DOCUMENT RESUME

ED 466 499 TM 034 238

AUTHOR Redfield, Doris
TITLE Critical Issues in Large-Scale Assessment: A Resource Guide.
INSTITUTION Council of Chief State School Officers, Washington, DC.; AEL, Inc., Charleston, WV.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
ISBN ISBN-1-884037-69-0
PUB DATE 2001-01-00
NOTE 118p.; Sponsored by the Technical Guidelines in Performance Assessment (TGPA) of the State Collaborative on Assessment and Student Standards.
CONTRACT R279A50006
AVAILABLE FROM Council of Chief State School Officers, Attn: Publications, One Massachusetts Ave., N.W., Suite 700, Washington, DC 20001 ($20). Tel: 202-336-7016; Fax: 202-408-8072; Web site: http://www.ccsso.org/publications.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Educational Assessment; Elementary Secondary Education; Reliability; *State Programs; *Student Evaluation; Test Bias; Test Construction; *Test Use; *Testing Programs; Validity
IDENTIFIERS *Large Scale Assessment; Large Scale Programs

ABSTRACT

The purpose of this document is to provide practical guidance and support for the design, development, and implementation of large-scale assessment systems that are grounded in research and best practice. Information is included about existing large-scale testing efforts, including national testing programs, state testing programs, and collaborative initiatives of states and organizations working to address issues related to large-scale testing. The guide also describes resources for additional technical information, especially research findings related to reliability, validity, fairness, and bias. Also discussed are critical issues in large-scale assessment and the trade-offs involved in making decisions about these issues. The document is designed so that directors of large-scale assessment programs can select and use sections as appropriate. The guide contains these chapters: (1) "Overview"; (2) "How Different Purposes for Assessment Systems Make a Difference"; (3) "Matching Tests to Purposes and Uses: Accountability vs. Instructional Planning"; (4) "What Should Be Measured? When? To What Extent?"; (5) "How Many Tests Make an Assessment System?"; (6) "Sampling"; (7) "Norm-Referenced versus Criterion-Referenced Test Results"; (8) "Test Formats"; (9) "Test Identification and Development"; (10) "Test Preparation"; (11) "Scoring the Tests: Reliability, Rubrics, and Reality"; (12) "Validity--Making Accurate Inferences from Test Results"; and (13) "Special Populations." (Contains 8 tables, 11 figures, and 14 references.) (SLD)

TM

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

SCASS

Critical Issues in Large-Scale Assessment: A Resource Guide

Doris Redfield
AEL, Inc.

SCASS TGPA
Technical Guidelines for Performance Assessment

A publication of the Council of Chief State School Officers

Critical Issues in Large-Scale Assessment: A Resource Guide

Doris Redfield
AEL, Inc.

Chair, Editorial Board
State Collaborative on Assessment and Student Standards:
Technical Guidelines for Performance Assessment

Council of Chief State School Officers
State Education Assessment Center
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone: 202-408-5505

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues. Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center is a permanent, central part of the Council of Chief State School Officers. The Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA). The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

Council of Chief State School Officers

Gordon M. Ambach
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards

ISBN 1-884037-69-0

This report was prepared for submission under contract with the Council of Chief State School Officers. The preparation of this report was financed with funds provided by the United States Department of Education, Office of Educational Research and Improvement, Grant number R279A50006.

The views and opinions expressed in this report are not necessarily those of the United States Department of Education, the Council of Chief State School Officers, or states participating in the SCASS TGPA.

This report is one of a collection of publications sponsored by the Technical Guidelines in Performance Assessment (TGPA) consortium of the Council of Chief State School Officer's State Collaborative on Assessment and Student Standards (SCASS). The report was developed with the assistance of a number of people, as noted below.

* In addition to the members of the Editorial Board, Frank Evans made significant conceptual contributions to the sections on alignment. Martha Thurlow authored the chapter on special populations. Phoebe Winter, Project Director of the TGPA SCASS, ably drafted early chapters, and guided and advised the project throughout. Untold credit for this document goes to Phoebe, members of the Board who endured and contributed to countless revisions, and members of the TGPA SCASS.

† Formerly

Contents

List of Tables

List of Figures

Chapter I. Overview

Tests[1] have traditionally been thought of as pencil-and-paper instruments for determining student knowledge or skill in a particular content area such as mathematical computations or spelling.

While the term **assessment** is commonly used as a synonym for test, an assessment is usually broader or more inclusive than a single test. Assessments may include a number of tests as well as other measures. For example, in assessing a student's knowledge and skill relative to language arts, the student might be asked to edit a document, write an essay, read a passage and answer a series of questions, and defend a position in a debate. Together, these "tests" form a fairly comprehensive assessment of the student's grasp of language arts and may be used to make evaluative judgments and/or inferences about the student's level of knowledge and skill relative to the content assessed.

Large-scale programs are those that test or assess relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale programs are in contrast to tests and other assessments administered on a smaller scale, for example, by classroom teachers for instructional purposes.

Purpose

The purpose of *Critical Issues in Large-scale Assessment: A Resource Guide* is to provide practical guidance and support for the sound design, development, and implementation of large-scale assessment systems—systems that are grounded in research and best practice. Hence, the following kinds of information are included:

- information about existing large-scale testing efforts, including national testing programs, state testing programs, and collaborative initiatives such as the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards (SCASS), whereby states that are facing common issues join forces and work together to address these issues;
- resources for additional technical information, particularly research findings on issues related to reliability, validity, fairness, and bias; and
- discussions of critical issues, examples of how the issues might be approached, and trade-offs involved in decision making relative to the issues.

Uses

The document is designed so that directors of large-scale assessment programs can select and use sections of the document, as appropriate, with their various constituencies. For example, testing directors are often called upon to provide information to policymakers, test developers, program evaluators, researchers, Title I coordinators, teachers, parents, and the press.

[1] Words in bold are defined in the chapter glossary and in the comprehensive glossary at the end of the document.

Critical Issues in Large-Scale Assessment: A Resource Guide is not a handbook. It is meant as a resource. Users are encouraged to apply what is useful, when appropriate, and in a manner that is beneficial.

Neither is *Critical Issues in Large-Scale Assessment: A Resource Guide* an answer book. That is because there is no one right way to approach the issues. Each approach has its trade-offs. The intent is to delineate key issues and to specify associated trade-offs.

Background and Rationale

At least five trends have fueled the development of *Critical Issues in Large-Scale Assessment.*

1. In recent years there has been a shift in the degree of reliance upon norm-referenced tests toward greater inclusion of criterion-referenced tests. This trend is largely fueled by state and national emphases on standards-based systems of curriculum and assessment that require tests designed around a particular set of curriculum objectives or learning standards. This does not mean that norm-referenced tests are no longer useful for their intended purposes. Neither does it mean that all tests are singularly used in a norm-referenced, criterion-referenced, or standards-based manner. Some tests are used in more than one way, and many assessment systems rightfully include several types of tests.

   The term "**norm-referenced**" does not indicate a test's format. For example, norm-referenced tests need not be restricted to multiple-choice formats and questions that require simple recognition and recall of facts. They may include constructed-response items and require complex reasoning and problem solving. Norm-referencing refers to the manner in which test results are interpreted. It indicates that individual test results or scores can be "referenced" (i.e., compared) to a "norm," the norm being the range of scores for all students who took the test when it was normed. Many commercial achievement tests have national norms for groups of students at different age or grade levels. Students can score above the national average, for example, on a norm-referenced test and still lack the skills necessary for success.

