously scored)
ECR = extended constructed response items (polytomously scored)

The Angoff method is an approach to standard setting employed in a variety of settings and for a variety of purposes. According to Berk (1986), the method is superior to most competing procedures, and is generally straightforward in the tasks required of judges and the interpretability of results. Although the Angoff method rarely has been used to set multiple test standards, it could be modified to provide such a result. The Angoff method has been employed almost universally to set standards for individual rather than aggregated test results, but could be modified to accommodate the latter task. Both of these 'modifications' were introduced into the NAEP approach to meeting the National Assessment Governing Board's policy requirements of three test standards, and to accommodate the particular characteristics of the NAEP assessment, which yields only aggregated scores.

For polytomously-scored exercises, NAEP employed an *Exemplar Response* method of setting performance standards in 1992. Judges were provided a random selection of examinee responses to each ECR question, stratified by each score point. So that judges would not be unduly influenced by the score scale, the scores for each paper were not shared with the judges during the initial round of ratings. Judges were required to select from the distribution of responses, two papers that exemplified minimally acceptable performance for each performance level (i.e., *Basic, Proficient,* and *Advanced). The* scores associated with these papers were averaged over judges and equated to test-score functions for the set of items.

There are few, if any, methodologies described in the literature for establishing standards on tests composed of polytomously-scored exercises. In considering the options for NAEP, three methods were considered: (1) set standards on the scoring rubric (i.e., have standard-setting panelists select test standards from among the values of the polytomous score scale); (2) estimate distributions on the score scale; and (3) select exemplar responses. Setting standards on the scoring rubric would have limited the range of choices for standard-setting panelists, and would

have linked the standards to what students *can do* and less to what students *should do* (the latter were an integral part of the NAGB's policy definitions). There was also a concern that, if selecting among score-scale values, panelists would be less likely to take differences in prompt-difficulty into account. The second option, estimating distributions on the score scale, was not used because it was feared that this task might be too difficult for panelists to complete. Therefore, the most viable option seemed to be the *exemplar response* method, which preserved the desirable features of (1) being less normative in its orientation, (2) not being bounded by the values on the score scale, and (3) minimizing potential systematic bias that clearly could become an issue if panelists' selected standards from among score-scale values. In reading and mathematics, judges' estimates from the dichotomous and polytomous exercises were combined using an information weighting procedure (Luecht, 1993b) before mapping the final estimates onto the NAEP scale. In the NAEP writing assessment, the judges' proposed standards were mapped directly to the NAEP scale.

The State of Technical Debate

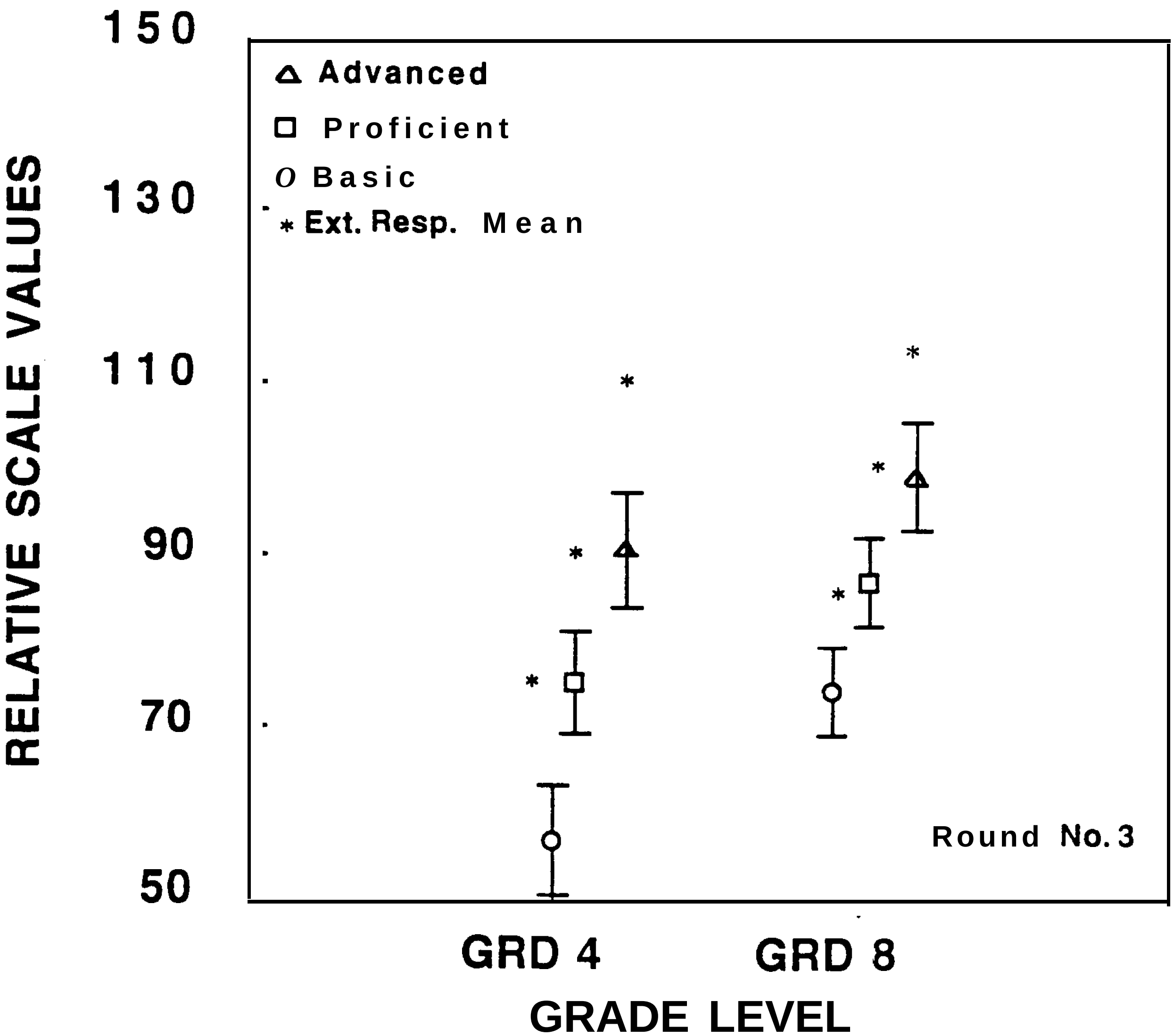
A host of technical issues came to light during the NAEP standard-setting process. The remainder of this section will attempt to describe several of these issues and offer some plausible explanations.

Discrepancy in Cut-Scores. Early in 1992, a standard-setting pilot study was conducted to explore various aspects of the process, including selection and training of panelists, programming and on-site data analysis, impact of feedback on panelists, achieving inter-rater and intra-rater consistency, and the standard-setting task itself. The 1992 reading assessment was selected for the pilot study because it employed three item formats, multiple-choice, short constructed-response, and extended constructed-response exercises. The first two were dichotomously scored, the third was scored using a rating scale from '1' to '4' (polytomous scoring). In order to examine the standard-setting approaches used with both scoring procedures, a separate standard was estimated for each. Figure 1 displays the results for the reading pilot study.

It is clear that performance standards for the polytomous exercises were set substantially higher than were standards for the dichotomous exercises. There was no empirical evidence to suggest why this was the case. And since it came to light only after the standard-setting meetings were over, it was not possible to interview panelists to ascertain why this happened. However, several plausible explanations were offered by the technical advisors to the project.

First, panelists were given a limited set of exemplar responses (16) from which to make their selections of papers that best represented performances at the thresholds of *Basic, Proficient, and Advanced.* In addition, the papers represented a rectangular distribution of responses, with roughly equal numbers at each score point. It is possible that there were no papers in the set at the threshold of each level, which had the effect of forcing the selection of higher test standards. A second plausible explanation revolves around the guessing factor. Guessing was virtually impossible on the polytomous exercises, and panelists were not instructed to take guessing into account in making their probability estimates on the dichotomous items (thus reducing their proposed standard on the dichotomous items). A third explanation offered was method effect. The task required of panelists on the polytomous exercises was very different from that required in the modified Angoff procedure applied to the dichotomous items. A fourth hypothesis was

Figure 1


150
130
110
90
70
50
RELATIVE SCALE VALUES
△ Advanced
□ Proficient
O Basic
* Ext. Resp. Mean
Round No. 3
GRD 4
GRD 8
GRADE LEVEL

that examinees simply don't try as hard to do well on the polytomous exercises as they might on the dichotomous items. Students are less motivated to respond when an exercise presents a production task. Therefore, panelists' estimates were not 'off target', but instead, examinees were not behaving the same way toward the polytomous exercises as the dichotomous items. Finally, the observed differences may be due to the possibility that the polytomous exercises (or at least as they are perceived by panelists) are assessing a different dimension of reading skills (or math or writing skills), in which case the multidimensional nature of the data is not being taken into account through the unidimensional IRT scaling used in NAEP.

Although these hypotheses remain untested due to lack of data, the concern is serious since NAEP continues to evolve and move toward more 'authentic' item types. The results suggest that more research is needed before adopting any particular methodology for setting standards on polytomous exercises. In 1992, this discrepancy in performance standards had minimal impact on the overall standards adopted by NAEP in mathematics, since the item pool contained a limited number of extended constructed-response exercises. The impact was slightly greater for reading, since about 15 percent of the item pool consisted of ECR items. In writing, however, the impact was so great that the achievement levels reported here were not officially adopted by the National Assessment Governing Board (Applebee, et al., 1994).

Alternative Methods for Polytomous Exercises. Luecht, (1993c) has done some work in the area of polytomous-item standard-setting simulations. A series of simulations was conducted to compare three methodologies for setting standards on polytomous exercises: *Poly-Angoff1*, *Poly-Angoff2*, and the *Exemplar Response* approach that had been used on the 1992 NAEP. Although other methods could have been devised, these three were selected for further research because they appeared to hold some potential for improving on the modified Angoff procedure used with dichotomous items. The simulations were also used to demonstrate the effect of various mapping procedures (discussed below).

The Poly-Angoff1 method asked panelists to estimate the percent of examinees at the *Basic* level who would score at each score point on a polytomous-item score scale (i.e., from '1' to '6'). The *Poly-Angoff2* method generated performance standards by summing the *Poly-Angoff1* percentage ratings for any score greater than or equal to '2', since a score of '1' was always incorrect, and a score of '2' or higher was always at least partially correct. (This method was also pilot tested during the standard-setting meetings in 1992.) According to Luecht (1993c), the *Poly-Angoff1* ratings were found to be slightly more robust than those resulting from the *Poly-Angoff2* method. The *Poly-Angoff2* method, the *Exemplar Response* method, and a *hybrid* method that combined elements of both were explored during the 1994 standard-setting pilot tests for NAEP.

Mapping Panelists' Ratings to the NAEP Scale. In reading and mathematics, mapping panelists' estimates (in the percent-correct metric) to the NAEP scale was a fairly straightforward procedure. After the panelists' estimates were aggregated across exercises, the means were numerically mapped to the NAEP test characteristic curve (TCC). This procedure, called the TCC method, resulted in a NAEP scale score for each of the nine performance standards mapped. The TCC procedure was also employed in determining the writing cut scores in 1992. These are shown in table 2.

For purposes of comparison, and at the urging of technical advisors to the NAEP standard-setting project, a second method was also used to determine the writing cut scores. This

Table 2—NAEP Writing Cut Scores Based on Test Characteristic Curve Methodology by Grade

Grade	Basic	Proficient	Advanced
4	187.0	282.4	354.0
8	222.2	299.9	371.4
12	246.4	332.9	416.3

Source: *ACT Writing Report, 1993.*

Table 3—NAEP Writing Cut Scores Based on Plausible Values Methodology by Grade

Grade	Basic	Proficient	Advanced
4	203.3	240.2	264.0
8	240.5	275.6	297.6
12	266.4	298.7	319.0

Source: *ACT Writing Report, 1993.*

procedure capitalized on the plausible values (PV) technology used in NAEP to generate unbiased group estimates of proficiency. In the PV method, examinee scores corresponding to the *exemplar responses* (papers) selected by the panelists were identified and aggregated to determine the performance standards. Since individual NAEP scores were not computed, the *plausible values* for each examinee became the score proxy. These values were left unweighted to avoid other assumptions involved in using weighted data. The PVs were summed over responses and judges to arrive at the performance standards that are shown in table 3.

The question to be asked here is, which method is appropriate for generating performance standards? There are conceptual differences between the two methods. It might be suggested that the Test Characteristic Curve (TCC) method operates from a definition *of Advanced,* for example, as *consistently* Advanced, since the performance standard is based on the total test score function. On the other hand, one might feel that the PV method operates from a definition of *Advanced, as typically* Advanced, but *not necessarily always* Advanced, since the performance standard is based on the plausible values of individual examinees, who may or may not achieve the same score on all exercises presented to them. This distinction suggests a policy question, "Is the *Advanced* standard defined as *always Advanced*, or is it defined as *typically*

Advanced?" Perhaps the same distinction could be made at the *Proficient* and *Basic* levels as well. Would this suggest lower performance standards instead of higher ones?

From a technical perspective, there is evidence to suggest that the apparent differences in the performance standards resulting from the TCC and PV methods shown in tables 2 and 3 are largely a function of *regression bias*, overpredicting in some cases, and underpredicting in others (Luecht, 1993a). However, consensus has not been reached on this point, and more research is needed.

Finally, it must be mentioned that a National Academy of Education Panel on the NAEP Trial State Assessment has suggested more-fundamental flaws in the methods used to set performance standards on NAEP. The Panel holds that item- or exercise-based procedures for setting performance standards produce invalid results, and that methods based on judgments of the quality and adequacy of entire test booklets are needed (The National Academy of Education, 1993). Commissioned critiques of the National Academy of Education report by Cizek (1993) and Kane (1993) dissent from these conclusions as well as the policy recommendations contained in the report. Here again, debate continues.

How Performance Standards Are Set by the National Board for Professional Teaching Standards

The National Board for Professional Teaching Standards (NBPTS) is developing performance assessments that are used to identify classroom teachers who have achieved advanced levels of accomplishment in their profession. These teachers are eligible to receive National Board Certification. The performance exercises used by NBPTS elicit complex responses from teachers, including written evaluations of the videotaped classroom performance of another teacher; submission of selected samples of the work of the teachers' students, together with reflective analyses of the quality of that work and the link between the teachers' instruction and the students' performances; videotapes of the teachers engaged in specified activities with their students, coupled to descriptions of plans and objectives and self-evaluative analyses of instructional success; written responses to problems involving appropriate structuring of curriculum; and demonstrations of effective use of instructional resources, among others.

Teachers' responses to exercises are scored by trained assessors using highly-structured rubrics and scales that yield scores ranging from "1-" to "4+" for each exercise. Training of assessors is extensive, and includes demonstrations of scoring competence and consistency.