   While norm-referenced interpretations compare student performance to a norm, individual students' results on **criterion-referenced tests** are compared to an established criterion or definition of performance. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance such as proper use of conventions and logical, supporting ideas for a point of view in writing.

   Standards-based systems of assessment include criterion-referenced tests. In such systems, test items reflect a pre-established set of **content standards** that specify the knowledge and skills students are expected to acquire as a function of schooling. Results are then interpreted against a set of criteria or **performance standards** that define student performance relative to the content standards represented by the test items. An excellent reference on performance standards is *Meeting the Requirements of Title I: Handbook for the Development of Performance Standards* (Hansche, Winter, & Redfield, 1998).

   Teachers have always drawn upon criterion and standards-based assessment practices to guide instruction. However, the use of such practices is becoming more attractive to large-scale programs that are under legislative mandate to develop standards-based accountability programs (e.g., Improving America's Schools Act, 1994). **Accountability** brings a whole new dimension to the interpretation of test results. For test results to be used in making responsible decisions about such things as school accreditation and student graduation, the results must be accurate, fair, and unbiased. Hence, much anxiety

and activity surround the sound design, development, and implementation of large-scale, standards-based assessment programs that can be defensibly used to hold schools accountable for instruction and students accountable for learning.

2. Closely related to trend #1 is the fact that states and school districts are changing the ways they use and report assessment results. For example, in the past, assessment results were primarily reported to parents, teachers, and students as information about the performance of those students. The results were used to inform decisions about areas of curriculum and instruction needing improvement or which students might be grouped together for instruction. Except for the use of test results in the placement of students, or in decisions related to college admissions, the stakes were usually relatively low.

   More recently, the stakes for students have become more pervasive and weighty, moving beyond decisions of placement to decisions of sports eligibility, promotion, and graduation. And the application of high stakes (e.g., accreditation, promotion, tenure) has increasingly spread beyond students to include educators and schools.

   These shifts in the ways test results are reported and used are especially evident since publication of the National Commission on Excellence in Education's *Nation at Risk Report* (Gardner, 1983), the inception of the National Goals (Office of Educational Research and Improvement, 1991), and the adoption of Goals 2000 (U.S. Department of Education, 1994). More and more schools are being evaluated based on the absolute degree to which their students meet criteria specified by content and performance standards, rather than on the rank of average student scores relative to the average scores of other or similar schools on the same test. For information about how states are using large-scale assessment results, readers are referred to *Trends in State Student Assessment Programs* (Council of Chief State Schools Officers, 1999).

3. There has been a shift away from the nearly exclusive use of multiple-choice items in large-scale assessment toward the inclusion of more performance-based assessment formats. **Performance-based** usually means that, in responding to the assessment items, the student must do something beyond selecting a correct response. Hence, performance items are sometimes referred to as **constructed-response** items, whereas items such as multiple-choice items are often referred to as **selected-response** items. Examples of commonly used performance assessments include open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

   Of these formats, constructed-response items are the most commonly used for large-scale assessment purposes (CCSSO, 1999), probably due to considerations of cost, time required for comprehensive testing, and scoring reliability. A discussion of the accuracy and validity of performance assessments compared to multiple-choice assessments is provided by Ragosa (1998).

4. Views on the relationship between and among **curriculum** (*what* is taught), **instruction** (*how* the curriculum is taught), and large-scale assessment are changing. For this reason, it can be useful to distinguish between teaching a test and teaching to a test. **Teaching the test** is exactly what the phrase implies—teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

   By contrast, **teaching to a test** means teaching the broad-based knowledge and skills represented by a test's underlying content standards. In these times of accountability,

teaching to these standards increases the probability of students' success relative to any assessment based on the standards, not just the items on a particular form of a particular test. To do anything less than teach to the standards would be irresponsible. Clearly, the use of standards-based tests to influence what is taught highlights the importance of having defensible standards.

5. Related to item #4 is the increasing use of assessment results to model, inspire, monitor, and/or require instructional change. For example, large-scale assessments built upon state content standards are often viewed as representing an assessment ideal, with teachers being encouraged to emulate them during instruction. There is also an increasing emphasis on the use of assessment data in evaluating the quality of curriculum and instruction for purposes of refining them.

The Challenge

The most important criterion for a defensible assessment program is the extent to which it supports effective learning and valid opportunities for children to demonstrate their learning. This is true whether the system depends more heavily on the use of norm-referenced or criterion-referenced reporting of results, and whether it draws more heavily upon multiple-choice or performance types of item formats. This criterion of defensibility, indeed, raises critical issues for responsible policy making and implementation. Clearly educational policy making has increasingly serious implications for large-scale assessment and vice versa.

Chapter I Glossary

Accountability. The systematic use of assessment data and other information to assure those inside and outside the educational system that schools are moving in desired directions. Commonly included elements are goals, indicators of progress toward meeting those goals, analysis of data, reporting procedures, and consequences or sanctions. Accountability often includes the use of assessment results and other data to determine program effectiveness and to make decisions about resources, rewards, and consequences.

Assessment. Any systematic method of obtaining evidence from tests and other sources that is used to draw inferences about characteristics of people, objects, or programs for a specific purpose.

Constructed-response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Criterion-referenced. The reference point for interpreting test results using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to a norm or the performance of others.

Curriculum. What is taught.

Instruction. The teaching methods used to deliver the curriculum to students.

Large-scale. Assessment programs that test or assess relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale

programs are in contrast to tests and other assessments administered on a smaller scale, for example, by classroom teachers for instructional purposes.

Norm-referenced. Test interpretations whose scores are based on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Performance-based or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing of a musical composition, and portfolios showing samples of work over time.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrate the range of performance for each performance level; and (4) cut scores that separate one level of performance from another.

Reference population. The population of test takers represented by a test's norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or other characteristics at the time of testing.

Selected-response. Test items that require students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Standards-based systems of assessment. Include criterion-referenced tests. In such systems, test items reflect a pre-established set of **content standards** that specify the knowledge and skills students are expected to acquire as a function of schooling. Results are then interpreted against a set of criteria or **performance standards** that define student performance relative to the content standards represented by the test items.

Systems of assessment. Complementary components that, together, provide an accurate profile of student achievement.

Teaching the test. Teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

Teaching to a test. Teaching the broad-based knowledge and skills represented by a test's underlying content standards. Compared to **teaching the test,** it is not cheating.

Tests. In contrast to **assessments,** tests include a number of measures that help create a more complete picture or profile of performance, are usually single instruments or procedures such as quizzes, standardized measures, questionnaires, surveys, observations, checklists, and the like. Thus, tests are typical components of aligned systems of assessment.

Chapter I References and Resources

Council of Chief State School Officers. (1999). *Trends in state student assessment programs.* Washington, DC: Author.

Gardner, D. P., Larsen, Y. W., Baker, W. O., & Campbell, A. (1983). *A nation at risk: The imperative for educational reform.* An open letter to the American people. A report to the nation and the Secretary of Education. Washington, DC: U.S. Department of Education.

Hansche, L. N., Winter, P., & Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington, DC.

Office of Educational Research and Improvement. (1991). *Striving for excellence: The national education goals.* Washington, DC: U.S. Department of Education.

Ragosa, D. R. (1998, May). *Accuracy of individual scores and group summaries.* Professional development session for Council of Chief State School Officers; State Collaborative on Assessment and Student Standards, Durham, NC.

U.S. Department of Education. (1994). *Goals 2000: A world-class education for every child.* Washington, DC: Author.

Chapter II. How Different Purposes for Assessment Systems Make a Difference

Background

In the final analysis, we must ask the most fundamental of questions: "What do we want the results of our assessment system to tell us about our students?" The purposes and intended uses of assessment systems are central in making decisions about the assessment instruments and procedures to be included in the system. And the importance of clearly articulating the intended purposes and uses of assessment results cannot be overstated. This is because there are certain things that certain kinds of tests and assessments can and cannot do relative to the understood purpose. It is also sometimes impossible to meet all goals, given practical limitations of time, money, staffing, and the technology of testing. Clarity about the ultimate goal can help with decisions about compromises and trade-offs.

The importance of clarity can be illustrated by the relationship between **laws** and ensuing **policies** and **guidelines.** When legislation is perceived as unclear, compliance can be hampered. For example, even though IASA/ Title I **legislation** has been in place since 1994, states and localities are individually and collectively still trying to interpret the law in ways that (1) meet the spirit of the legislation and (2) are feasible and technically sound. State and local policy and guidelines have the opportunity to provide clarifying links to the state and federal law. They also have implications for the kinds of assessments that can be used as well as when and how they are used.

Not all assessments are driven by legislation. However, when the results of an assessment carry consequences, it is advisable to have enabling legislation. Such legislation can influence the process in ways that help ensure the **defensibility** of the assessments and can also provide resources for building defensibility into the assessments. A caution in the development of enabling legislation is that it be clear without being overly specific. It is best when the legislation can be clear about purpose and provide technical experts with the authority to propose **technically sound** details in keeping with the stated purpose.

Ultimately, discussions with those who will be affected by the assessments—students, parents, teachers, administrators, and workforce leaders, for example—are important to developing clear legislation, policies, and guidelines as well as responsive and responsible assessment systems. **Stakeholders** can be enormously helpful in determinations about (1) what an assessment system should do and (2) the level of investment they are willing to make in its success.

Critical Design Issues

1. What are *all* of the goals of the assessment program?
2. What interpretations do we want the assessment results to support?
3. What attributes will make the assessment system suitable for *all* students?
4. How will assessment results be used? By whom?
5. How will assessment results be reported?

The purposes and uses of an assessment system are central in making decisions about the assessment instruments and tasks included in the system. As implied at the outset of this

chapter, clear statements about purpose and use will influence more technical decisions ranging from content coverage to the defensible degree of technical soundness required for **high-stakes** decision making. While one might argue that all assessment results carry high stakes such as social stigma, high stakes generally refer to binding consequences to schools (e.g., accreditation) and students (e.g., graduation).

1. What are the goals of the assessment program?

The goals of assessment programs can vary according to degree of explicitness, specificity, and audience. A program's goals may include few, some, or many of the following:

- improve student learning on state or district content standards through improved instruction
- influence curriculum content and/or teaching methods
- motivate students and/or teachers
- inform the public about school performance
- verify school-based information, such as teacher-assigned grades
- influence decisions about
  - √ student promotion and/or graduation
  - √ teacher effectiveness
  - √ school quality
- Provide data for comparisons of
  - √ students, schools, and/or states to other students, schools, states, and/or nations
  - √ students, schools, and/or states to past performance
  - √ students, schools, and/or states to a pre-set criterion or standards of proficiency or excellence

Each possible goal has implications for the design of the system. Each may also have implications for policy or implementation, particularly with regard to issues of cost, time, and staffing.

If a goal is to *improve student learning on state or district content standards,* there needs to be an agreed-upon set of content standards. These standards should reflect a consensus of what the constituents of the state or district consider as important knowledge and skills for students to learn as a function of schooling, the breadth of such knowledge and skills, and the depth to which they are to be learned. The assessments need to align with these content standards.

The structure of content standards—the grade levels and subject areas covered—can differ according to the underlying goal. For example, some states have standards in core subject matter areas such as reading, writing, mathematics, science, and history for each grade, kindergarten through 12. Presumably this is because they want to ensure that certain content and skills are taught and learned in each of the designated content areas at specific grade levels. Hence, they may choose to administer certain tests at each grade level.

Other states have standards for **benchmark** grades only, such as grades four, eight, and eleven. This structure allows flexibility as to when the designated content and skills are taught and learned, yet still requires that they be accomplished within the prescribed timeframes. Benchmark testing is adequate when the goal is a general measure of the effectiveness of states, school districts, or schools relative to the tests.

When the standards cover only benchmark grades, it limits the grade levels at which large-scale assessments can be developed and administered. In such cases, local school districts sometimes develop and administer tests for other grade levels. Regardless of who develops any off-grade tests, if they will be used for making decisions about students, personnel, or schools, they must be technically sound. Readers are referred to the *Standards for Educational and Psychological Testing* (American Educational Research Association, et al., 1999) for information about required levels of technical rigor.

Influencing curriculum content and/or teaching methods can be accomplished through careful crafting of the content standards. Performance standards that include clear descriptions and examples of the range of student performances within each performance level can also influence the depth of teaching. Assessment systems that are carefully aligned with content standards, measure the full breadth and depth of these standards, and allow students at all levels of performance to demonstrate their proficiency can have a strong, positive impact on curriculum and instruction.

If a goal of the assessment system is to *increase and sustain students' motivation to learn and teachers' motivation to teach,* then it is important to include these stakeholders in the design process. This can range from participation in the drafting of content standards, to review of potential items, to professional development programs for teachers, and recognition of students for desired levels of achievement. Sometimes sanctions such as not promoting students or not accrediting schools are applied as motivators. For the most part, sanctions are sticks rather than carrots. However, sanctions can be useful if they include provisions for support, such as remediation programs for students and school improvement programs for schools.

The results of assessments are *used by the public to judge how well schools are doing.* Results often are printed in newspapers and used by real estate agents to influence the sale of properties. If a goal is to inform the public, then careful attention must be given to the kinds of information the assessments are designed to provide. The reporting aspect of the system must be designed to support the kinds of things we want to be able to say. Many states now use school report cards. However, they vary from state to state, depending upon their purposes. School report cards might contain information about how well, on the average, students in each grade tested performed on each content area tested. Whether this information is shown in relationship to a national norm or a state standard depends on the goals of the assessment program. Most school report cards also contain other information that could be used to judge the quality of a school, e.g., information about attendance or dropout rates. The 1999 issue of *Quality Counts* (Olson, 1999) contains a wealth of information about the variety of ways states have approached school report cards.

It can be useful to compare how students perform on large-scale assessments to *how well they are doing on other measures of school achievement such as teacher-assigned grades.* Such comparisons can provide important information for influencing policy changes. The change process might involve a careful look at grading practices, the match between the assessments and the underlying content standards, the match between the curriculum and the content standards, or the efficacy of the performance standards applied to student performance on the assessments.

If the goals involve *making decisions about student promotion and/or graduation, teacher effectiveness, and/or school quality,* then the following issues must be considered.

- If a goal of the assessment system is to *influence decisions about student promotion and/or graduation,* the technical soundness of the assessment instruments is of utmost importance. The tests must be fair and free from bias, reliably scored, and valid for the inferences or interpretations that will be made on the basis of results. High stakes tests, such as those used to determine promotion or graduation, also carry ethical and legal implications:

  1. Students must have been provided with adequate opportunity to learn the knowledge and skills covered by the content and performance standards upon which the test is based.