Two approaches to performance standard-setting have been evaluated by NBPTS. The first, termed *Iterative, Judgmental Policy Capturing,* elicits panelists' implicit standard-setting policies by fitting mathematical models to their responses to profiles of the performances of hypothetical candidates for certification on a set of exercises. The second, termed the *Multi-stage Dominant Profile Procedure,* elicits panelists' explicit standard-setting policies. These procedures are described in turn.

The Iterative, Judgmental Policy Capturing Procedure. In this procedure, standard-setting panelists receive extensive training on the nature of the exercises that compose an assessment and on the meaning of each possible score level associated with the scoring of each exercise. In distinct contrast to the standard-setting methods described earlier for NAEP, the methods used

by NBPTS do not assume that exercises are scored on a unidimensional scale. Thus the score scales used with different exercises can be unique.

During a pilot study, a panel of 20 middle-school teachers applied the Iterative, Judgmental Policy Capturing Procedure (JPC) to performance exercises contained in the NBPTS Early Adolescence Generalist teacher assessment. Panelists applied Judgmental Policy Capturing to 200 profiles composed of the scores of hypothetical candidates for National Board Certification on six of the exercises that compose the Early Adolescence Generalist assessment. Panelists responded independently to graphical profiles of the performances of hypothetical teachers by specifying for each profile, whether the overall performance of a teacher with that profile should be considered "Deficient" (1), "Competent" (2), "Accomplished" (3), or "Highly Accomplished" (4). The National Board (NBPTS) has asserted that only "highly accomplished" teachers will receive its certification. A sample profile is shown in figure 2.

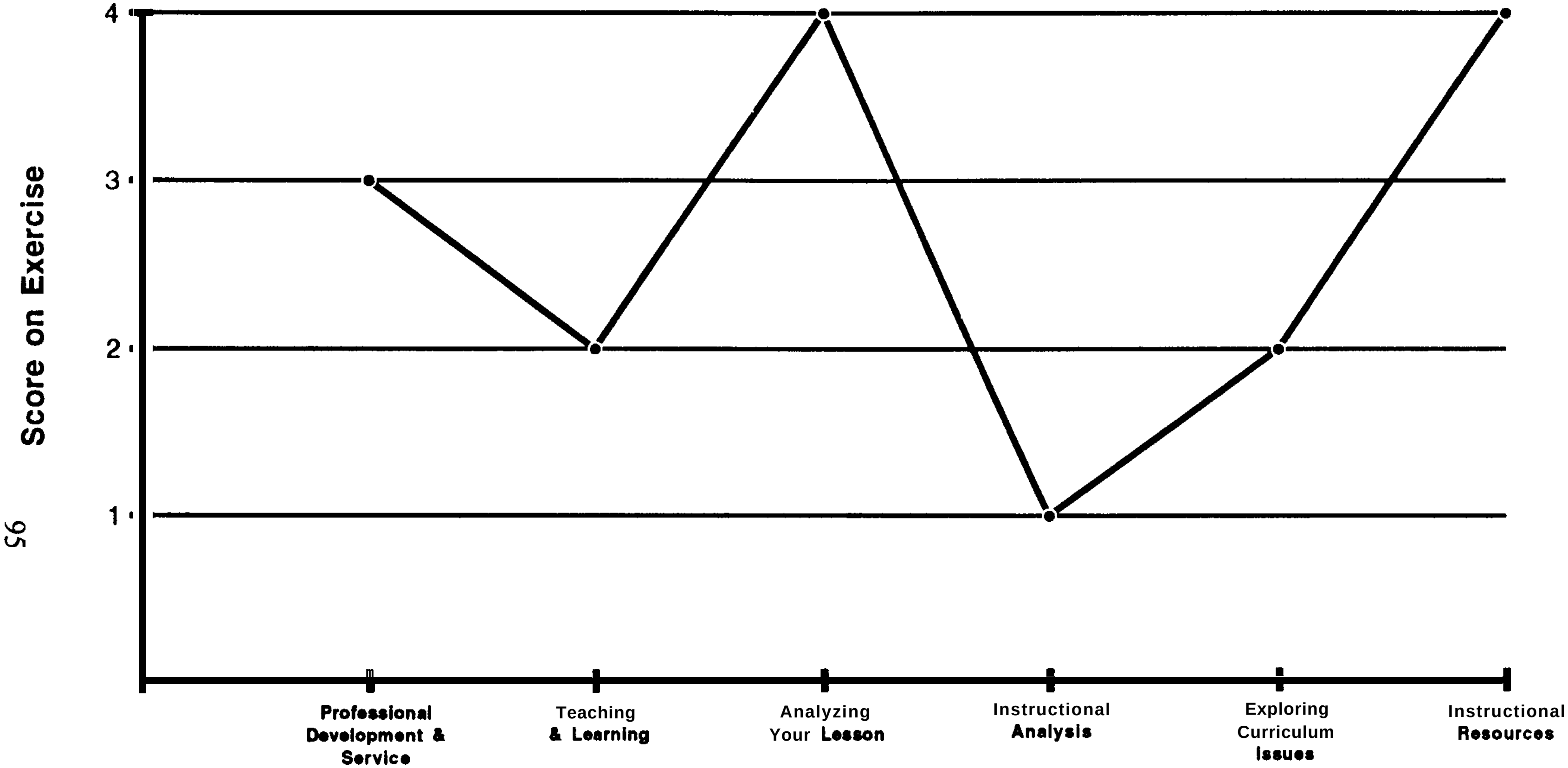
The relative weights each panelist applied to the six exercises in specifying the overall performances of hypothetical candidates were estimated using each of two analytic models. The first, a compensatory analytic model, was derived using ordinary least squares multiple regression. The second, a conjunctive analytic model, was derived using a multiplicative equation which, following a logarithmic transformation, also was evaluated through the use of ordinary least squares multiple regression. Thus two analytic models of the "captured policy" each panelist used in specifying performance standards and the magnitudes of relative weights that should be applied to the six exercises of the Early Adolescence Generalist assessment were created. Presumably, one could use a panelist's captured policy to predict the overall score (s)he would give a candidate for National Board Certification who earned any profile of scores on the six Early Adolescence Generalist exercises.

After each panelist had responded independently to a set of 200 profiles and models of panelists' policies had been created, the distributive weight each panelist applied to each of the six exercises was estimated. The distributive weights derived for each panelist summed to unity and could, therefore, be interpreted as proportions. For example, if a panelist assigned a distributive weight of 0.4 to the "Teaching and Learning" Exercise and a distributive weight of 0.2 to the "Analyzing Your Lesson" Exercise, we would conclude that (s)he placed twice as much weight on "Teaching and Learning" as on "Analyzing Your Lesson."

Distributions of the distributive weights each panelist applied to each of the six exercises were compiled and graphed. These graphs were provided to each panelist, with their own profile of distributive weights superimposed in graphical form, so they could see the weights their fellow panelists applied to each exercise, and determine whether their own distributive weights were similar to, or substantially different from, those applied by others. This form of normative feedback was followed by instruction on how to read and interpret the graphs, and an opportunity for panelists to discuss with their colleagues the relative importance they attached to each of the six exercises and their rationales for doing so. Initial discussion took place in groups of three or four panelists and was followed by a discussion by the entire panel.

After panelists had reviewed their own patterns of distributive weights and engaged in discussion with their colleagues, they once again completed a Judgmental Policy Capturing exercise. They again responded to the original set of 200 profiles, specifying for each, an overall performance rating on a 1-to-4 scale, for a teacher with the indicated profile of performances on six exercises of the NBPTS Early Adolescence Generalist assessment.

Figure 2. Simulated Profile of a Candidate's Performances on Six Early Adolescence Generalist Assessment Exercises



Overall Evaluation of Candidate →	A (1) Deficient	B (2) Competent	C (3) Accomplished	D (4) Highly Accomplished

Fill the appropriate bubble (A, B, C or D) on the bubble sheet to indicate your judgment.

The model-fitting analyses described earlier were applied to the judgments panelists recorded in their second round of Judgmental Policy Capturing. The results of this new model fitting were used to estimate a new set of distributive weights for each panelist and, in addition, a table for each panelist that indicated the outcome of applying that panelist's captured policy to 114 hypothetical profiles of candidate performance. For each profile of performances on the six exercises, a panelist was able to see whether the application of her/his policy would result in the candidate being certified by the National Board or being denied certification.

Prior to a final round of Judgmental Policy Capturing, the results of which were analyzed as described earlier, panelists were once again given an opportunity to compare their profiles of distributive weights to graphically-presented distributions of the weights applied by their fellow panelists. They were also given an opportunity to discuss with fellow panelists, their rationales for applying greater weight to some exercises than to others.

Since an objective of the standard-setting study was to produce equations that would represent the judgments of all panelists, weighted averages of individual panelists' regression weights were computed to form a single, integrative equation. Again, separate equations were produced using each of two analytic models.

Some Results. Distributions of distributive weights that resulted from applying a compensatory analytic model to panelists' judgments elicited during the third round of Judgmental Policy Capturing are shown in figure 3. Extreme values, the first and ninth deciles, and the quartiles of the distributions are portrayed in box-and-whisker charts. Results obtained using a conjunctive analytic model were so similar to those obtained with a compensatory model that little would be gained from displaying them. They are, therefore, not shown.

The results of Judgmental Policy Capturing indicate that panelists assigned great weight to candidates' performances on the "Teaching and Learning" exercise in setting an overall performance standard (see figure 3). In contrast to traditional standard-setting procedures, which set a unidimensional test standard that is based on a simple or transformed sum of candidates' performances on all assessment exercises, JPC acknowledges the multidimensional nature of candidates' performances. It bases an assessment standard on profiles of candidates' performances.

A critical issue in the use of Judgmental Policy Capturing concerns the reasonableness of representing the captured policies of a group of panelists through a single policy that represents the central tendency of their distribution. At issue is the assumption that the policies of individual panelists can be regarded as randomly-differing observations drawn from a single distribution. To explore this issue, the vectors of weights that composed the 20 panelists' policies were subjected to nonmetric multidimensional scaling analyses (Schiffman, Reynolds & Young, 1981). In addition, the matrix of correlations among overall performance scores assigned by the panelists to the 200 profiles of exercise scores were factor analyzed, using panelists as variables (Cattell, 1966).

The multidimensional scaling results did not reveal the presence of definable clusters of panelists, suggesting instead, a single distribution of distributive weights across all panelists. The results of factor analyzing the 20-by-20 matrix of correlations among panelists' overall ratings of profiles of exercise scores clearly demonstrated that their policies could be considered random

Figure 3

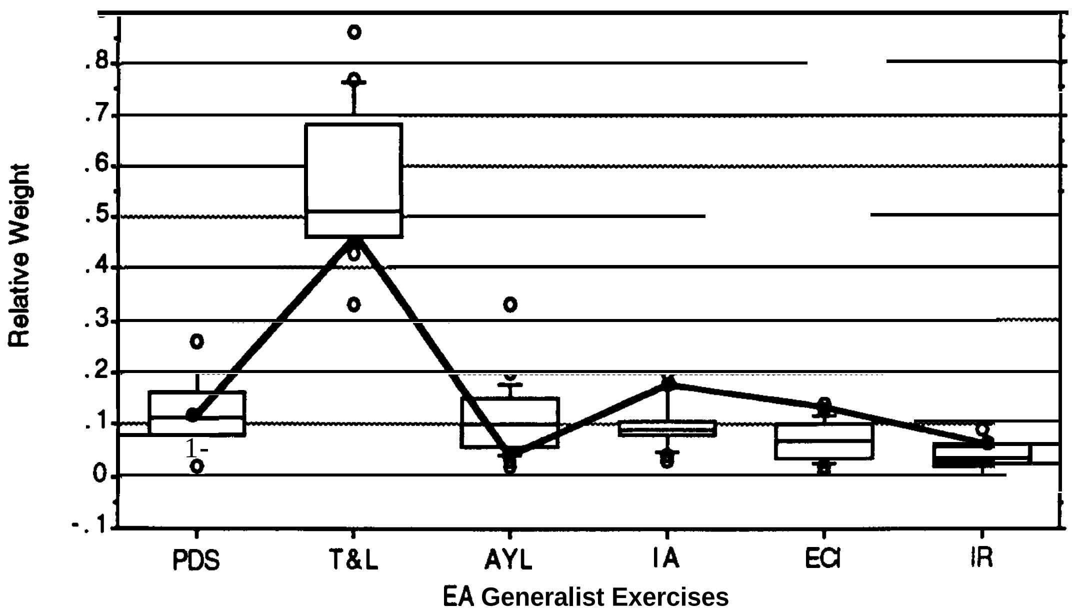
DISTRIBUTIONS OF RELATIVE WEIGHTS ASSIGNED TO EA GENERALIST EXERCISES BY <u>YOU</u> AND BY THE ENTIRE STANDARD-SETTING PANEL

JUDGE 20

PERCENT OF VARIANCE EXPLAINED = 71%

THE RELATIVE WEIGHTS YOU ASSIGNED TO EXERCISES

0.11 Professional Development and Service (PDS)
0.46 Teaching and Learning (T&L)
0.05 Analyzing Your Lesson (AYL)
0.18 Instructional Analysis (IA)
0.14 Exploring Curriculum Issues (ECI)
0.07 Instructional Resources (IR)



observations from a single distribution. A principal components analysis resulted in single factor that accounted for approximately 82 percent of the variance among panelists' policies.

These results were used to justify the derivation of a single standard-setting policy for the entire panel of 20 teachers. Analytically, the compensatory policy equations derived for individual panelists were combined through a weighted averaging procedure. The weights used were the inverses of estimated variances of estimates of corresponding regression coefficients. Through this procedure, more-precise estimates of regression coefficients were given greater weight than were less-precise estimates. Following the computation of a weighted average regression question, a linear transformation was applied so that a profile containing scores of four on each of the six exercises of the Early Adolescence Generalist assessment would yield an estimated performance score of four and a profile containing scores of one on each of the six exercises of the Early Adolescence Generalist assessment would yield an estimated performance score of one. This transformation was consistent with the desired result that perfect score performance on all exercises, should, with certainty, lead to National Board Certification.

In light of the scale definition noted earlier, the transformed JPC model is interpreted as follows: Any candidate with a predicted overall score, following transformation, that equals or exceeds 3.5 would receive National Board Certification. Candidates with lower predicted overall scores would not be certified. When this rule was applied to the Round 3 results of judgmental policy capturing, it was found that candidates who scored well on the "Teaching and Learning" Exercise would be certified, even though they performed less well on several other exercises in the assessment. For example, a candidate who earned scores of four on the "Teaching and Learning" Exercise, and the "Professional Development and Service" Exercise, but scores of three on the other four exercises would be certified, as would candidates who earned scores of four on the "Teaching and Learning" Exercise, the "Professional Development and Service" Exercise and the "Instructional Analysis" Exercise, scores of three on the "Exploring Curriculum Issues" Exercise and the "Instructional Resources" exercise, and a score of two on the "Analyzing Your Lesson" Exercise.