  2. The assessment process should accurately reflect the knowledge, skills, and level of cognitive demands represented by the content standards. For example, if the content standard calls for "description," then the assessment should also require description, rather than another skill such as recall or problem solving.

3. Students must be provided with opportunities to demonstrate their knowledge and skills. This means that the assessments must be accessible to all students.

4. Students must not fail a grade or be denied graduation on the basis of only one opportunity to take the test.

5. A single test score must not be the sole basis for making a decision about a student.

There are also cost implications associated with high-stakes assessments for students. Remediation and multiple testing opportunities can add significant costs. This is especially true if alternate forms of the test must be developed for "make-up" purposes or if the tests include performance items such as essay responses. Performance items are more costly to score than machine-scorable multiple-choice tests.

- The goal of using student achievement data to *influence judgments about teacher effectiveness* has long been controversial. Yet, there can be little denial that teachers are significant factors in students' learning experiences. The key question is not, "Should teachers be held accountable for student learning?" Rather, it is, "To what extent should student achievement on large-scale assessments be used to evaluate teacher effectiveness?"

  Clearly, using a single measure of student achievement to evaluate teachers or using student achievement as the only measure in evaluating teachers is indefensible. Teachers, after all, are not with all students for all of their learning, and not all that a student does or does not learn is the function of one year's worth of schooling. Texas is one state that provides an interesting perspective on the inclusion, but not exclusive use, of student achievement in the evaluation of teachers. Teachers participate in the Texas Professional Development and Appraisal Program (Texas Education Agency, 1998) wherein one domain of their evaluation includes students' average improvement on the statewide student assessments.

  Also, holding a teacher at one particular grade responsible for student attainment of skills that require several years of instruction to acquire is patently unfair. For example, it would be unfair to hold a third-grade teacher solely accountable for how well a student reads, since reading involves an accumulation of skills, some of which should have been taught in previous grades. Some states approach this issue by using measures of school accountability rather than measures of teacher accountability. In such cases, the results from students of all teachers who are expected to contribute to specified learning outcomes are combined to derive a measure of school accountability.

- The issue of *using assessment results to measure school effectiveness* is closely linked with the issue of reporting to the public. Implicitly or explicitly, assessment results are commonly viewed as the quintessential measure of school effectiveness. Here we carry the issue a step further. Do we have performance standards for schools? How well must the students in a school be doing for the school to be considered effective, accreditable, or in crisis? What measures besides student achievement on standardized tests will be considered in judging the effectiveness of a school?

  A typical method for dealing with these issues is to use an indicator system whereby different indicators (e.g., attendance, student test scores) are assigned weights in the overall system. The student achievement indicator might be something like "70% of the students tested must meet the proficient level relative to the content standards on which they are tested." Information on state indicators nationwide is available from the Southern Regional Education Board (Creech, 1998) and the Council of Chief State School Officers (Blank, Manise, & Brathwaite, 1999).

  Once we answer these questions, we have entered the realm of explicit accountability. What will be the consequences for schools that do not meet the standard? Will there be

rewards for schools that do? We know students must be given multiple opportunities to be assessed before the stigma of failure can be defended. How will we generalize the principles of multiple opportunity to succeed and remediation to schools? What will be the consequences of repeatedly not meeting the standard for school quality? The answers to these questions—whether they involve sanctions, rewards, or the implementation of school improvement plans—can have substantial cost implications.

Whether or not the ability to make *comparisons* is a stated purpose of assessment systems, comparisons are nearly always made in practice. Being clear about what comparisons we do and do not wish to make can go a long way toward guiding the design and development of assessment systems and how the results are reported.

- If we wish to *compare students, groups of students, schools, and/or states to other students, schools, states, and/or nations,* it is necessary to use a standardized test that has been designed for such comparisons. To compare student performance from one state to another, the same kinds of students from the states being compared must have taken equivalent tests at similar times and under similar circumstances. This is why many large-scale assessment programs include a nationally normed test component, even if they have state content standards that do not match the nationally normed test.

  In other cases, states that do not wish to make comparisons below the state level may include results of the National Assessment of Educational Progress (NAEP) as part of their system. This test allows comparisons on the basis of the percentage of students performing at the basic, proficient, and advanced achievement levels to performance of the nation as a whole and from one state to another. However, to date, not all states participate in NAEP.

  Whether the comparisons are to be made on the basis of a norm or a performance standard, it is clear that the results yielded by a test designed around a specific set of content standards cannot be directly compared to the results of tests designed around a different set of standards.

- *Comparing students, schools, and/or states to past performance* requires the use of **scaled scores,** sometimes referred to as standard scores, which can be used from year to year with different groups of students and still carry the same meaning. Grade-equivalent and stanine scores are examples of scaled scores.

  To illustrate the usefulness of scaled scores, imagine that two forms of a test are used in two different years. The tests cover the same content standards, but the items are a little different so that **test security** can be maintained. Further consider that one of the tests is slightly more difficult than the other. If **raw scores** were reported, a comparison of the results between the two tests would be unfair. However, if the tests are **equated** and scaled (using statistical procedures that account for the differences in difficulty between the two test forms), then the scaled scores can be directly compared.

  Of course, no matter what kinds of scores are used to make comparisons over time, a critical substantive issue remains: "How much improvement represents actual improvement?" At the student level, measuring improvement requires pre- and posttesting on alternate forms of the same assessment or on an assessment that has been designed to accurately measure changes in learning over repeated administrations. It is necessary to use equated tests if comparisons will be made across time (e.g., from one year to the next).

  At the school level and above, pre- and posttesting may not be required, so long as it is understood that different cohorts of students are being compared. For example, comparing the scores of this year's third-grade students to the scores of last year's third-grade students provides an indicator of change on the third-grade test for different groups of

students. This can be useful information if data for more than two years are available and if something is known about the stability of the population.

- Finally, if a goal of the system is to *compare students, schools, or states to a preset criterion or standard,* it is necessary to have standards of performance as described previously. Many large-scale assessment systems accommodate for more than one type of comparison. For example, a state assessment system may include several components. Students in selected grades take a nationally norm-referenced standardized achievement test. This allows students, schools, and the state to be compared to national averages. However, these results are not used as part of the school or student accountability program. Instead, a standards-based test, aligned with the state's content standards in mathematics, language arts, and science, is administered to students in one grade in elementary, middle, and high school, and standards-based tests are administered at the end of certain courses in high school. Students must reach certain levels of performance on the end-of-course tests in order to graduate. School accreditation is based on the proportion of students reaching the proficient level or higher on the standards-based tests.

An issue that has not yet been addressed as a critical design issue concerns the *inclusion of students with special needs,* such as students with learning disabilities and students of limited English proficiency. Three aspects of inclusion are (1) which students get assessed, (2) what testing **accommodations** are allowable, and (3) which students are included for school accountability purposes. Due to the importance of this topic, it will be treated separately in Chapter XIII.

2. What do we want to be able to say on the basis of the assessment results?

Surely, what we want to say about assessment results always involves student scores. However, student scores can be described on at least two dimensions: (1) level of aggregation (e.g., classroom, school, district, state, nation); and (2) compared to a reference point (e.g., norms vs. performance criteria or standards). Table II-1 illustrates the intersection of these two dimensions and may help with making decisions about what kinds of tests to use and how to report the results.

Table II-1. Examples of Ways to Report Student Scores

Levels of Aggregation	Interpretations		
	Compared to a Norm	Compared to a Performance Standard	Compared to Prior Performance (Baseline)
Student			
Classroom			
School			
District			
State			
Nation			

If we check off the cells in the table that reflect what we want our assessment results to tell us, we can design the system accordingly. For example, if we want to be able to compare student level results to a performance standard, we know that we need assessment instruments that are valid at the individual student level, that we need to assess all students for whom we want a score, and that we need defensible performance standards.

If we want our results to tell us how well schools are doing relative to a particular criterion, but we do not need individual student scores for purposes of diagnosis or the application of consequences, then we may not need to test every student on every test item.

Provided a school has enough students, we may be able to use sampling procedures that allow either all students or some students to take different subsets of the items. This idea applies at the district, state, and national levels as well. Using sampling, more aspects of the curriculum can be assessed. It typically takes hundreds or thousands of items to assess an entire curriculum domain. In this case, students would not get individual scores; even if they did, it would be impossible to compare them. The National Assessment of Educational Progress (NAEP) is based on sampling procedures where only some students are tested and not every student who is tested takes the same items (http://nces.ed.gov/nationsreportcard/).

If we want our results to tell us how much improvement students, schools, districts, or states are making, we need baseline information, i.e., prior results against which subsequent performance can be compared. As noted previously, the assessments must have technical properties that allow for valid comparisons over time. Baselines can be used in combination with other forms of test interpretation. For example, in the early years of an assessment program, we may want to know how much progress schools are making toward a performance standard even though the standard is not required for accreditation purposes for several years.

Finally, if we want to be able to report student performance on subscales of the total test, such as computation and problem solving in mathematics, the test must contain enough items in each reporting category to yield reliable scores. If subscale scores are provided, the technical manual should be consulted regarding the level of reliability associated with the subscale scores.

3. What attributes will make the assessment system suitable for *all* students?

While this topic warrants its own chapter, a few issues critical to the design phase of an assessment system are addressed. The foremost consideration is that the assessment system be designed so that the characteristics of the *entire* student population are taken into account. Factors to consider include the full range and depth of knowledge and skills to be included in the content standards, appropriate assessment techniques for students with limited English proficiency or for students with various disabilities, and the inclusion of content and contexts appropriate for the cultural and ethnic diversity represented by the student population. In some cases, testing **accommodations** or **alternate assessments** may be warranted. These issues are further discussed in Chapter XIII. Readers are also referred to the work of the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards (SCASS). Three of these collaboratives are working on issues related to the inclusion of students with special needs in large-scale assessment programs: the Limited English Proficiency SCASS, the Comprehensive Assessment Systems for Title I SCASS, and the Assessment of Special Education Students SCASS.

4. How will assessment results be used? By whom?

In thinking about this, it is useful to consider the intersection between (1) who is interested in the results of large-scale student assessments and (2) how these interested parties might want to use the results. The goals of the assessment program should clarify the match between the program's underlying purposes and the possible uses as shown in Table II-2. Check marks indicate the uses for which particular interest groups might want assessment results.

In examining the table, consider parents and members of the community as an example. We can see that they are interested in factors affecting how education tax dollars are spent and the quality of neighborhood schools. We can also see that parents are likely interested in some of the things that interest local school boards and colleges. It is clear that the assessment system design must be attentive to the needs of parents. It appears that parents will be interested in having norm-referenced information on their children as well as the schools their children attend. They may also be interested in the rigor of the curriculum as

Table II-2. Possible Uses of Assessments by Various Interest Groups

How Used?	Who?					
	Teachers	Administrators & Curriculum Supervisors	Local School Boards	Parents/ Community/ Taxpayers	State/Federal; Legislators	Colleges/ Universities
Guide Instruction	√	√	√	√		
Program/School Evaluation	√	√	√	√	√	
Curriculum Revision	√	√				
Personnel Placement		√	√			
Professional Development	√	√	√	√	√	
Funding Decisions		√	√	√	√	
Housing/Location Decisions				√		
Acceptance into programs/ schools				√		√
Placement	√	√		√		√

evidenced in the content standards and reflected in the accompanying standards-based assessment. They will likely want their children's schools to be held to a standard of quality. All of these factors have implications for designing valid assessments. Similar analyses could be applied to groups other than parents and community members.

5. How will assessment results be reported?

The considerations discussed relative to issue #4 have implications for issue #5. Determining who is interested in using the results for particular purposes can guide our thinking about how to present the results. While individual student results should never be reported publicly, the students, their teachers, and their parents must have access to the results in readily understandable formats that tell them what they need to know. For example, a student report that goes home to parents might include the student's scaled scores and percentile ranks on a nationally norm-referenced test taken by fifth-grade students in the content areas of reading comprehension, word recognition, mathematics computation, and mathematical problem solving. This report might also show whether or not the student met the criterion for proficiency in the areas tested on a state-developed standards-based test. Either of these might show improvement since the student was last tested, providing that the assessment instruments allow for such comparisons.

It can be useful to report information about student averages for an entire school to parents and the community. In such cases, the focus of interest will likely be comparisons to a national average or to the performance of other communities within the state relative to the state's standards, as well as the degree to which students are meeting pre-set standards of performance.

An important issue will be how to include the results for students with special needs. This is an issue at both district and state levels. The determining factor is often whether or not the results will be used for accountability purposes. IASA/Title I and the Individuals with Disabilities Education Act (IDEA) legislation makes it clear that schools are responsible for the education of *all* students, and reports that do not include results from *all* students do not allow for valid inferences about how well schools are educating *all* students.

Some states report results in two ways: (1) a report showing the average results for all students; and (2) average results for various categories of students such as students with learning disabilities, LEP students, and students receiving free or reduced meals at school. At the district and state levels, it is also useful to disaggregate results by ethnic or cultural group to determine if certain populations are being underserved.

Of course, if there are a small number of students in any category, results should not be publicly reported. The students' privacy should be protected at all costs. Most states have policies governing the number of students that must be in a category before results can be publicly reported. Jaeger and Tucker (1997) recommend that results for groups containing fewer than 10 students not be reported publicly.

Experience is also teaching us that special forms of reporting are helpful in dealing with busy policymakers and members of the media who are keenly interested in the results but do not have a technical background and are up against deadlines. Not only are short, direct, clear reports helpful in these instances, it can also be useful to hold briefing sessions prior to the release of results. Some large-scale assessment directors release **embargoed** copies of results to the media so that they have more time to consider the information before meeting critical deadlines. It is also wise to release embargoed, preview copies of results to anyone who may be affected by them, e.g., school superintendents, so that they may prepare to respond accurately to questions that will undoubtedly arise.

Illustrative Scenario

The state described in the scenario (see pp. 16-17) reflects a composite of legislation and subsequent actions from a number of states. The particulars have been chosen to illustrate particular issues. The intent of the scenario is to demonstrate how the purpose of a system drives policy decisions which, in turn, influence the design, implementation, and defensibility of an assessment system relative to its purpose.

Legislative decisions constitute the law. They can range in level of specificity.

The state board of education has responsibility for developing policies for implementing the law. The less specific the law, the more room for interpretation or misinterpretation on the part of the board.

The state department of education, led by the state superintendent of public instruction is responsible for implementing the board's policies. In ideal situations, the superintendent and department staff advise the board throughout the policy-making process.