The pilot study demonstrated the feasibility of using Judgmental Policy Capturing as a method for setting performance standards on assessments composed of complex, multiply-scored performance-based exercises. The results of the study indicated (1) that the Judgmental Policy Capturing procedure and associated instructions used could be applied readily by panelists following a short period of instruction, (2) that panelists could provide judgments in response to as many as 200 six-dimensional profiles in a reasonably short time without undue fatigue (on average, panelists required 41 minutes to respond to 200 profiles), (3) that most panelists could achieve high levels of consistency in their responses to profiles, as reflected in coefficients of multiple determination between 0.70 and 0.87 (median of 0.79), and (4) that iterative use of Judgmental Policy Capturing reduces the variability among the captured policies of standard-setting panelists and leads to a reasonable degree of consensus around a collective policy. The results of the study also confirm that panelists applied differential importance weights to the exercises that composed the NBPTS Early Adolescence Generalist assessment package. Thus differential weighting of exercises in determining a performance standard (at least on the NBPTS assessment) appears to be essential.

The Multi-stage Dominant Profile Procedure. *The* Multi-Stage Dominant Profile procedure consists of four stages: individual policy creation, revision of individual policies following

discussion and reconsideration, application of individual policies to profiles of candidate performance (administered through a tailored Judgmental Policy Capturing procedure), and reaction to summarized policies created through detailed analyses of the application of individual policies to profiles of performance on an assessment. Whereas the Judgmental Policy Capturing procedure requires panelists to react to profiles of candidate performance that are presented to them so that their standard-setting policies can be inferred, the Multi-stage Dominant Profile procedure requires panelists to formulate and state the "bottom-line" standard-setting policies they would apply in screening candidates for some valued reward, such as National Board Certification. Panelists' policy statements are then refined through discussion, and by analyzing their application to profiles of performance that challenge or confirm panelists' initial policies. A more-detailed description of the procedure follows. This procedure was applied in a recent pilot study conducted on behalf of the National Board for Professional Teaching Standards. Twenty middle-school teachers applied the procedure to six performance assessment exercises from the National Board's Early Adolescence Generalist teacher assessment.

Policy Creation. During the first stage of the Multi-Stage Dominant Profile procedure, individual panelists stated policies that defined the lowest levels of performance on the six exercises that, in their view, warranted National Board Certification. Depending on the nature of their policies, panelists constructed one or more profiles of performance that, in their view, characterized the performance of a "just barely certifiable" Early Adolescence Generalist teacher. Following extensive instruction on the exercises for which performance standards were to be set, the meaning of each possible score on each exercise, and the nature of compensatory and conjunctive standard-setting policies, panelists were given ten blank profile forms on which to convey their "bottom-line" profiles. Panelists were told that only one profile would be necessary to convey a strictly conjunctive standard-setting policy, but that many "bottom-line" profiles could exemplify a compensatory policy. Panelists were asked to use as many profile forms as they needed to illustrate their standard-setting policy, and to write a paragraph that explained their policy.

Discussion and Reflection. Following their initial specification, panelists were invited to describe and discuss their standard-setting policies. Each panelist was provided with an opportunity to present her/his policy to the standard-setting panel and to describe the rationale underlying the presented policy. Stated policies were summarized on an overhead projector so that all could note how their own standard-setting policies compared to those of their colleagues. Panelists were then asked to reconsider their initial standard-setting policies in light of the discussion and what they had learned about their fellow panelists' recommendations and rationales.

Panelists then worked independently, just as they had originally, to specify and illustrate their own standard-setting policies. Panelists completed as many illustrative profile forms as they needed (ten blank forms were given to each panelist, and no panelist used all ten) and again wrote one or two paragraphs describing their standard-setting policies.

Construction of Confirmatory and "Challenge" Profiles, and Summary Policies. Immediately following the standard-setting session, panelists' policies were analyzed to identify areas of agreement and disagreement. Many panelists provided one or more illustrative profiles, and all prepared brief descriptive statements of their policies. Content analysis of panelists' written policy statements and their accompanying profiles revealed several sophisticated policies

for determining candidates' qualifications for National Board Certification. These policies reflected what panelists had learned during their engagement in the Judgmental Policy Capturing procedure.

Using analyses of candidates' stated and illustrated policies, 105 profiles of performance on the six exercises that composed the NBPTS Early Adolescence Generalist assessment were constructed. These performance profiles either conformed to the "bottom-line" standard-setting policies that one or more panelists had specified or "challenged" those policies in the sense of conforming to all but one or two policy specifications. For example, if a panelist had specified in a "bottom-line" policy that, to be certified, a candidate could earn no more than two scores of two, must earn a score of four on a particular exercise, and must earn scores of three or more on the remaining exercises, a profile that contained three scores of two but otherwise conformed to the policy specifications was constructed. These profiles were used to examine the consistency with which panelists applied their "bottom-line" policies to the profiles of performance of hypothetical candidates. Panelists were sent the 105 profiles one week following the standard-setting session and were asked to state for each profile, whether a candidate with the indicated profile of performance should, or should not, receive National Board Certification as an Early Adolescence Generalist teacher.

Panelists were also sent three policy statements that represented the standard-setting policies of major subgroups of the standard-setting panel. They were asked, for each statement, whether or not they would accept the adoption of the statement as the National Board's standard-setting policy for certifying Early Adolescence Generalist teachers.

Some ***Results***. Three standard-setting policies reflected the judgments of almost all panelists. These policies combined compensatory and conjunctive elements, in that panelists specified threshold levels of performance for some exercises, but allowed high scores on other exercises to compensate for lower-levels of performance on the balance of the assessment. When asked whether they would accept the adoption of these standard-setting policies by the National Board, 15 of 20 panelists endorsed a single policy, and four more endorsed minor variants of the majority policy. The final panelist created an entirely new policy that was at substantial variance with the judgments held by the balance of the panel.

Analyses of the responses of panelists to the 105 profiles of candidate performance created to confirm or challenge their policies revealed some variation in the consistency of panelists' application of their stated policies. However, on average, panelists were 77 percent consistent.

The results of the pilot study of standard-setting were very encouraging. The Judgmental Policy Capturing procedure was regarded by panelists as an important and helpful learning experience and as a useful strategy for eliciting their standard-setting policies. Following the Judgmental Policy Capturing procedure with the Multi-Stage Dominant Profile procedure permitted panelists to create explicit policies that accurately conveyed their judgments and to test and revise those policies when confronted with the ideas of their fellow panelists and with explicit profiles of candidate performance. The convergence of 75 percent of the panelists around a single standard-setting policy was a particularly encouraging result.

How Standards Were Set for the Maryland Statewide Assessment

Educators in Maryland have developed descriptions of what students should know and be able to do by the year 2000[4]. These descriptions together are called the Maryland Learning Outcomes. Content areas include reading, mathematics, science, social studies, writing and language usage.

The State of Maryland also developed a new test, the Maryland School Performance Assessment Program (MSPAP) to assess the ability of students to master the Maryland Learning Outcomes. The MSPAP is one element in Maryland's School Performance Program reform initiative. The test is designed to measure students' ability to apply what they have learned to real-world problems, and to relate and use knowledge from different subject areas.

Administered in grades 3, 5, and 8, the MSPAP measures critical thinking and problem-solving abilities. MSPAP standards are set for satisfactory and excellent levels of performance. By the 1996 MSPAP test, 70 percent of students in each school are expected to meet satisfactory standards.

The Standard-Setting Process. MSPAP standards were set through a process that relied on the professional judgment of expert educators. The standards were reviewed and refined by a council that included representatives of business, educational organizations, and the legislature. Opportunity for comment was provided for the public at large.

In May 1993, a MSPAP Standards Committee was convened to set proficiency levels for satisfactory and excellent performance. The Standards Committee consisted of 17 content area and assessment experts from 11 school systems within Maryland and the Maryland State Department of Education. The Committee reviewed background material on the MSPAP assessments, including the learning outcomes that are assessed in MSPAP; examples of actual test questions, scoring criteria, and student responses; and 1992 MSPAP school performance results.

MSPAP scores fall into five levels of proficiency, with Level 1 the highest. The Standards Committee selected proficiency Level 3 or higher as the standard for satisfactory performance, and proficiency Level 2 or higher as the standard for excellent performance in each of the content areas/grade levels. The Standards Committee recommended a percentage range of students that must meet these levels by 1996. These ranges are: satisfactory—70 percent to 80 percent; excellent—20 to 30 percent.

The standards proposed by the Standards Committee were reviewed by the MSPAP Standards Council, including 12 persons from local boards of education, the Maryland State Teachers' Association, the Maryland Business Roundtable, the business community, students, parents, and the State Legislature. This council set specific goals for schools, within the ranges recommended by the Standards Committee, as to the percentage of students expected to meet the satisfactory and excellent levels of performance. The Standards Council recommended that by 1996 at least 70 percent of students should perform at the satisfactory level and 25 percent should perform at the excellent level. By 2000, 95 percent of students should perform at the satisfactory level in each content area, and 50 percent should perform at the excellent level. These standards will be reviewed after the 1996 test to see if adjustments are needed.

The State Superintendent of Schools reviewed the recommendations of the Standards Council and briefed local superintendents. The State Board of Education reviewed the Superintendent's recommendations toward the end of May. Regional meetings for the public were held in June and July, 1993, and the State Board of Education included the new MSPAP standards in the regulations covering public school standards in July, 1993.

Use of MSPAP Results. The main purpose of MSPAP is to assess school performance, not the performance of individual students. Individual students' scores, however, are available to be used in conjunction with other information to inform parents of their child's performance in particular content areas.

How Standards Were Set for the Kentucky Instructional Results Information System

In 1990, Kentucky's General Assembly adopted the Kentucky Education Reform Act (KERA)[5]. KERA established goals for the educational system and provided a procedure for assessing progress toward the goals.

One of the six education goals is:

"Develop students' ability to:

1. Use basic communication and mathematics skills;
2. Apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, and practical living to situations they will encounter throughout their lives;
3. Become self-sufficient individuals;
4. Become responsible members of a family, work group, or community;
5. Think and solve problems in a variety of situations; and
6. Connect and integrate experiences and knowledge from all subject- matter fields."

A Council on School Performance Standards was created to further define these cognitive goals. The Council developed 75 performance goals, or "valued outcomes," which the State Board of Education approved.

Kentucky proceeded to develop a new assessment program to measure students' performance on the valued outcomes. The assessment program was to include multiple-choice and open-ended items, performance tasks that can be administered in one class period, and longer tasks or projects that can be placed in a student's portfolio. The valued outcomes were to be the foundation for new models of school curricula. NAEP frameworks were also used to supplement the valued outcomes where needed. KERA stipulated that a transitional assessment would occur at Grades 4, 8, and 12 in 1991-92, with a full-scale, primarily performance-based assessment to be implemented soon thereafter. Information on development of the tests, transition to performance-based assessment, and the mechanics of test administration is available in a *Technical Report*.

Setting Performance Standards. It was decided to set four levels of performance standards, termed "novice," "apprentice," "proficient," and "distinguished." Considerable thought went

into deciding the appropriate number of levels as well as their names. Too many levels would make differences between adjacent categories somewhat trivial and would be hard to communicate. Too few levels would result in setting the standard for minimum competency at a very low threshold, removing the motivation for improvement.

Standard setting for writing portfolios was completed first by a Writing Advisory Committee consisting of approximately 10 teachers per grade level (Grades 4, 8, and 12), staff from the Kentucky Writing Project, university faculty, and members of the Kentucky Department of Education. The committee reviewed student portfolios, selected sample portfolios that exemplified the standards, and completed the scoring guide to operationalize the standards. These were incorporated into a set of training materials distributed to teachers statewide.

The first year's test was transitional in nature, containing common questions (given to all examinees) and matrix-sampled questions (each examinee is tested on only some of the items). Standards were set first on the common items; standards for the matrix-sampled questions were derived from data describing the relationship between the two types of questions.

For the common items, standard-setting was done for performance (open-ended) items only, and was based on examining actual student work. Scores for performance items ranged from O for no response or an irrelevant response, to 4 for a "complete, insightful, strongly supported response, demonstrating in-depth understanding of relevant concepts and processes. "

Scoring guides for performance items were translated into standards. Standard-setting panelists were first trained on the writing scoring guides. After training, they were provided with several examples of scored writing samples. The committee members were then given several papers that had already been assigned scores. For each question, they independently decided which score reflected each level of performance. They then debated their decisions as a group, and reconsidered their decisions until consensus was reached. They arrived at a decision that scores of O and 1 "generally indicated Novice work," 2 was Novice, 3 was Proficient, and 4 was Distinguished.

Committee members were given several papers at the upper and lower ends of the range for each of the above score values, to discuss whether these papers cast doubt on their score-to-standard translation scheme. Their decision rules remained unchanged following this discussion. Finally, the committees were given information about the percentage of students attaining each performance level, so they could see that very small percentages of students achieved the Proficient or Distinguished levels. In spite of this, they decided not to change their original decision rules.

Standards for the matrix-sampled questions were derived from those established on the common items by using a regression approach. The methodology is fully explained in the *Technical Report*. The procedure resulted in 60 predicted performance-level scores. Cutoffs for each performance level were then determined so that the distribution of students classified into the various performance levels on the matrix-sampled items was as similar as possible to the distribution of students classified on the common items. Once cutoff points were established, each student was assigned to a performance level based on his/her score on the matrix-sampled items.

Standards were set for performance events by assuming that the distribution of scores across levels would be as close as possible to those obtained for the common items. Regression analysis

was not used for these events, however. Students with the highest scores on the performance events were placed in the Distinguished category until the percent of students at this level approximately equaled the percentage of students classified as Distinguished on the common items. Students with lower scores on performance events were subsequently placed in the Proficient, Apprentice, and Novice categories in approximately the same percentages as with the common items.

Use of Assessment Results. Assessment results were used to determine the percentage of students in each school whose scores warranted placement in each of the four categories—Novice, Apprentice, Proficient, and Distinguished. Schools received rewards or sanctions based upon their improvement or failure to improve over time. The 1991–92 test established each school's baseline, from which targets for future improvement were determined. Test results collected during the 1992–93 and 1993–94 school years were used to measure each school's progress toward its target and to establish a new baseline for future performance targets.

Some Technical Dilemmas and Procedural Issues

Although constructive work on setting performance standards for assessments containing performance exercises is ongoing, and some progress has clearly been made, many dilemmas and procedural concerns remain. In this section we review some troubling issues and unresolved dilemmas, thus, by implication, defining a research agenda for the coming years.