Local school boards are responsible for local policy in accordance with state law and policy. For example, the state board may choose to give local school boards policy-making authority over certain issues that may require local variation. School district superintendents and their staffs, including principals and teachers, are responsible for implementing local policy, in keeping with state policy.

The italicized words and phrases in the scenario indicate critical issues in large-scale assessment as they may affect this scenario.

Summary and Lessons Learned

The state described in the scenario developed its content standards and established its regional resource centers four years ago. It began administering a new norm-referenced test three years ago. Two years ago, the standards-based tests were piloted and performance levels and associated cut scores were established. Beginning this year, students who will be high school seniors in five years must meet the criteria for passing end-of-course tests, and schools must meet the criteria for accreditation. These dates were chosen to allow for staff development and adequate exposure to the new standards and assessments before imposing high-stakes consequences.

This state was fortunate in that its legislative mandates illustrate a balance in specificity. The legislation is clear about purpose, provides guidance regarding its intent, and provides the board with authority to consult with technical experts in designing and implementing a program that meets the intent of the legislation.

Clearly, this state board took great care in designing a system to meet the intent of the

Illustrative Scenario

Key Legislative Decisions

- The general assembly and the board of education believe that the fundamental *goal* of the public schools of the state must be to enable *each* student to develop the skills necessary for success in school and preparation for life, and that the quality of education is dependent upon an appropriate working environment, benefits, and salaries necessary to ensure the availability of highly qualified instructional personnel, and adequate commitment of *resources.*
- The board of education shall establish *educational objectives* to implement the development of skills that are necessary for success in school and for preparation for life in the years beyond. The board of education may, from time to time, revise these educational objectives to maintain *academic rigor.* The Board shall seek to ensure that the educational objectives are consistent with the world's highest educational standards. *These objectives shall include,* but not be limited to, basic skills of communication, computation, and critical reasoning, and the development of *personal qualities* such as social responsibility, self-management, integrity, and honesty.
- In order to provide appropriate opportunity for *input* from the general public, teachers, and local school boards, the board of education shall conduct public hearings prior to establishing the educational objectives.
- With such funds as are available, the state board of education may prescribe *assessment methods* to determine the level of achievement of the educational objectives by *all students. Such assessments shall evaluate* knowledge, applications of knowledge, critical thinking, and skills related to the educational objectives being assessed. The board, with the assistance of independent testing *experts,* shall conduct a regular analysis and *validation* process for these assessments.
- Local school boards shall develop and implement *programs of prevention, intervention, or remediation* for students who are educationally at risk, including, but not limited to, those whose scores are in the bottom national quartile on the state-adopted nationally *norm-referenced test* and *those who do not pass the state tests* based upon the state's educational objectives in grades 3, 5, 8, and 11. District superintendents shall require such students to take special programs of prevention, intervention, or remediation, which may include attendance in public summer school programs. Based on the number of students attending and the State's share of the per pupil costs, additional *state funds* shall be provided for summer school and other remediation programs.
- In establishing course and credit *requirements for a high school diploma,* the state board shall include in the student outcome measures end-of-course or end-of-grade tests for various grade levels and classes, as determined by the board. These assessments shall include, but need not be limited to, end-of-course or end-of-grade tests for English, mathematics, science, and social studies.
- Local school districts shall also implement the following: (1) programs in grades K-3 that emphasize developmentally appropriate learning to enhance success; (2) career education programs infused into the K-12 curricula; (3) *early identification* of students with disabilities and enrollment of such students in appropriate instructional programs consistent with state and federal law; (4) *early identification* of gifted students and enrollment of such students in appropriately differentiated instructional programs; and (5) a plan to make achievement for students who are educationally at risk a districtwide priority, which shall include *procedures for measuring the academic achievement progress of such students.*
- The state department of education shall provide to the local school districts *technical assistance* in the delivery of those support services necessary for the operation and maintenance of the public schools including, but not limited to, in-service training of staff.
- The general assembly recognizes the need for the board of education to prescribe requirements to ensure that student progress is measured and that school boards and school personnel are *accountable.*
- The superintendent of public instruction shall develop and the board of education shall approve criteria for determining and recognizing educational performance in the state's public school districts and schools.

Key Policy Decisions

Following are the state board's policy decisions relative to the key legislative decisions listed above.

- By using the term "*each* student," the *goal* of the legislation clearly indicates that every student is to be educated. Hence, testing samples of students would be in conflict with the goal. Testing all students carries cost implications and may imply using testing formats that are relatively inexpensive to administer and score, e.g., multiple-choice type items.
- The legislation also recognizes the state's responsibility to commit *resources* that can support an appropriate working environment, including benefits and salaries necessary for attracting and maintaining high-quality instructional personnel. The board's policy role is to provide accurate information to lawmakers for making decisions about funding priorities and to develop budget requests that reflect these priorities.
- While the law requires that *educational objectives,* i.e., content standards, be developed, it gives the state board the authority to determine what those standards will be. It is clear that the

Key Policy Decisions *(continued)*

content standards are expected to be rigorous because the law gives the board authority to revise them from time to time in order to ensure their *academic rigor.* In other words, the standards are to go beyond basic skills or minimum competencies. This decision has important implications for related decisions about who shall be held accountable for what and the consequences for not meeting the accountability or performance standard.

- The law is clear that the content standards must be developed with *public input;* that in *grades* 3, 5, 8, and 11 they *must include* the *content areas* of English/language arts and mathematics; that they must call for *critical thinking and practical problem solving;* and they must foster the development of *personal qualities* such as self-management and honesty. The law, here, definitely provides some challenges for the board, which must decide whether to restrict testing to the specified content areas and grade levels, what is meant by critical thinking and problem solving, and the implications of personal quality standards for assessment. The board ultimately decided that the state's standards should reflect rigorous academic content and be measurable. Hence, the large-scale assessment program does not include measures of personal qualities.
- The legislation was not clear about the amount of funds available for the state board's prescription of assessment *methods.* Ultimately, in *consultation with assessment experts,* the board determined to develop a series of standards-based tests consisting primarily of multiple-choice items, with some constructed-response items in each content area. While the board, based on public input, determined to develop content standards at every grade level, they decided to request funding to develop assessments in the content areas of English, mathematics, science, and social studies at grades 3, 5, 8, and 11, and at the end of certain high school courses. The decision for using primarily multiple-choice items in the tests was based on cost concerns as well as concerns about the reliability of scoring more performance-based assessment formats. While the board was confident that the state would not fund the development, administration, and scoring at all grade levels, the board placed no restrictions on additional local testing. This, of course, raised the issue of what local school districts might use for testing grades not included in the new testing program.
- The board was given the legal responsibility for developing and implementing *programs of prevention, intervention, or remediation* for students who are educationally at risk. Realizing the associated cost implications, the legislation provides funding support for such programs. The legislation gave further clues as to its expectations for the state assessment program by defining students at risk as including, but not limited to, those whose scores are in the bottom national quartile on the state-adopted nationally *norm-referenced test* and *those who do not pass the state's standards-based tests.* Clearly, the state testing program would need to include a nationally normed test. Decision issues included how much testing was appropriate for particular grade levels or students. It was decided to adopt a nationally norm-referenced test that most closely matched the state's content standards and to administer abbreviated forms of the test in grades 4, 8, and 11 for the content areas of English and mathematics only.
- The legislation also clarifies that the standards-based tests should bear high stakes by establishing that course and credit *requirements for a high school diploma* include student outcome measures via end-of-course or end-of-grade tests for various grade levels and classes, as determined by the board. Legally, the board was bound to include, but not necessarily limit such testing to the content areas of English, mathematics, science, and social studies. The board decided to make passing four end-of-course exams, one in each of the designated content areas, a requirement for graduation. Among other issues, such as differences in instruction, this raises the issue of determining how to fund and administer multiple opportunities for students to take the tests before denying a diploma.
- In specifying the *early identification* and provision of services for students with disabilities and gifted students, the law is clear in its intent that the board shall develop *procedures for measuring the progress of such students.* The board directed the department to work with testing experts to develop methods for assessing gifted and other special needs students. The only policy decision to date is that special education students must be tested according to the specifications of their Individual Educational Plans (IEPs). In some cases this means that test items are read aloud to students, that students dictate responses, or that visually impaired students use large print versions of the tests. The accommodations made in testing must be those also made during instruction and must not alter the content being assessed. This state, like all states, is under federal mandate to develop assessments that are appropriate for students who are unable to take existing tests even if accommodations are applied.
- To meet the requirement for providing *technical assistance* to local school districts, the state board of education requested state funding for the department to establish, staff, and implement regional resource centers with expertise in curriculum, instruction, assessment, and technology to provide technical assistance, such as staff training, upon request.
- *Accountability* was described by the legislature as board-prescribed requirements for holding local school boards and schools accountable for student progress. Policy-wise, this has translated into: (1) prescribing four levels of proficiency on the standards-based tests—advanced, proficient, approaching proficient, and below proficient—along with associated cut scores for each of the standards-based tests; (2) requiring that students meet the proficient level of the performance standards for four end-of-course tests in order to receive a high school diploma; and (3) establishing a performance criterion whereby schools must increase every year the proportion of students meeting the proficient level or higher and decrease the proportion scoring "below proficient" on the designated tests in order for the school to be accredited. The norm-referenced test results are not used for accountability purposes.