Artificial Dichotomization (Polytomization) of Performance Continua

The central problem in setting cut-points or standards for levels of performance is that they establish artificial dichotomies on a continuum of proficiency (Shepard, 1984). Proficiency in a population does not occur at discrete, easily recognizable levels. It is, rather, a matter of degree. The problem is how to treat the gray areas around the cut-points, since a certain proportion of examinees just above or just below a cut-point will almost inevitably be misclassified due to measurement error.

Most judgmental methods of setting performance standards begin with the concept of a minimally competent examinee to assist in establishing a passing threshold for a test. Judges must either decide which of the response categories in a multiple-choice item the minimally competent examinee would reject (Nedelsky, 1954), which items he/she would answer correctly (Jaeger, 1982), or the probability, for each item (or parcel of similar items), that he/she would give the correct answer (Angoff, 1971). These methods result in an informed but arbitrary (i. e., determined by judgment) performance standard. Examinees achieving a score above this standard are classified as competent; those below are classified as incompetent. The problem arises in the cases of false positives and false negatives (i. e., those persons who are misclassified on the basis of the test versus their actual abilities). Taking measurement error into account in establishing a performance standard can partially diminish the likelihood of committing one kind of error or the other.

Other methods combine judgment with empirical data on actual test performance. The borderline-group method furnishes judges with information on the actual performance of a group of examinees believed to be on the borderline of minimal competence. The median score of this group is adopted as the standard. In the contrasting-groups method, groups of "masters" and "non masters" are identified and given the test. Frequency polygons are developed for the scores

of each group. The standard for passing the test is set at the score where the frequency polygons cross. If it is found that an unreasonably high percentage of examinees fail the test, the threshold can be lowered. The degree to which the threshold is lowered can be based on how high the stakes are for a test. For example, if the test is used to qualify examinees for promotion to the next grade, or for professional licensing, the cost to examinees who fail even though they are truly qualified is high. The passing threshold should be lowered accordingly, but not so much that it is no longer a standard of minimally acceptable performance.

Ways to minimize the misclassification problem include:

- Consideration of the examinee population, not only the test items, in setting standards;
- Supplementing judgments with empirical data (e.g., whether the test actually distinguishes the competent from the incompetent);
- Use of mathematical models for adjusting cut points to minimize classification errors. Such models include the binomial error model developed by Subkoviak, Huynh's beta-binomial model, Bayesian decision-theoretic models, and linear loss models;
- Pilot tests of examinations, with revision of performance standards as appropriate; and
- Occasional review of performance standards as conditions change (e.g., when curriculum changes).

Shepard recommends seeking insight from several models of standard-setting. Evidence from different approaches, combined with normative evidence (actual performance), should be used to establish a range of plausible performance standards.

Performance Standards Tend To Be Method-Dependent

It is widely agreed that different methods of setting performance standards yield different results. The problem seems thus far to be intractable. The inconsistencies are inherent in the way judgments are made according to the various methods, as well as in the composition of panels of judges. In a review of research, Jaeger (cited in Phillips, 1993) summarized the results of 32 comparisons culled from 12 separate studies. He found that almost six times as many students would fail when using one standard-setting method rather than another. Inconsistencies arose even when different methods were used under seemingly identical conditions.

As already suggested, the results of several methods should be used in any given study, and these should be tempered with statistical techniques, use of empirical data, and knowledge of the examinees. Consideration of the use to which the performance standard will be put should also have weight in setting the cut-point(s). Shepard comments, "It is important to see that it is the nature of the problem rather than lack of effort which prevents us from finding the preferred model for standard setting" (Shepard, 1984, p. 187).

How Large Should a Standard-Setting Panel Be?

Although we hold no belief that performance standards are population parameters waiting to be discovered, we recognize that any standard-setting panel is composed of a sample of individuals, selected from a population of persons who might have been invited to serve. We also recognize that a second panel, selected using the same procedures as the first and presented with identical instructions and stimulus materials, would likely recommend a somewhat different

performance standard. Examinees whose performance is in the region bounded by the performance standards recommended by two different panels are victims of the process used to select a panel. They would have received some valued reward (admission, certification, classification as "advanced" learners, etc.) had their performances been judged against the performance standard specified by one panel, but denied those rewards if judged against the performance standard of the second.

Determining the number of panelists who should recommend a performance standard is not unlike other sample size problems. The larger the number of panelists, the more stable will be the resulting performance standard across samples of panelists. Since all popular standard-setting procedures have as a final step, computing the central tendency of a distribution of recommendations provided by individual panelists, the standard error of the final performance standard will vary inversely as the square root of the number of panelists. A seemingly reasonable criterion for determining the number of panelists to be sampled would be to produce a standard error of the mean recommended standard that was small, compared to the standard error of measurement of the assessment for which a standard was sought. Since variation in the recommended performance standard and measurement error variance contribute additively to errors of examinee classification, the size of the standard-setting panel should be such that the standard error of the mean (or median) recommended performance standard does not add appreciably to the overall standard error. Jaeger (1990) noted that if the standard error of the mean recommended performance standard is no more than a fourth as large as the standard error of measurement of the assessment, the overall standard deviation of errors due to sampling of panelists and the unreliability of the assessment would only be 3 percent larger than the standard error of measurement alone. By this criterion, random error due to sampling of panelists would contribute minimally to errors of classification of examinees. To apply this criterion in practice, both the standard error of measurement of the assessment for which a standard is sought, and the standard deviation of panelists' recommended performance standards must be known. Standard errors of measurement are readily estimable. As experience with setting performance standards on assessments composed of performance exercises is gained, typical standard errors of recommended performance standards should become available as well.

Who Should Compose a Standard-Setting Panel?

Most professionals in the performance testing arena recommend that a standard-setting panel should be representative of a broad range of interested and informed groups. Broad representation avoids the selective bias that might result if only one concerned group (e.g., only members of the business community or only school administrators) constituted a standard-setting panel. It is important to identify the relevant groups from which judges should be chosen, and to sample the judges from these groups in such a way that they are representative of the relevant population. However, drawing on research concerning the nature of expertise (Chi, Glaser & Farr, 1988; Minsky & Papert, 1974; Chi, Glaser & Rees, 1982; Glaser, 1987), Jaeger (1991) argued that judges who recommend or set performance standards should be experts in the domain assessed by the test and in the roles sought by successful examinees. He recommended the use of standard-setting panels that were composed of experts, and not necessarily representative of stakeholder populations.

The issue here is quite important. Democratic values support the argument that a performance standards should be set through a process of negotiation that involves all

constituencies having an interest in the qualifications of those who would "pass" the test or assessment on the basis of the performance standard set. In the case of achievement testing, those constituencies would likely include parents, students, teachers, school administrators, potential employers of students, civic leaders, university admissions officers, and so on.

The principal counterargument is based on the contention that standard-setting panels should be restricted to those with solid understanding of (1) the test or assessment for which a standard is to be set, (2) the cognitive demands imposed by the assessment, and (3) the capabilities of examinees, given the opportunities they have been provided to learn the material tested. Lacking firm understanding of these issues, a standard-setting panel could readily set performance standards that were unrealistic, unfair to students and educators, and unlikely to influence instructional improvement. The 1985 *Standards for Educational and Psychological Testing* require that the qualifications of standard-setting panelists be documented (Standard 6.9). This requirement suggests that legitimate interest alone should not determine the composition of standard-setting panels; expertise should be required as well.

There are some standard-setting initiatives where the term "expert" has been broadly defined to include not only those who have expertise in the content area (typically gained through formal academic preparation), but additionally, members of the non-academic community who use the skills and knowledge of the content area in their trade or profession. When performance standards were set for the National Assessment of Educational Progress, "expertise" was defined very broadly. The National Assessment Governing Board regularly has included, and now by statute (P. L.. 103-382, 1994) must include on its standard-setting panels, school administrators, parents, and concerned members of the general public.

The Appropriate Uses of Normative Information

The desire for criterion-referenced performance standards. The nation's governors and two Presidents have reached a consensus that there should be high expectations of what students should know and be able to do, and clear indicators of how students' actual performances measure up to these expectations. In other words, it is no longer enough to find out what students know and can do; new public policy demands an assessment of whether students' performance is "good enough." Therefore, criteria are needed for deciding what is "good enough," in the sense that students know what they should and are able to do what is expected of them.

In many applications of performance assessment, attempts are made to set absolute standards of performance, quite apart from current norm distributions of performance. The work of the National Board for Professional Teaching Standards is a case in point. The National Board has defined standards of "accomplished performance" in the teaching profession for more than a dozen teaching fields and will eventually define such standards for more than 30 teaching fields. These standards represent the National Board's vision of accomplished teaching performance and are not grounded in typical current practice. In no sense are these performance standards normatively based.

The need to ground performance standards in reality. The dilemma arises when it is recognized that standard-setting panelists must make use of normative anchors if they are to set performance standards at levels that are practical and defensible. Much of our judgment in life is normatively grounded, and it is difficult to establish absolute standards apart from our normative experience. For example, we regard men who are over 78-inches tall as "very tall" not because of

any absolute sense of the accumulation of inches of height that warrant that appellation, but because, in our normative experience, very few men achieve that height.

Normative data are essential in the standard-setting process for ensuring that performance standards will be appropriate to the population for which they are being set. Setting a performance standard in a vacuum, without reference to actual performance, could easily lead to thresholds that result in large classification errors. This would be particularly harmful in the case of high-stakes tests in which large numbers of examinees could be disqualified even though they are competent in the skills being tested. Even in group assessments, however, the standards would lack credibility if they did not reflect actual abilities. Such a situation could arise, for example, on a mathematics assessment if it was found that examinees who scored at an "advanced" level were actually poor mathematics students.

It was noted earlier that different methods of deriving performance standards yield different, sometimes widely divergent standards. A standard-setting process that used several methods as well as incorporating normative feedback to judges should result in more stable and realistic performance standards. Such normative data would help judges decide whether a particular standard actually separates examinees according to relevant skill levels, as desired. Judges should be given an opportunity to revise standards in light of empirical consequences. This iterative process might be repeated several times with more or different types of feedback. As Shepard observed, "The standard we are groping to express is a psychological construct in the judges' minds rather than in the methods ... Normative expectations are always the starting point, because norms are the source of validity evidence and ultimately determine our psychological absolutes" (Shepard, 1984, p. 188).

So the dilemma of wanting to set absolute standards coupled with the reality of knowing from experience that totally unrealistic standards may result remains. It seems best, in light of research on the effects of normative feedback on performance standards (cf., Busch & Jaeger, 1990) to temper the desire for absolutes with the reality that normative feedback to standard-setting panelists provides.

Essential Training of Standard-Setting Panelists

Defining minimally acceptable performance. Methods of setting performance standards often require panelists to formulate some idea of a minimally acceptable examinee. Then for each test item or exercise, panelists must estimate whether that hypothetical examinee would give the correct answer, or must estimate the percentage of such examinees who would answer the item correctly. In setting multi-level standards, panelists must repeat this process for each performance level. Some methods use an iterative process in which panelists are given empirical information on the difficulty of each item or exercise and/or share each others' ratings in the previous round.

These classical approaches to standard-setting are most useful when items are dichotomously scored rather than being scored along a polytomous scale. As noted earlier, performance assessments in which essays, portfolios, and extended constructed responses are used do not lend themselves to classical standard-setting methods. Further, some assessments now include both multiple-choice and performance items, resulting in differential standards being set. The picture is further complicated by the multidimensionality of performance items, in which more than one skill or area of knowledge is integrated. The task becomes overwhelming for judges, who must

keep the test frameworks in mind while minimizing the role of personal experience and intuition in their judgments.

The difficulty in employing methods centered on defining minimally acceptable performance underscores the advisability of incorporating empirical evidence of the validity of standards. The contrasting-groups and borderline-group methods described above may be helpful in validating standards that are derived through other methods (in the sense of providing convergent evidence). Bootstrapping is another recently-developed method, in which the performance standard is set at the point at which it has been empirically determined that examinees who score above the standard will, with high probability, succeed in the task being measured. Benchmarking is a variant of this approach. In benchmarking, the standard-setter attempts to find examples of high performance as an aid in setting the standard, as, for example, the scores attained by high-performing math students in other countries.

Training on the nature of the standard-setting task. It is very important that panelists be adequately trained in the process of setting standards that they are expected to use. They must understand fully, and internalize, the test or assessment framework before rating items or exercises as to what they measure and their levels of difficulty. Panelists must have clear definitions of the knowledge and skills expected of minimally acceptable examinees, or of examinees at each of multiple performance standards. Panelists must also understand fully the content domain to be assessed. Because performance standards are based on judgment, a final decision necessarily will be arbitrary. With the requisite training, however, the performance standard will not be capricious.

Sources of Error and "Adjustment" of Standard-Setting Recommendations

If a performance standard is set to the value recommended by a standard-setting panel, half of the examinees whose true performance level is precisely equal to the standard will fail the assessment, and half the examinees whose true performance level is minutely below the standard will pass the assessment. This is true because errors of measurement influence the observed performances of examinees and sampling errors in the selection of panelists influence the performance standard. These sources of error are independent of each other.

It is possible to adjust recommended performance standards for either or both of these sources of error, but doing so requires conscious consideration of the relative costs of false-positive and false-negative classification errors. An adjustment for both sources of error can be made by calculating the standard error of measurement of the assessment and the standard error of the mean recommended performance standard. These standard errors must then be squared and summed, and the square root of the sum of squares must be calculated. We will call this result the "total standard error."

If the recommended performance standard were to be increased by one total standard error, the percent of examinees with true performance just below the original performance standard who would pass, due to these errors, would be reduced from 50 percent to 16 percent (assuming normally-distributed errors). Thus the rate of false-positive errors would be reduced substantially. However, the rate of false-negative errors for those whose true performance was equal to the original standard would be increased to 84 percent. Lowering the original performance standard by one total standard error would have precisely the opposite effect. The

rate of false-negative errors would be diminished substantially, but the rate of false-positive errors would be similarly increased.

Whether performance standards should be adjusted to compensate for errors of measurement and sampling errors, and if so, in which direction, is a policy matter. Alerting policy-makers to the existence of the errors and to potential adjustments of performance standards that could be used to address the errors is the responsibility of professionals who recommend performance standards.

Summary

The need to establish performance standards is unlikely to diminish in the coming years. Education policy at the federal and state levels is firmly grounded in the notion that students and teachers should be held to high standards and that they should demonstrate their achievement of such standards. As long as policy-makers express the need for standardized evidence of student and teacher performance in terms of pre-established expectations and define public information needs similarly, those who establish performance standards will busily ply their trade.

In this chapter we have described the myriad ways performance standards are used and have addressed the need for new methods of establishing such standards for performance assessments of students and teachers. We have reviewed some strategies employed by those who have heretofore encountered the problem of setting performance standards on assessments that incorporate performance exercises—a literature that is principally fugitive and an area of inquiry that is embryonic. Finally, we have concluded with a review of technical dilemmas and procedural issues that plague all attempts to set performance standards, but are particularly important when standards are set for performance assessments. We hope that this latter section will be regarded as a challenge to those who would advance this important line of inquiry.

Notes

[1] The word arbitrary is used here in accordance with the second meaning provided in the Oxford English Dictionary (19701): "Relating to, or dependent on the discretion of the arbiter, arbitrator, or other legall y-recognized authority; discretionary, not fixed (p. 107). ” Even though judges used to recommend test standards rarely have legal authority to do so, those who set standards of performance typically have statutory authority to establish education policy. In the case of the National Assessment of Educational Progress, for example, the National Assessment Governing Board has exercised authority provided in the NAEP authorization to adopt performance standards that were recommended by a standard-setting panel and its contractor.

[2]The word capricious also used in accordance with the second meaning provided in the Oxford English Dictionary (1971): "Full of, subject to, or characterized by caprice; guided by whim or fancy rather than by judgment or settled purpose; whimsical, humoursome (p. 335)." Although the words arbitrary and capricious are sometimes used interchangeably, their connotative meanings are quite different. Reflective consideration and a well-defined judgment process often underlies arbitrariness, but never capriciousness. This distinction is central to the contrast made here.

[3]Proficient is defined as competency over challenging subject matter, *Basic* as partial mastery, and *Advanced* as superior performance.

[4]The following information about Maryland's MSPAP is taken from a Memorandum of July 27, 1993 from Nancy S. Grasmick, State Superintendent of Schools, to Members of the State Board of Education, and appendices.

[5]Information about Kentucky's performance assessment standards is derived from Kentucky Instructional Results Information System. 1991-1992 Technical Report, Kentucky Department of Education, Frankfort, KY, January 1993.

References

American Psychological Association (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd. ed.). Washington, DC: American Council on Education.

Applebee, A.N., Langer, J. A., Mullis, I. V. S., Latham, A. S., & Gentile, C.A. (1994). *NAEP 1992 Writing Report Card.* Washington, DC: National Center for Education Statistics.

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

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Busch, J. C. & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Cannell, J. J. (1987). *Nationally normed achievement testing in America's public schools: How all fifty states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago, Rand-McNally.

Chi, M., Glaser, R., & Farr, M. (1988). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

Chi, M., Glaser, R., & Rees, F. (1982). *Advances in the psychology of human intelligence,* vol. I (pp. 17-76). Hillsdale, NJ: Lawrence Erlbaum.

Cizek, G. (1993). Reactions to National Academy report, "Setting Performance Standards for Student Achievement." Washington, D. C.: National Assessment Governing Board.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R. (1987). Thoughts on expertise. In C. Schooler & W. Schaie (Eds.), *Cognitive functioning and social structure over the life course* (pp. 81-94). Norwood, NJ: Ablex.

Glass, G. V (1978). Standards and criteria. *Journal of Educational Measurement, 15, 237-261.*

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed.). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1990). Setting standards on teacher certification tests. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: assessing elementary and secondary school teachers.* Newbury Park, CA: Sage Publications.

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice, 10,* 3-6, 10, 14.

Kane, M. (1993). Comments on the NAE Evaluation of the NAGB Achievement Levels. Washington, D. C.: National Assessment Governing Board.

Linn, R. L., Kiplinger, V. L., Chapman, C. W., and LeMahieu, P. G. (1991, December). Cross-State Comparability of Judgments of Student Writing: Results from the New Standards Project.

Luecht, R.M. (1993a). Some Technical Issues and Results of the 1992 Scaling of the Achievement Levels in Writing." In *Setting achievement levels on the 1992 National Assessment of Educational Progress in Writing: a final report.* Iowa City, IA: American College Testing.

Luecht, R.M. (1993b). Interpolating theta for the partial credit model. In *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing. A technical report on reliability and validity.* Iowa City, IA: American College Testing.

Luecht, R.M. (1993c). Using IRT to improve the standard setting process for dichotomous and polytomous items. Paper presented at the annual meeting of the National Council for Measurement in Education, Atlanta, GA.

Minsky, M. & Papert, S. (1974). *Artificial intelligence.* Eugene, OR: State System of Higher Education.

National Academy of Education (1993). *Setting performance standards for student achievement- A report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels.* Stanford, CA: The National Academy of Education.

National Academy of Education (1993). *Setting performance standards for student achievement: Background studies.* Stanford, CA: The National Academy of Education.

National Council on Education Standards and Testing (1992). *Raising Standards for American Education.* Washington, DC: (author).

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics.*. Reston, VA: (author).

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Office of Technology Assessment (1992). *Testing in American Schools: Asking the Right Questions* (Washington, DC: (author).

Phillips, G. (1993). *Methods and issues in setting performance standards.* Washington, DC: National Center for Education Statistics, unpublished.

Public Law 103-182. (1994). Improving American's Schools Act of 1994. Washington, D.C.

Resnick, L. (1990, April). Psychometricians' beliefs about learning: Discussion. Presented at the annual meeting of the American Educational Research Association, Boston.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for education reform. In B. R. Gifford & M. C. O'Connor (Eds.). *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction.* Boston: Kluwer.

Schiffman, S. S., Reynolds, M. L. & Young, F. W. (1981). *Introduction to multidimensional scaling.* Orlando, FL: Academic Press

Shepard, L. A. (1990). Psychometricians' beliefs about learning. Presented at the annual meeting of the American Educational Research Association, Boston.

Smith, M. L. (1989). The *Role of External Testing in Elementary Schools.* Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.

The Secretary's Commission on Achieving Necessary Skills (1992). *Learning A Living: A Blueprint for High Performance.* Washington, DC: U.S. Department of Labor.

U.S. General Accounting Office (1993). *Education achievement standards: NAGB's approach yields misleading interpretations.* Report No. GAO/PEMD-93-12. Washington, DC: U.S. General Accounting Office.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity, Phi Delta Kappan, November 1993.

Fairness in Large-Scale Performance Assessment[1]

Lloyd Bond
University of North Carolina, Greensboro

Pamela Moss
University of Michigan

Peggy Carr
National Center for Education Statistics

Defining Fairness

In the context of assessment, fairness is integrally related to the concept of validity. Concerns about fairness are, in large part, concerns about the "differential validity" (Cole and Moss, 1989) of an assessment for any person or group of concern. However, concerns about fairness extend beyond the consensual boundaries of validity, which encompass the soundness of an assessment-based interpretation or action, to include concerns about the soundness of the educational or social policy of which the assessment is a part. In this chapter, we consider these two broad aspects of fairness: those focusing on the validity of an assessment, which we characterize as concerns about *bias*, and those focusing on the soundness of the educational or social policy to which the assessment contributes, which we characterize as concerns about *equity*.

Fairness is a concern with respect to all persons who are assessed, whether considered as individuals or as members of some definable group. Much of the relevant research has focused on fairness with respect to groups defined by social background: typically race/ethnicity, primary language, gender, and handicapping conditions, although any relevant grouping might be studied, for example, based on age, curriculum experienced, type of school attended, and so on. No validity research can warrant the validity of an interpretation or action for a particular individual unless it is carefully evaluated in the context of other relevant information about that individual. The *1985 Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practice*, jointly prepared by AERA, APA, and NCME, delineate obligations for tests users and for assessment development professionals to those who are assessed. Our focus here is primarily on studying issues of bias and equity across groups differing in social background.

Bias, like validity, refers to the soundness of an interpretation or action based on a test score—not to the assessment instrument itself. An assessment may be biased for some purposes but not for others. So, bias must always be evaluated in the context of the way in which an assessment instrument is interpreted and used. Validity refers to the extent to which the available evidence supports the intended interpretation and action over alternative interpretations and actions given the intended purpose. Bias refers to the extent to which there is evidence of *differential validity* for any relevant subgroup of persons assessed. For every validity issue raised in the chapter by Messick and colleagues, there is a related bias issue that may be raised with respect to a group of concern. An assessment-based interpretation or action is biased when it is not equally valid for all relevant subgroups.

Concerns about *equity* spill over the consensual bounds of validity and bias to include questions about the fairness of the educational system in which the assessment was used. It is possible for an assessment to be considered unbiased in a technical sense—in the sense that the intended interpretation is equally valid across various groups of concern-and yet be used in service of a policy that fails to promote equity. Clearly, equal access to a high quality education is a minimum requirement of a fair educational system, although, as we will discuss below, the means of documenting and eliminating inequalities in the opportunity to learn are complex and problematic. In addition, differences in educational outcomes are also relevant indicators of fairness-there should be evidence that students are benefiting from their education. Ideally, in a fair system, the average achievement is high and social distribution of achievement is equivalent for all groups of concern (Lee and Bryk, 1989). The same might be said for the range of other valued outcomes of schooling, such as self-esteem, citizenship, or access to employment or higher education. Where differences exist between socially-defined groups, there is an on-going effort to understand the reasons for the difference and to implement policies that move in the direction of closing the gap and raising the overall level of success. The question for assessment evaluators is whether an assessment is contributing to or detracting from the fairness of the educational system of which it is a part.

The fairness of an assessment system must always be evaluated in light of the purpose it is intended to serve and the way in which the results are interpreted and used. An assessment may be "fair" for some purposes, but patently unfair for others. For instance, if an assessment is used for instructional planning, evidence that students have had the opportunity to learn the material assessed is not necessary, although it may be useful; however, if the same assessment is used to inform decision about promotion or certification for graduation, such information is crucial. Moreover, simple differences in summary statistics between socially defined groups are not, in and of themselves, indicators of bias or inequity; they do, however, highlight the necessity of a careful search for additional evidence, in light of the purpose of assessment, to explain the source of the differences and to inform decisions about the appropriateness of the assessment system and the policy that surrounds it. Our intent is to offer readers a framework for studying fairness-collecting and evaluating evidence to inform conclusions about bias and equity—and, where existing research permits, to offer suggestions for designing and using assessments in ways that enhance fairness.

In the section entitled "Fairness and Bias," we offer suggestions for minimizing bias during assessment development, administration, and evaluation. In the section entitled "Fairness and Equity," we cover three interrelated sets of issues. First, we summarize evidence on differences in performance between groups differing in social background and offer recommendations for responsible reporting of such differences. Second, we highlight the importance of understanding differential performance in light of the differences in antecedent conditions, especially the opportunity to learn. Here, we emphasize the legal and ethical importance of ensuring opportunity to learn when students are held accountable for their performance as in cases of decisions about promotion or graduation. We close with a consideration of the potential for differential consequences of performance assessments across socially defined groups, highlighting the need for on-going critical dialogue among all stakeholders in the assessment process.

Fairness and Bias

As mentioned earlier, the term *bias* will be used here to refer to the extent to which a test score and test use are valid for all intended individuals and groups. As such, if an assessment results in scores that systematically underestimate the status of members of a particular group on the construct in question, then the test is biased against members of that group. An assessment is also biased if it purports to measure a single construct across groups, but in fact measures different constructs in different groups. To take a familiar extreme example, an assessment, in English, of quantitative reasoning administered to two groups of students who differ significantly in English proficiency, may be a relatively pure measure of quantitative reasoning in English-proficient group, but probably reflects a mixture of knowledge of English and quantitative reasoning in the less English-proficient group.

The term *bias* will also be used here to apply to situations where an assessment is differentially valid in selecting individuals for favored treatment (e.g., employment or college admissions) and in predicting some future criterion. We should note in passing that bias in selection and prediction is by no means a straightforward matter, and competing models of selection and prediction bias are still being debated (see Hartigan & Wigdor, 1989; and Journal of Educational Measurement, Special Issue on Bias in Selection, Vol. 13, 1976).

Finally, it is worth repeating that what counts as "bias" ultimately depends upon the purpose, intended interpretations, and context of the assessment. Assessment exercises on NAEP, for example, have no consequences for individuals, schools, or districts (at least not yet!) and are very different from assessments for individual certification or placement, and lead to different concerns of the what constitutes bias and the nature of that bias.

With these caveats, we may ask, what specific factors affect bias in performance assessment and what procedures can be undertaken to minimize their effects? Biases in performance assessment may be conceived as falling under two broad rubrics. The first, biases emanating from *internal* sources, encompasses the actual content and administration of the assessment itself. The second, *external sources* of bias, refer to the relationship between the assessment and some outside criterion of practical or theoretical interest.

Internal Sources of Bias in Performance Assessment

In virtually all aspects of the assessment situation, internal biases can enter the assessment. Biases may be present in the specification of the content framework, in the development of actual tasks and exercises, in the selection and training of scorers, and in the actual scoring rubrics. Moreover, performance assessments do not follow a compensatory model. Poorly conceived exercises that draw heavily upon construct-irrelevant abilities cannot be rescued by high quality scorer training. A well-thought-out and comprehensive content framework can be negated by exercises that do not reflect the framework, or by an inadequate scoring scheme. Performance assessment is a good illustration of the old saw that a chain is only as strong as its weakest link.

Diversity of development and scoring panels. First, diversity among the experts selected to specify the content framework, to develop and evaluate candidate exercises, and to devise valid and reliable scoring schemes is critical. This is especially true for those subject areas where differences in the skill required for successful performance vary as a function of context. The

conduct of an arts assessment is a good example. Individuals of different racial, ethnic, or gender groups will probably vary substantially in their opinions about what range of content should be included for an arts assessment; what performance tasks should be included; and how the scoring should be developed and applied. The problems associated with a lack of diversity among experts are of particular concern with performance assessments, as there is far more room for subjective judgment with open-ended and extended response questions than with traditional multiple choice or true or false questions. Hence, it is particularly critical that a full array of opinions from experts of different cultures, ethnic, and social backgrounds be reflected in the process of test development for performance assessment.

Specifying the content framework. The first source of possible bias in assessment is in the specification of the content framework. If the initial conceptualization of what constitutes understanding of geometric concepts, for example, is flawed, then the tasks devised to assess that understanding will in all probability be similarly flawed. Even if the specification of the content framework is sufficiently rich and diverse to encompass a wide variety of original and creative demonstrations of understanding, the exercises designed to reflect that understanding may contain sources of irrelevant difficulty that disadvantage members of subgroups of the population. To continue with the above example, the understanding of a concept in geometry, and the ability to use that concept in problem solving, can be distinguished from the ability to write reflectivel y about one's understanding. Much of performance assessment, however, is dominated by precisely this kind of expressive and reflective ability. (To be sure, such abilities are important in their own right, but to the extent possible they should be distinguished from other abilities.) A given teacher can often write eloquently about what she or he wants students to learn and about "constructivist approaches" to teaching, but the actual behavior before the classroom may belie this eloquence. The opposite is also true. In a research project co-directed by the first author involving the development of expertise in aircraft engine mechanics, the recognized shop aces were often unable to explain an assembly or tear-town procedures away from the shop. They insisted upon going to the job site and *demonstrating* the procedure in question.