legislation. Key challenges faced by this state, four years after establishing its content standards, include: (1) delivering meaningful and high-quality technical assistance in a timely manner; (2) building the capacity of local school districts to meet the requirements of the new system; (3) continuing research to demonstrate and maintain the validity of the assessments; (4) handling public relations; and (5) testing all students in a fair and inclusive manner, especially those with limited English proficiency and students with disabilities.

Chapter II Glossary

Accommodations. (1) Changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To be appropriate, assessment accommodations must be those also made during instruction and must not alter the construct intended to be measured or the meaning of the resulting scores. (2) Specific changes in testing conditions, procedures and/or formatting that do not alter the validity or reliability of a state standard. Policies and procedures must ensure that the accommodations do not compromise the security of the test and are consistent with the student's Individualized Educational Plan (IEP), 504, and/or Limited English Proficient (LEP) plan. Accommodations can be made available for use in both instruction and statewide assessments. These may include accommodations for scheduling, setting, equipment, presentations, and/or responses. Allowable accommodations for states' assessments are generally identified in State Education Agency (SEA) documentation. (3) Alteration in *how* a test is presented to the test taker or in how a test taker is allowed to respond; includes a variety of alterations in presentation format, response format, setting in which the test is taken, scheduling or timing, and/or specialized equipment required by the student. The alterations do not substantially change level, content, or performance criteria. The changes are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate what is known. (4) Change in *how* a student accesses information and/or demonstrates learning; does not substantially change the content, instructional level, or performance expectations; provides for equal opportunity to demonstrate knowledge and skills.

Alternate assessments. An approach used in gathering information on the performance and progress of students whose disabilities preclude them from valid and reliable participation in typical state assessments as used with the majority of students who attend school. Under the re-authorized Individuals with Disabilities Education Act (IDEA, 1997), alternate assessments are to be used to measure the performance of a relatively small population of students who are unable to participate in the regular assessment system, even with **accommodations** or **modifications.**

Benchmarks. Specific statement of knowledge and skills to be demonstrated at the end of a specified range of grades. For example, benchmark content standards may be set at the end of grades 4, 8, and 12 to specify standards to be met by the end of primary, middle, and high school grade ranges. Benchmarks are located on a performance continuum and are used as checkpoints to monitor progress from one level to the next.

Construct. The underlying theoretical concept or characteristic a test is designed to measure.

Defensibility. The technical properties of an assessment that make its use for a particular purpose appropriate. Such properties include validity, reliability, fairness, and lack of bias.

Derived scores or scaled scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Embargoed. Test results prohibited from release until a specified date/time.

Equated. Two or more forms of a test that yield equivalent or parallel scores for specified groups of test takers. Equating involves converting the score scale of one form of test to the score scale of another form so that the scores are equivalent or parallel.

Guidelines. Information and the description of procedures that can be used by local school districts in implementing state board policies.

High-stakes. Tests whose results have important, direct, or lasting consequences for examinees, programs, or institutions.

Laws. Legislative mandates that carry negative legal consequences when violated.

Legislation. The result of lawmaking activity; law.

Modifications. (1) Changes made in the content and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. (2) Changes in the administration of an assessment that may cause the construct being measured to differ from the construct as measured under standard administration conditions, or produce a score that means something different from scores yielded by the standard administration. Unlike *accommodations*, modifications may directly or indirectly compromise either the validity or reliability of the state standard. Modifications may compromise test security and therefore are not recommended for statewide assessments. Modifications are more appropriate for instruction and classroom tests and include a much wider range of supports and instructional scaffolding than do accommodations. Modifications can be identified on the student's IEP, 504, and/or LEP plan. Modifications can be effectively used in combination with accommodations in instructional and assessment situations when individualized to the student's strengths and needs. (3) Changes in what a student is expected to learn, such as changes in content, instructional level, and/or performance expectations. The intent of modifications is to allow for meaningful participation and enhanced learning.

Policies. Procedures for implementing laws.

Raw score. The number of items correct.

Scaled scores or derived scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Stakeholders. Persons holding a vested interest in the outcomes of the assessment program. These likely include parents, students, educators, and taxpayers.

Technically sound. Defensible assessments; they are reliable (consistent in their measurement and in the application of scoring procedures), valid for the purposes for which the results will be used, and are fair and unbiased.

Test security. The need to keep tests safeguarded so all students have equal exposure to the test materials and equal opportunities for success. If test security is violated, then some students can be placed at an unfair advantage or disadvantage. When this happens, the validity of tests is violated.

Chapter II References and Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological tests).* Washington, DC: American Educational Research Association.

Blank, R., Manise, J., & Brathwaite, B. C. (1999). *Indicators Report.* Washington, DC: Council of Chief State School Officers.

Creech, J. (1998). *Annual benchmarks report.* Atlanta, GA: Southern Regional Education Board.

Jaeger, R. M. & Tucker, C. G. (1997). *Analyzing, disaggregating, reporting, and interpreting students' achievement test results: A guide to practice for Title I and beyond.* Washington, DC: Council of Chief State School Officers.

National Assessment of Educational Progress. (n.d.) http://nces.ed.gov/nationsreportcard/ (retrieved 1-28-01).

Olson, L. (1999). Quality Counts '99. *Education Week*. Washington, DC: Editorial Projects in Education.

Texas Education Agency. (1998). *Professional development and appraisal system implementation manual for appraisers and teachers*. Austin, TX: Texas Education Agency.