In some subject matters, specification of the content framework would appear to be a relatively straightforward, if time-consuming, matter. The disciplinary consensus achieved in the National Council of Teachers of Mathematics Standards is a case in point. These standards represent the work of an incredibly diverse set of individuals from a variety of philosophical, social, and ethnic backgrounds who teach in widely different cultural, social, and economic contexts. Yet consensus on what constitutes desired mathematical competencies in students and desired practices in teaching mathematics was achieved. Whether a similar consensus will be achieved in other disciplines remains to be seen. History and social studies, to take what are perhaps the most obvious examples, are largely "constructive" enterprises, involving matters of interpretation that will vary as a function of ethnicity, gender, and social class. Consensus regarding what constitutes the very "stuff' of history and social studies is problematical, not to mention what constitutes evidence of "understanding" social and historical phenomena. To repeat, it is important to ensure that a diversity of viewpoints be represented in the initial specification of the content framework.

Selection ***of tasks and exercises***. ***The*** unidimensionality of most paper-and-pencil measures of cognitive ability constitute both a strength and, ironically, an essential weakness. On the one hand, unidimensionality facilitates test use and the interpretation of test results. Whether the test

model underlying performance is classically-based or based upon item-response-theory, persons with the same "observed score" or "theta" are presumed to be approximately equal in the ability being measured. In performance assessment, however, the situation is fundamentally different. Generalizability studies in a wide variety of contexts indicate that the variance components due to exercises and the exercise-by-examinee interaction constitute significant portions of the total variance in performance. The situation is exacerbated by the fact that the number of exercises are usually much smaller in typical performance assessments than in standardized multiple-choice group tests. Thus, even if a "general factor" in complex performance could emerge, it is unlikely that it would be reliable given the number of exercises involved.

The multidimensionality of performance assessments stems not only from the inherent complexity of such assessments, but also from the very real effects of ***context***. Posing a task one way, as opposed to another, or in one context as opposed to another, can result in surprisingly different results. Horizontally presented addition problems turn out to be significantly more difficult for young children than the same problems presented vertically. This finding is no doubt traceable in part to instruction, but it highlights the fact that early knowledge is precarious, and seemingly trivial, surface differences in exercise formats can have substantial effects on performance.

One implication of the relatively large variance components due to exercises and the exercise-by-candidate interaction in performance assessment is that *which* exercises are included in an assessment, and the wording and context of those exercises, becomes important and has perhaps more implications for fairness and equity than would be the case in standardized group tests.

Sensitivity Review. Assessment stimuli must, of necessity, be about ***something***. Whether the assessment is a writing sample, an extended response to a practical problem in mathematics, or a portfolio entry, the actual exercise will contain words that describe situations encountered in our culture. Sound testing practice, not to mention simple common sense, requires that assessment materials not include wording, references, or situations that are offensive and insensitive to test-takers. "Sensitivity review" is a generic term for set procedures for ensuring (1) that stimulus materials used in assessment reflect the diversity in our society and the diversity of contributions to our culture, and (2) that the assessment stimuli are free of wording and/or situations that are sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population. Sensitivity review is becoming a routine part of many large-scale test development agencies. Ramsey (1993) provides an excellent review of sensitivity review procedures at ETS, a leader in this area.

Selection and training of administrators and scorers. Sound administration and scoring of performance assessments are much more difficult to achieve in performance assessment than in multiple-choice tests. In the latter, standardization of *instructions* in administering the test is presumed to be a part of the definition of fairness, and standardized scoring is assured by a machine. In performance assessment, however, the selection and training of administrators and scorers are critical features of the overall validity of the assessment. The necessity for a diverse group of administrators and scorers has already been mentioned. In addition, however, close attention must be paid to administrators and scorers, and their biases.

The objective of assessment should be not so much the standardization of ***instructions***, as ensuring that examinees have a common understanding of the tasks involved. Because

performance assessment is, or can be, richly interactive, it is vitally important that administrators not only understand the constructs being assessed, but it is essential that they know how to discern when an examinee does not understand what is being asked and what kinds of additional explanation is needed. Regarding scorers, it is surprising how many people continue, even after training, to equate and confuse "writing ability" with "understanding." Many people find misspellings so annoying that it is impossible for them fairly to ignore this construct-irrelevant ability in essays where the intent is understanding a discipline, rather than spelling, per se. It is, moreover, often impossible to "train out" personal biases some persons hold about the abilities of boys vs. girls, or blacks vs. whites.

The emphasis in training should not be restricted solely to eliminating sources of construct-irrelevant difficulty for minority group members. It may also happen that assessment exercises contain construct-irrelevant elements that specifically *advantage* certain individuals. The "neatness" and visual attractiveness of a portfolio, for example, are related to the resources available to a student or teacher, but are typically unrelated to the ability being measured. Yet, despite training, these factors often figure unconsciously in scoring. The bottom line is that often the thought and hard work that go into the specification of the content framework and into the construction of high quality exercises that are tied to that framework, can be rendered useless by poorly chosen or inadequately trained scorers.

Bias in performance exercises. The sophisticated statistical machinery that has been developed for the identification of biased items in paper-and-pencil tests cannot unfortunately be applied in a straightforward manner to performance assessment. For example, all of the techniques for identifying differentially functioning items by subgroups using item response theory (Holland & Wainer, 1993) require large numbers of examinees and more items than are typically found in performance assessment. The same applies to classically-based procedures such as the Mantel-Haenszel technique (Holland & Thayer, 1988), and the standardization method (Dorans & Holland, 1988).

A repeated and rather baffling finding in studies that attempt to identify patterns of DIF in standardized aptitude and achievement measures is that items identified as functioning differentially do not appear to differ in any discernible way from items not so identified. O.Neill and McPeek (1993) have attempted to summarized patterns of DIF for the Scholastic Aptitude Test, and although some of their conclusions are encouraging, the general rule is that patterns of DIF traceable to substantive characteristics of the items or item-formats are difficult to come by. Moreover, the connection between DIF (as a technical characteristic of test items defined by reference to internal properties of the test), and *test bias* (a much larger issue involving the opportunity to learn, among other things) is not at all clear. Space does not allow a thoroughgoing discussion of this complex issue. For an incisive discussion of the relation between *item bias* (a term that predated DIF by some 35 years), DIF, and test bias broadly conceived, see Camilli Shepard (1994).

Absent any agreed-upon methodology for the systematic examination of sources of bias in performance assessment, we must rely upon expert judgment and a close adherence to sound assessment practices. One of the significant promises of performance assessment is that the sources of any biases in the assessment, as well as the sources of any genuine deficiencies of examinees, are more likely to be discovered. Unlike multiple choice test, where the only

permanent record is a series of blackened ovals by the examinee, performance assessment allows at least some insight into an examinee's thinking.

External Sources of Bias in Performance Assessment

In many applied uses of tests, the relationship between test performance and some external criterion is crucial, and the extent to which test scores of various social groups are *differentially* related to the external criterion becomes a potential source of bias.

The behaviors elicited by traditional group-administered paper-and-pencil tests are often different from the behaviors the test is intended to predict. The differences are quite apparent in such applications as certification and licensure and in employment contexts. In the performance of their professional legal and accounting duties, for example, lawyers and accountants are virtually never required to perform the kinds of tasks typically found on group administered multiple-choice. Nor are the circumstances under which the tests are administered reflective of the circumstances under which their actual duties are performed. (Most lawyers, for example, would never decide on a specific legal strategy without extensive prior research).

In the context of student assessment, much of the criticism of traditional paper-and-pencil measures of academic achievement derives from the impoverished, decontextualized nature of the stimulus questions. Their predictive validity for future academic performance notwithstanding, verbal analogies are not what teachers of English Language Arts should be about. It is entirely likely that excellent instruction in English Arts instruction coupled with performance assessment (which properly viewed are complementary) would result in students who do well on verbal analogies, but that is a derivative side effect, rather than an object of instruction. One point appears certain: to the extent that instruction and assessment become merged into one coherent process, performance assessment has the potential to remove one major criticism of current testing practice, to wit, that biases in the prediction of external criteria are more likely when the relationship between teaching, testing, and later performance is only implicit.

Fairness and Equity

An unbiased assessment—one that is equally valid for all relevant subgroups—is, of course, a minimal requirement for fairness. Here we move beyond questions of bias and validity to raise questions about the fairness of the educational system in which the assessment is used. Many of the state and national assessment-based reform efforts reflect high expectations about the extent to which assessments can facilitate excellence and equity in education. These expectations must be carefully and continually evaluated as assessments are designed and implemented. Here, we will focus on differences between groups varying in social background—differences in performance on the assessments, differences in opportunity to learn the material assessed as well as other valued goals of education, and difference in the consequences of assessment use. In particular, we focus on differences between underserved minority groups and the majority group to which they are compared. By underserved minority groups, we refer to students from poor communities and from black, Hispanic, and American Indian racial/ethnic backgrounds. We characterize these groups as underserved because of the large body of evidence (e.g., Oakes, Gamoran and Page, 1992) that documents inequalities in access to-high quality education—an issue to which we will return in the section on opportunity to learn.

Differential Performance

A large body of evidence shows that, on the average, students from undersexed minority groups-those from poor communities and from black, Hispanic, and American Indian racial/ethnic backgrounds-have scored lower than students from the relevant majority comparison group on multiple choice tests across a wide range of assessment purposes involving students in elementary, secondary, post-secondary and vocational education (National Commission on Testing and Public Policy, 1991). This difference in performance between socially defined groups is often referred to as the "*achievement gap.* " Some educators have theorized that performance assessments will narrow the achievement gap because they sample a broader range of capabilities, provide examinees with the opportunity to explain their responses, permit them more latitude in choosing tasks that best represent their capabilities, and, in the case of hands on assessment, rely less heavily on language skills (see Office of Technology Assessment, 1991; Hambleton and Murphy, 1992). Others raise concerns that these unfamiliar assessment formats may disadvantage students when compared to the more familiar multiple choice format (see Badger, in press; Office of Technology Assessment, 1991). These expectations must be carefully evaluated.

Relevant evidence is scanty and mixed. (Baker and O'Neill, in press; Bond, 1992, Office of Technology Assessment, 1991; Darling-Hammond, 1994; Linn, 1993; Linn, Baker, and Dunbar, 1991; Madaus and Kelleghan, 1992; Madaus, 1994; in press; National Commission on Testing and Public Policy, 1990; Nettles, in press; Winfield and Woodard, 1994.) Typically, differences in performance between groups are compared for traditional and alternative assessments in terms of effect size or by comparing the proportions passing (or scoring above or below some cut score). Effect sizes are computed by subtracting the mean for the minority group of concern from the mean for the majority comparison group and dividing by the standard deviation of the majority comparison group. A narrowing of the "achievement gap" would be indicated by a smaller effect size or greater equivalence across groups in proportions passing when performance assessments are compared to more traditional assessments.

Some studies have found no difference in the achievement gap between performance assessments and more traditional assessments. For instance, Linn, Baker and Dunbar (1991), using results from the 1988 NAEP, found that differences in achievement between black and white students in grades 4, 8 and 12 were essentially the same for the open-ended essay items in the writing assessment and the primarily multiple choice items in the reading assessment. Other studies have found a narrowing of the achievement gap. For instance, Badger (in press), using results on open-ended and multiple choice items from the Massachusetts testing program found smaller performance gaps for open-ended items when comparing students in low- and high-SES schools, and when comparing black and Hispanic students to white and Asian students. Her results were consistent across grades and subject areas. A comparison of the magnitude of differences between black and white military personnel on paper and pencil (Armed Services Vocational Aptitude Battery (ASVAB) and other job knowledge tests) versus hands on tests resulted in smaller differences for the hands on tests (Wigdor & Green, 1991). Others have found an increase in the achievement gap. Elliott (1993), comparing differences in performance for white and Asian students to those for black and Hispanic students on multiple choice and constructed response items for the 1992 NAEP in mathematics found a larger gap for constructed response items. LeMahieu (1992) reported larger differences between black and white students when portfolio scores, reflecting self-selected pieces of writing, were compared to

an independent measure of writing. LeMahieu attributes the differences to self-selection: black students did not choose material from their writing folders that best represented their writing.

Given the existing evidence, there is no reason to believe that differences observed with traditional assessments between underserved minority groups and the majority comparison group will necessarily diminish with performance-based assessment. To offer any other generalization based on the existing evidence would be inappropriate. The studies differ widely in terms of the purposes of assessment, the constructs assessed, the format in which performances are presented and evaluated, the antecedent instructional conditions, and the social and academic characteristics of the students assessed. The only appropriate advice is to highlight the crucial importance of investigating and then attempting to understand the reason for group differences for *each* context of assessment.

Careful attention must be given to the way in which group differences are reported and interpreted. Reporting simple differences in performance between schools or districts, between racial/ethnic groups, between students from poor and wealthier families, without additional information to assist in *explaining* the differences, may result in serious misinterpretations. For instance, changes in assessment scores from year to year may simply reflect changes in the student population (dropouts or transfers, for instance) rather than changes in the capabilities of students. Differences in assessment scores across ethnic groups may reflect differences in socioeconomic status of the communities in which they live. Differences in assessment scores from school to school may reflect differences in resources and activities such as the qualification of teachers or the number of advanced course offerings. Of most serious concern here is the potential misinterpretation of differences between racial/ethnic groups. Lee cautions (personal communication, April 14, 1994) that providing racial breakdowns may "overestimate the importance of racial differences in academic achievement ... because they are confounded with uncontrolled social class differences"; this risks misinforming the nation by "allowing people to conclude that it is only race (and not poverty) which is driving these differences". Haertel (1989) similarly notes that controls for socio-economic status reduce apparent disparities based on race/ethnicity. Moreover, differences in achievement between underserved minority and majority groups (whether defined by race/ethnicity or socio-economic status) can be substantially explained by differences in access to high quality education (Darling-Hammond, 1994; Barr and Dreeben, 1983; Lee and Bryk, 1988, 1989; Oakes, 1985; Oakes, Gamoran, and Page, 1992). As Darling-Hammond and Ascher (1991) note, "comparisons of test scores that ignore these factors hold little promise of directing policy makers' attention to the real sources of the problem, so that it can be rectified" (p. 16). This is the issue to which we turn next.

Opportunity to Learn and Other Antecedent Conditions

All of the policy research we reviewed appropriately places a high premium on providing a "level playing field" for students and schools. In the context of assessment, there are at least four reasons why it is important to provide information about opportunity to learn. First, when assessments are used for high stakes decisions, such as promotion for graduation, it is a legal and ethical responsibility to ensure that students have had the opportunity to acquire the capabilities for which they are held accountable. Second, regardless of whether the stakes of assessment are high or low, such information is essential both for understanding the results and for directing attention to needed reforms in policy and practice. Third, it must be recognized that assessments not only document the success of learning opportunities, they constrain and enable future

learning opportunities (an issue to which we will return in the section on differential consequences). And so, when assessments are introduced into the system, it is essential to study both the antecedent and consequent conditions of their use. Finally, the issue of "opportunity to learn" is no longer solely a measurement or even an ethical one, but a matter of legal precedent. The most directly relevant case law is, of course, Debra P. vs. Burlington (730 F. 2d 140 11th Cir. 1984), in which the court held that a state's minimum level of performance on a standardized competency test as a prerequisite for a high school diploma is proper only if the assessment tested what was actually taught in the schools. The Due Process Clause of the Fifth Amendment, Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, and Section 504 of the Rehabilitation Act of 1973 are implied if student assessments have an adverse impact on students because of their race, national origin, gender, or disability. If, for example, a test results in substantially different rates of eligibility for educational benefits or services on the basis of a student's group membership, then the use of that test may be in violation of civil rights laws and statutes.

Unquestionably, differential learning opportunities play a major role in determining student achievement (Barr and Dreeben, 1983; Gamoran, 1987; Lee and Bryk, 1988, 1989; Oakes, 1985; Oakes, Gamoran, and Page, 1992). Moreover, there is evidence of social stratification-differences attributable to race/ethnicity and socioeconomic status-in the learning opportunities available to students and consequently in educational outcomes. Differences in access to high quality education occur within schools, in course taking patterns, for instance, as well as between schools in the resources they have available. Oakes, Gamoran, and Page (1992), summarizing literature on curriculum differentiation, report that "disproportionate percentages of poor and minority students (principally black and Hispanic) are found in curricula designed for low ability or non-college bound students ....Further, minority students are consistently underrepresented in programs for the gifted and talented .. .. Although it seems clear that consistently lower performance on tests by non-Asian minorities under girds this pattern, the result is that curriculum differentiation leads to considerable race and class separation and race- and class-linked differences in opportunities to learn" (p. 58x, 590). Darling-Hammond (1994) summarizing evidence on the social distribution of qualified teachers notes that "because teacher salaries and working conditions are inadequate to ensure a steady supply of qualified teachers in poor districts, low-income and minority students are routinely taught by the least experienced and least prepared teachers" (Darling-Hammond, 1994, p. 16).

No consequential decision about individual students or educational programs should be made on the basis of a single assessment. With respect to individuals, additional information about the student's achievement and past experience should be considered. As Haertel (1989) notes in his paper for the National Commission on Testing and Public Policy, "simplistic policies, where action is triggered by scores above or below a cutting point on a single test. . . are contrary to the consensus of professional practice in testing." (p. 32). At the program level, assessment scores should be part of a comprehensive system of indicators, that includes information about student characteristics and about school activities and resources over which policymakers have some control.

While it is beyond the scope of this chapter to provide specific advice about how to design an informative and equitable indicator system, we highlight some general guidelines from policy researchers and direct readers to the original articles for more specific advice about how to operationalize their suggestions. We note, as does Oakes (1989), that there are many obstacles to

overcome in developing standards and indicators of opportunity to learn, including insufficient measurement technology and feasibility. Porter (1993) raises the question of whether school delivery standards should be intended to inspire and to provide a vision for educators, students, parents, and the public of what good practice might be, or whether delivery standards should provide prescriptions of required practice that can be used to police the action of teachers, school administrators, and public officials. He suggests distinguishing between school process standards, which would provide the vision, and school performance standards, which would support school monitoring, thus meeting accountability requirements for ethical and legal purposes. Our goal here is simply to suggest some possible frameworks for studying these issues.

Some of the advice on documenting opportunity to learn focuses on whether students have had the opportunity to learn the material *on which they are tested.* Mehrens and Popham (1992), focusing on legal defensibility, suggest gathering the following evidence regarding students opportunity to learn the material covered by a particular test or assessment:

> (1) Self report from teachers regarding the extent to which instruction for tested content was supplied....(2) Self reports from students regarding whether the students have been taught 'how to answer the illustrated kinds of test questions....(3) Analysis of required textbooks to determine if test content is addressed....(4) Analyses of curricular documents such as course syllabi to establish whether tested content is supposed to be addressed by teachers....(5) Analyses of teacher's classroom tests to discern if the content contained in the high-stakes examination is also addressed in teachers' routine classroom tests. (p. 275-276).

They advise against use of classroom observation because the expense and the difficulty of obtaining an adequately representative sample to permit inferences about all the tested material. While this advice addresses the opportunity to learn the tested material, it does not address the quality of education more broadly nor, with the possible exception of examining teachers tests, does it address the way in which the material is taught-an issue crucial to performance assessments that encourage students to interpret and evaluate material for themselves. (See Pullin, 1994, for a more extensive review of the legal issues involved in current national educational reform proposals.)

Moving beyond the content of tested material, the recently passed goals 2000 legislation offers the following guidelines for elements to be included in voluntary national opportunity to learn standards

- the quality and availability to all students of curricula, instructional materials, and technologies, including distance learning;
- the capability of teachers to provide high-quality instruction to meet diverse learning needs in each content area to all students;
- the extent to which teachers, principals, and administrators have ready and continuing access to professional development, including the best knowledge about teaching, learning, and school improvement;
- the extent to which curriculum, instructional practices, and assessments are aligned to voluntary national content standards;

- the extent to which school facilities provide a safe and secure environment for learning and instruction and have the requisite libraries, laboratories, and other resources necessary to provide an opportunity-to-learn; and

- the extent to which schools utilize policies, curricula, and instructional practices which ensure nondiscrimination on the basis of gender. (Goals 2000 Legislation, Sec. 213(c)(2)(A-F)).

Factors that effect opportunity to learn occur at a variety of levels. At the school level, factors such as financial resources, qualification of teachers, curriculum offerings, teachers' work loads, materials resources such as libraries, laboratories, and computers, and so on, should be investigated. However, equality of opportunity cannot be adequately addressed by simply looking at school level resources and assuming the students within schools have equal access to those resources. Substantial variation in access to resources exists within schools and even within classes that school level indicators can't observe (e.g., Barr and Dreeben, 1983; Lee and Bryk, 1989; Oakes, 1989; Porter, 1993).

Oakes (1989) notes that school resources are mediated by organization policies and structures as well as by the school culture that is reflected in norms and relationships.

> Even though we do not fully understand how schools produce the results we want, context information may provide clues to policymakers about why we get the outcomes we do. Measures of what goes on in schools can add important information to the political discussion about how to improve them (Oakes, 1989, p. 182).

Similarly, Porter (1993) notes that evidence of opportunity to learn must include evidence of effective pedagogy. With performance assessment, this covers instructional strategies such as "a need for instruction to emphasize active student learning, where students take responsibility for constructing knowledge through writing, discussion, lab work, manipulatives, and computer simulations" (p. 26). Moreover, knowledge of school resources, such as libraries and labs must be accompanied by evidence of course specific resources indicating access for all students in all subjects.

Oakes (1989) suggests gathering information about the following sets of "context indicators" to monitor schooling resources and processes. *Access to knowledge* refers to the "extent to which schools provide students with opportunities to learn various domains of knowledge and skills" (p. 186). Indicators might include: teacher qualifications, instructional time, course offerings, class grouping practices, materials/laboratories/equipment, academic support programs, enrichment activities, parent involvement, staff development, and faculty beliefs. *Press for achievement* refers to the "pressure which the school exerts to get students to work hard and achieve" (p. 186). Possible indicators include: focus on academics, graduation requirements, graduation rates, enrollment in rigorous programs, recognition of academic accomplishments, academic expectations for students, uninterrupted class instruction, administrative involvement in academics, quantity and type of homework, and teacher evaluation emphasizing learning. *Professional teaching conditions are the* "conditions that can empower or constrain teachers and administrators as they attempt to create and implement instruction programs" p. 186). Here, potential indicators include teacher salaries, pupil load/class size, teacher time for planning collegial work, teacher involvement in decision making, teacher certainty, teacher autonomy/flexibility, administrative support for innovation, and clerical support. Clearly, to

address the equity question, these indicators must be examined to see whether they are equally distributed across all groups of concern.

Factors that exist outside of school influence students readiness to take advantage of the opportunities school affords them. These include factors such as health, nutrition, living conditions, family support for education, educational resources in the home, and so on (Lee and Croninger, in press; Madaus, 1994, in press; Koretz, et al., 1992). Also of concern are differences between the cultural context of the home and the school, which are too often viewed as deficits rather than differences (Garcia and Pearson, 1994; O'Connor, 1989; LaCelle-Peterson and Rivera, 1994). Darling-Hammond (1994), speaking in the context of assessment, notes, "the choice of items, responses deemed appropriate, and content deemed important are the product of culturally and contextually determined judgments, as well as the privileging of certain ways of knowing and modes of performance over others. (p. 17; See also, Garcia & Pearson, 1994; Gardener, 1983; O'Connor, 1989; Steinberg, 1985)." (Darling-Hammond, 1994, p. 17) Gordon, while affirming his commitment to standards of competence that are the same for all populations, highlights the importance of allowing differential indicators of progress toward those standards: "The task is to find assessment probes which measure the same criterion from different contexts and perspectives which reflect the life space and values of the learner ....Thus options and choice become critical features in any assessment system created to be responsive to equity" (Gordon, 1992, pp. 5-6). These issues concerning potential disjunctions among cultural contexts are equally relevant to any school-based learning or assessment activity.

Frameworks for more comprehensive indicator systems that tie together background information, information about school context and processes, and information about valued learning outcomes have been suggested by Darling-Hammond and Ascher (1991), Porter (1991, 1993) and Bryk and Hermanson (1993). Porter, drawing on Shavelson (1987, in Porter, 1991), suggests a comprehensive model organized around inputs, processes, and outputs of education. Inputs include fiscal and other resources, teacher quality, student background, and parental/community norms. Processes are divided into organizational and instructional characteristics of schooling. Organization characteristics of schooling include national, state, district, and school quality; instructional characteristics include general curriculum/content and teaching/pedagogy quality, student nonacademic activities, course specific teacher quality, and course specific resources. Outputs include achievement, participation, and attitudes and aspirations.

Bryk and Hermanson (1993), drawing on the work of the Special Study Panel on Education Indicators (NCES, 1991, in Bryk and Hermanson, 1993), caution that inputs-processes-outputs models risk conveying a false sense of control based upon external manipulation. They suggest that a comprehensive indicator system should focus *on the* following "enduring concerns" in education, the purpose of which is to inform public discourse about the means and ends of education:

> Clearly the system must focus on *student learning* and the *quality of its educational institutions. The* system should, as well, report on the larger social context that also educates. This implied a concern for *children's readiness to learn as they* enter the formal educational system and, in more general terms, *societal support for learning*. It must also inform us about the broader aspirations we hold for education in a modem democratic

society. This meant a dual focus on *educational equity* and the contributions of education to *economic productivity.* (Bryk and Hermanson, 1993, p. 468).

They goon to suggest that the system should use a number of different reporting and data collection levels.

> As a trend in a particular indicator engages our attention, it should naturally lead to more detailed statistical information, including in-depth studies and case analyses that might illumine some of the forces at work .... The system should have a strong conceptual organization, capturing both established means-ends generalizations from social science and the best clinical expertise. (p. 468).

As a comparison of these more comprehensive indicators systems suggest, one of the major issues underlying assessment-based reform efforts is the way in which information is expected (and intended) to influence the educational system. That is the issue to which we turn next.

Differential Consequences[2]

It is widely hoped that performance assessments will not only permit valid inferences about the quality of education but serve as instruments of reform by raising standards for all students, thus promoting excellence and equity. Given our past experience with high stakes assessment and the current state of our knowledge about the construct validity of performance assessments, this anticipation is optimistic at best—the assumptions about the quality of information and the consequences of the assessment must be carefully evaluated. In understanding and evaluating consequences, it is important to look not just at the format of the assessment (e.g., multiple choice versus performance based) but also at the way in which the assessment is used to promote reform.

A growing body of evidence indicates that when assessments are visible and have consequences for individuals or programs, they alter educational practice, sending an unequivocal message to teachers and students about what is important to teach and learn (e.g., Johnston, Weiss, and Afflerbach, 1990; National Commission on Testing and Public Policy, 1989; Resnick and Resnick, 1992; Smith, 1991). In a review of literature on the impact of classroom evaluation on students, Crooks (1988) concluded assessment not only affects students' judgments about what is important to learn, but also their motivation, perceptions of competence, approaches to personal study, and development of enduring learning strategies. Similar conclusions have been drawn about the impact of district and state mandated assessment on the judgment, perceptions, and instructional strategies of teachers. The salience of this influence seems to be directly related to the importance of the consequences of testing to students and teachers and to the administrative and supervisory practices of a school or district. In a paper prepared for the National Commission on Testing and Public Policy, Resnick and Resnick concluded that "when the stakes are high—when schools' ratings and budgets or teachers' salaries depend on test scores—efforts to improve performance on a particular assessment seem to drive out most other educational concerns" and "to progressively restrict curricular attention to the objectives that are tested and even the particular item forms that will appear on the test." (1992, p. 58).

Evidence suggests that the narrowing of the curriculum associated with high stakes standardized assessment may be falling disproportionately on certain groups of students for

whom concerns about equality of education have been most salient. Neill and Medina (1989) found that standardized testing was more prevalent in large urban school systems. Madaus, West, Harmon, Lomax, and Viator (1992, in Madaus, in press) report national survey results showing that teachers with greater than 60 percent minority students in their classrooms were more likely to teach to standardized tests, to spend time in direct test preparation, and to spend more of their class time on these activities than were teachers in predominately white classrooms. Similar results were found for teachers of students in Chapter 1 programs. Herman and colleagues (Herman and Golan, 1993; Dorr-Bremme and Herman, 1986), using teachers' and principals' self-reports, found that in low income communities, teachers felt a greater need to spend time preparing students for tests and principals felt that tests counted far more in decisions such as planning curriculum, making class assignments, allocating funds, and reporting to district officials and the community. To the extent that testing undergirds decisions about educational placement, studies on the effects of tracking reviewed by Oakes, Gamoran, and Page (1992) also support concerns about the differential impact of testing on underserved minority students. They report that qualitative differences exist in the educational experiences provided students in different tracks, with lower track students progressing more slowly through the curriculum, having less experience with inquiry skills, problem solving, and autonomy in their work, and losing more educational time to classroom management; that the achievement gap between students in higher and lower tracks increases over years of schooling; and that track placement can have a long lasting impact on the life chances of students after high school. Taken together with our knowledge about the impact of high stakes testing on the curriculum, these observations raise substantial concerns about differential access to knowledge for students from underserved minority groups.

It is anticipated that performance assessment can overcome such influences by providing targets toward which we want teachers to teach. Clearly, providing a broader range of valued educational outcomes is essential to address concerns about narrowing the curriculum, as most advocates of performance assessment have argued, as is a careful consideration of what outcomes remain unaddressed. However, past experience with high stakes uses of multiple choice tests suggests the need for caution. Although high-stakes testing programs frequently result in improved test scores, such improvement does not necessarily imply a rise in the quality of education or a better educated student population (Darling-Hammond and Snyder, 1992; Haertel, 1989; National Commission on Testing and Public Policy, 1990; Shepard, 1992). At best, test scores can reflect only a small subset of valued education goals. When educators focus their attention on improving test scores, they not only narrow the curriculum, they undermine the validity of the tests as indicators of a broader range of achievements (Shepard, 1992). Koretz, Linn, Dunbar, and Shepard (1991, in Linn, 1993) showed evidence of test score inflation comparing performance in two districts between the district mandated tests in reading and math and tests constructed to cover the same content objectives. Students in the high-stakes testing districts scored considerably lower on the alternative tests in reading for both districts and in mathematics for one district. When particular tests become targets for instruction they become less valid indicators of the broader capabilities they were intended to tap. As Madaus (in press) notes, "there is no evidence to support the belief that performance-based measures will not be as corruptible as any multiple choice measure when used in the context of measurement driven instruction" (p. 41).

Further, evidence suggests that test driven reforms may undermine attempts at genuine educational reform by diverting attention from fundamental educational problems. Ellwein, Glass, and Smith (1989) conducted extended case studies of five competency testing programs at the state and district level. They concluded that competency tests and standards served more as symbolic and political gestures rather than as instrumental reforms-focusing attention on the tests themselves rather than on their impact, utility, or value. Similarly, Corbett and Wilson (1992), who studied competency testing programs in two states, focusing on six districts per state, found that the pressure to do well on tests did not encourage fundamental consideration of the structures, processes, or purposes of education, rather it caused "knee-jerk" reactions designed to improve test scores quickly—actions which many of the educators involved considered counter-productive.

When performance is tied to rewards and sanctions, including public disclosure, the system sometimes results in peculiar and counter-productive incentives to improve the appearance of progress, such as retaining students so that their scores will be compared with a younger cohort, placing students in special education programs so that their scores don't count, targeting instruction at those closest to the passing level and ignoring the needs of students unlikely to pass the exam, or failing to discourage low-scoring students from dropping out. (e.g., Haertel, 1989, Madaus, in press; Slavin and Madden, 1991). Such practices exacerbate inequities in educational opportunity.

Some suggest reporting information in ways that will minimize the negative effects of high-stakes standardized assessments. Haertel (1989) suggests preparing reports that attend to the entire distribution of achievement (e.g., including the 25th, 50th, and 75th percentiles) so that progress is not just defined in terms of increases in measures of central tendency. He also suggests tracking the population of students—counting dropouts and transfers—and, where possible, looking at the progress of individual students. Aggregate indices of individual growth are better indicators of program effectiveness than are differences in average scores which may simply reflect differences in the population of students. Moreover, they are less likely to confound program effectiveness with incoming capabilities of students. Lee (personal communication, April 14, 1994) notes that even in the absence of longitudinal achievement data, it is possible to control for academic background using proxy measures such as self reports of students' previous grades and whether or not students had repeated or skipped a grade level. And, as we noted before, interpreting educational outcomes in light of school resources and processes is important in controlling misinterpretations leading to counter-productive policies. (Darling-Hammond and Ascher, 1991; Darling-Hammond and Snyder, 1992; Haertel, 1989; Oakes, 1989; Porter, 1991, 1993).

However, some researchers argue that simply changing the format and reporting requirements of assessments is insufficient; rather, what is needed is a change in the way assessments are used to promote reform (Bryk and Hermanson, 1993; Darling-Hammond, 1994; Madaus, 1993, 1994, in press). Darling-Hammond (1994) raises the concern that, "if performance assessments are used in the same fashion as current externally developed and mandated tests are used, they are likely to highlight differences in students' learning even more keenly, but they will be unlikely to help teachers revamp their teaching or schools rethink their ways of operating" (p.19). She contrasts two different approaches to assessment reform, which reflect different theories of organizational change and different views of educational purposes. One view seeks to induce change through extrinsic rewards and sanctions for both schools and

students ..... The other view seeks to induce change by building knowledge among school practitioners and parents about alternative methods and by stimulating organization rethinking through opportunities to work together on the design of teaching and schooling and to experiment with new approaches. (Darling-Hammond, in press, p.114).

The recently passed Goals 2000 legislation suggests purposes for assessment that appear intended to increase the control of externally imposed assessments on teaching and learning. Its authors indicate that among the appropriate purposes for state assessments are "measuring and motivating individual students, schools, districts, state, and the nation to improve educational performance" (Sec. 213, (f) (1) (B) (iv)) and that after a period of five years, such assessments may be used "to make decisions regarding graduation, grade promotion, or retention of students" (Sec. 213, (f) (1) (C) (II)). In contrast, Darling-Hammond describes proposals in states like New York, Vermont, Connecticut, and California, where assessments used for policy purposes are distinct from those used for individual students.

"These states envision carefully targeted state assessments at a few key developmental points that will provide data for informing policy makers about program successes and needs, areas where assistance and investment are needed, and assessment models for local schools. Meanwhile, locally implemented assessment systems—including portfolios, projects, performance tasks, and structured teachers observations of learning-will provide the multiple forms of evidence about student learning needed to make sound judgments about instruction." (Darling Hammond, 1994, p. 20)

As research into the consequences of high stakes assessment suggests, choices made in designing assessment systems not only impact the nature of teaching and learning (in both intended and unintended ways) but also the nature of the discourse about the purposes and processes of education. Bryk and Hermanson (1993) offer a useful distinction between two different views of the ways in which indicators enter and influence the discourse and practice of educational reform: an "instrumental use" model and an "enlightenment" model. In "the instrumental use" model, the goals are: to develop a comprehensive set of outcome measures; to examine the relationship between these outcomes and indicators of school resources and processes; and, based upon that generalized knowledge, to control schools through allocation of resources, rewards and sanctions, and regulation so as to maximize performance on the outcomes. As they note, the instrumental use model characterizes much of the current rhetoric about the potential of indicators to improve schools. In criticizing this conceptualization, they argue first that there are many valued outcomes for which available measures do not exist. As our past experience suggests, any model which attempts to maximize measurable outcomes is likely to result in a variety of unintended, possibly undesirable, effects, including the undermining of progress in areas not addressed. More fundamentally, the instrumental use model, with its reliance on generalizations about the relationship between processes and outcomes, under-represents the complexity of schools. While "external policy-making and administrative action shape schools' structure and function" (p. 453), the "behavior, attitudes, and beliefs of actors inside the school-professional staff, parents, and students-influence its operations" (p. 453). "Schools are places where personal meaning and human intentionality matter. " (p. 457) An "enlightenment model" reflects a view of schools where interaction among individuals is fundamental and reform requires "changing the values and tacit understandings that ground these interactions". From this perspective, the goal of an indicator system is not to manipulate such interactions through external controls, but rather to "enrich and encourage

sustained conversation about education, its institutions, and its processes in order ultimately to improve them" (p. 467).

Notes

[1]The authors took responsibility for different aspects of the paper. Bond and Carr drafted the section on "Fairness and Bias;" Carr provided a discussion of legal issues that informed both sections of the paper; Moss drafted the introduction and the section on "Fairness and Equity;" and Bond coordinated the paper.

[2]This section draws heavily on Moss (in press).

References

Badger, E. (in press). The role of assumptions in a statewide testing program: Lessons from Massachusetts. In M. T. Nettles (Ed.), *Equity and excellence in educational testing and assessment* (pp. 329-372). Boston: Kluwer Academic Publishers.

Baker, E. L., & O'Neil, Jr., H. F. (in press). Diversity, assessment, and equity in educational reform. In M. T. Nettles (Ed.), *Equity and excellence in educational testing and assessment* (pp.77-94). Boston: Kluwer Academic Publishers.

Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: The University of Chicago Press.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), 429-444. New York: American Council on Education/MacMillan

Bond, L. (1993). Making innovative assessments fair and valid. *What we can learn from performance assessment for the professions,* Proceedings of the 1993 ETS Invitational Conference, Princeton, NJ: Educational Testing Service.

Bryk, A. S., & Hermanson, K. L. (1993). Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19*, 451-484.

Camilli, G. & Shepard, L. (1987). Inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12, 87-89.*

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (pp. 201-219). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Congress of the United States, Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex Publishing Corporation.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85(4), 438-481.*

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64(1)*, 5-30.

Darling-Hammond, L. (in press). Equity issues in performance-based assessment. In M. T. Nettles (Ed.), *Equity and excellence in educational testing and assessment* (pp. 97-121). Boston: Kluwer Academic Publishers.

Darling-Hammond, L., & Ascher, C. (1991). *Accountability in Urban Schools. New* York: National Center for Restructuring Education, Schools, and Teaching, Teachers College, Columbia and ERIC Clearinghouse on Urban Education.

Darling-Hammond, L., & Snyder, J. (1992). Reframing accountability: Creating learner-centered schools. In A. Lieberman (Ed.), *The changing contexts of teaching: Ninety-first Yearbook of the National Society for the Study of Education* (pp. 11-36). Chicago: University of Chicago Press.

Darling-Hammond (in press)

Dorans, N., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. Holland & H. Wainer (Eds.) *Differential Item Functioning,* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Holland, P. W. (1988). DIF detection and description: Mantel-Heanszel and standardization. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices.* University of California, Center for the study of Evaluation, Los Angeles, CA.

Elliott, E. (in press). National testing and assessment strategies: Equity implications of leading proposals for national examinations. In M. T. Nettles (Ed.), *Equity and excellence in educational testing and assessment* (pp. 415-424). Boston: Kluwer Academic Publishers.

Ellwein, M. C., Glass, G. V., & Smith, M. L. (1988). Standards of competence: Propositions on the nature of testing reforms. *Educational Researcher, 17(8), 4-9.*

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education, 60, 135-155.*

Garcia, G. E., & Peterson, P. D. (1994). Assessment and diversity. *Review of Research in Education, 20,337-391.*

Gardner, H. (1983). *Frames of mind. New* York: Basic Books.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses.* London: The Falmer Press.

Gordon, E. W. (1992). *Implications of diversity in human characteristics for authentic assessment* (CSE Technical Report 341). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California Graduate School of Education, Los Angeles, CA.

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 25-50). Boston: Kluwer Academic Publishers.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5(1),* 1-16.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practices, 12(4), 20-25,41-42.*

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Heanszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity.* pp. 147-169, Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnston, P. H., Weiss, P., and Afflerbach, P. (1990) *Teachers' evaluation of the teaching and learning in literacy and literature* (Report Series 3.4). Albany, NY: Center for the Learning and Teaching of Literature, State University of New York at Albany.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Joint Committee on Testing Practices, American Psychological Association.

Koretz, D. M., Madaus, G. F., Haertel, E., & Beaton, A. E. (1992). *National educational standards and testing: A response to the recommendations of the National Council on Education Standards and Testing* (Congressional testimony). Santa Monica, CA: Rand Institute on Education and Training,

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75.

Lee, V. E., & Bryk, A. S. (1988). Curriculum tracking as mediating the social distribution of high school achievement. *Sociology of Education5, 6(2)*, 78-94.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*, 172-192.

Lee, V. E., & Croninger, R. G. (in press). The relative importance of home and school in the development of literacy skills for middle-grade students. *American Journal of Education.*

LeMahieu, P., Eresh, J. T., & Wallace, R. C. (1992). Using student portfolios for public accounting. Paper presented at the conference *Diversifying student assessment: From vision to practice.* Center for Testing, Evaluation and Educational Policy, Boston College.

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

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

Madaus, G. F. (1993, October). *Assessment issues around the re-authorization of Chapter 1.* Paper presented to the National Academy of Education, University of Michigan, School of Education, Ann Arbor, MI.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64(1)*, 76-95.

Madaus, G. F. (in press). A technological and historical consideration of equity issues associated with proposal to change the nation's testing policy. In M. T. Nettles (Ed.), *Equity and*

excellence in educational testing and assessment (pp. 27-74). Boston: Kluwer Academic Publishers.

Madaus, G. F., & Kelleghan, (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 119-154). New York: Macmillan Publishing Company.

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

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Moss, P. A. (in press). Validity in high stakes writing assessment: Problems and possibilities. *Writing Assessment*.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to geteway: Transforming testing in America*. Chestnut Hill, MA: Author.

Neil, D. M., & Medina, N. J. (1989). (1989, May). Standardized testing: Harmful to educational health. *Phi Delta Kappan*, pp. 688-697.

Nettles, M. T. (Ed.). *Equity and excellence in educational testing and assessment. Ann* Arbor, MI: University of Michigan, School of Education.

O'Connor, M. C. (1989). Aspects of differential performance by minorities on standardized test: Linguistic and sociocultural factors. In B. R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 129-181). Boston: Kluwer Academic Publishers.

Oakes, J. (1985). *Keeping track: How schools structure inequality. New* Haven, CT: Yale University Press.

Oakes, J. (1989). What educational indicators? The case of assessing the school context. *Educational Evaluation and Policy Analysis, 11(2)*, 181-199.

Oakes, J., Gamoran, A., & Page, R. N. (1992). Curriculum, differentiation: Opportunities, outcomes, and meaning. In P. W. Jackson (Ed.), *Handbook of research on curriculum: A project of the American Educational Research Association* (pp. 570-608). New York: Macmillan Publishing Company.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA--SET-520). Washington, DC: Congress of the United States.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. Holland & H. Wainer (Eds.) *Differential Item Functioning*, (255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Porter, A. C. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis, 13(1)*, 13-29.

Porter, A. C. (1993). School delivery standards. *Educational Researcher, 22(5)*, 24-30.

Pullin, D. C. (1994). Learning to work: The impact of curriculum and assessment standards on educational opportunity. *Harvard Educational Review, 64(1)*, 31-54.

Ramsey, P. A. (1993). Sensivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Resnick, L. B., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 301-328). Boston: Kluwer Academic Publishers.

Shepard, L. A. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 301-328). Boston: Kluwer Academic Publishers.

Slavin, R. E., & Madden, N. A. (1991). Modifying Chapter 1 program improvement guidelines to reward appropriate practices. *Educational Evaluation and Policy Analysis*, *13(4)*, 369-379.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, *20(5)*, 8-11.

Sternberg, R. J. (1985). *Beyond IQ*. New York: Cambridge University Press.

Wigdor, A., & Green, B. F., Jr. (1991). *Performance assessment for the workplace*. (Vol. 1). Washington, DC: National Academy Press

Winfield, L. F., & Woodard, M. D. (1994). *Assessment, equity, and diversity in reforming America's schools* (CSE Technical Report 372). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California Graduate School of Education, Los Angeles.

*U.S. Government Printing Office: 1996 - 409-949/50264