Virginia Department of Education (1995). *Standards of quality and standards of accreditation*. Richmond, VA: Author.

Chapter III. Matching Tests to Purposes and Uses: Accountability Versus Instructional Planning[1]

- Purpose
- Design Considerations
- Summary/Conclusions
- Chapter Glossary
- Chapter References and Resources

Purpose

Chapter II focused on the need to be clear about the purpose for an assessment system. This chapter focuses on how to match the purpose to the components of the system, once the purpose has been clarified.

As learned in Chapter II, the bedrock of a sound, defensible assessment system is clarity about purpose. Purposes tend to fall into the following three categories (Almond, 1999).

1. *Systems **accountability***—using assessment information to hold programs, schools, and/or districts accountable for student achievement.
2. *Student accountability*—the application of consequences to individual students on the basis of their demonstrated achievement levels. Consequences might include such things as promotion or graduation.
3. *Instructional improvement*—the use of assessment information in determining needs for instructional improvement at the program, classroom, grade, and/or individual student level(s).

The purposes and uses of the assessment system are central in making decisions about the assessment instruments and tasks included in the system. Purpose and use will influence decisions ranging from content coverage to the degree and nature of technical documentation needed. Specifying the purposes of the assessment program as the first step in design and development will (1) increase the likelihood of addressing the intent of legislation and policy, and (2) lessen the likelihood of omitting important components of the system or of trying to meet multiple incompatible goals with a single component.

If purpose and use are not thoroughly and carefully considered, an assessment system intended to improve instruction through accountability, for example, may instead weaken instruction by inadvertently encouraging constriction of the range of instructional techniques. Or, a system designed with the intent of diagnosing student needs may not provide information appropriate for school-level accountability.

While legislation and school board regulations and policies often provide a broad framework for defining the purposes and uses of the assessment system, many of the decisions that will affect how the system is implemented and used are made by the system designers—state departments of education and local school districts.

Gribbons and Winter (1998) use a framework (see Table III-1) for the kinds of questions we should be addressing according to the purposes of our assessment systems.

[1] The information in this chapter draws heavily upon the work of Hansche, Stubits, and Winter (1998).

Table III-1. Framework for Considering the Purposes of Assessment Systems

	Student	School	District	State
Accountability	Has the student attained the standards?	Is the school (or, which schools are) making progress relative to all students' attaining high standards, including students who vary in terms of income, ethnicity, gender, language proficiency, and disabilities status?	Is the district making progress toward the goal of all students attaining high standards?	Is the state making progress toward all students' attaining high standards?
Program Improvement	What programs or services could the student benefit from?	What programs or services, including parent involvement, professional development, or extended day or year programs, need to be modified or added to enable the school's attainment of the goal?	What support does the district need to change or increase to help enable schools' attainment of standards?	What programs does the state need to add to meet the needs and enable attaining the goals?
Instruction	What areas has the student done well in, and in what areas does the student need more assistance?	What are the curricular and instructional problem areas?	In what specific areas does the district need to focus support for improving curriculum and instruction?	What areas of curriculum and instruction need additional attention at the state level?

Design Considerations

Designers should consider several interrelated questions in the initial stages of design. All of these questions point to the importance of validating an assessment for its intended purposes:

1. What are the ultimate goals of the assessment program?

Do they include systems accountability, student accountability, and/or instructional improvement? Purposes vary in their degree of specificity and might include improving student learning on common content standards; influencing instructional content and methods; motivating students and teachers; informing the public about school performance; and illuminating school-based information such as course-work grades. For example:

- Improving student learning on common content standards implies an agreed-upon set of learning expectations that are taught in all classrooms and assessed in a manner that reflects the knowledge and skills, including cognitive skills, contained in the standards. It also requires standards of proficiency, i.e., performance standards, that explicate the range of possible and acceptable evidence of student learning relative to the content standards. Since improving learning often necessitates changes and improvements in teaching, any assessment system that aligns with standards that suggest nontraditional forms of assessment, such as performance assessments, carries the obligation of providing adequate staff development. The process of developing agreed-upon standards requires expertise, the ability to work with various constituent groups, and time. Fortunately, we are at a time nationally when states can learn from each other about the process of standards development. See, for example, Hansche, Stubits, & Winter, 1998.

- Influencing instructional content and methods implies that the content standards must be specific enough that teachers, students, and parents clearly understand what is expected to be taught and the content and skills that will be tested. The content standards should be rigorous enough to raise the level of expectation for what is taught and learned, but they should also be attainable. The performance standards should provide adequate exemplars for the kinds of performances deemed proficient as well as below and above proficient. Sometimes changes in curriculum and instruc-

tion are motivated by accountability requirements. However, the quality of change and the ease with which it is implemented is highly dependent on support from central administration. Teachers need time to work together to efficiently and effectively improve curriculum and instruction. They may also need staff development support, especially if they are dealing with the alignment of curriculum and instruction to new content standards and assessments. We sometimes forget that classroom assessment is an important component of effective educational systems. When a purpose of assessment is to change instruction, teachers often need help in seeing and acting upon the relationships between state assessments and classroom practices. Providing teachers with meaningful staff development and planning time requires creative scheduling, collaboration, and, often, additional resources.

- Motivating students and teachers is facilitated by challenging and attainable learning expectations that they and their peers value. Accountability policies that involve the application of incentives and/or sanctions are on the rise, especially with regard to student graduation and school accreditation.

  One might consider the fairness of holding students accountable without holding schools accountable. On the other hand, some states have added student accountability components because students did not seem to take tests without personal consequences as seriously as they might.

  The application of consequences to students or schools carries numerous implications for fairness. For example, schools should have fair warning of the expectations of the accountability system, information about their status relative to the expectations, and assistance in meeting the criteria for effectiveness. Likewise, students must be given multiple opportunities to take high-stakes tests and given ample opportunities for instruction and remediation between test administrations. Obviously, there are costs associated with repeated test administrations and remediation programs.

- Informing the public about school performance may require an assessment that goes beyond student tests, especially if the schools' constituents believe that other indicators of school effectiveness exist in addition to student test scores. Several compendiums of the kinds of indicators used nationally to describe schools are available (Blank; Creech). Informing the public about school performance requires careful consideration to whether it is desirable for the information to be presented in ways that allow for comparisons between schools, for example. Results can be presented in a manner that facilitates appropriate interpretation. The information should be accurate, easily understood, clear about how *all* students are included, and protect the anonymity of individuals.

- Clarifying the meaning of school-based information, such as course-work grades, suggests using additional sources of information to better understand students' learning strengths and weaknesses. For example, if, on average, the students in a school have high course grades but low test scores, it may indicate that the instructional level is not as challenging as it might be or that the test is inappropriately difficult.

  In the case of a standards-based system, it might mean that classroom instruction and the assessments are not aligned in terms of performance expectations or standards. In the case of norm-referenced test scores, it may simply indicate that the test measures different learning than the classroom assessments. The informational value of this discrepancy depends on the purpose for administering the norm-referenced test. It would be inappropriate to use it as an indicator of individual student accountability; however, it might be appropriate to use aggregated results as one indicator of school effectiveness. At the school, program, and district levels it is useful to look at multiple indicators in planning for improvement.

2. What do you want the assessment results to tell you and at what level do you want this information?

Do you want the information to tell you about progress, performance relative to a standard, and/or performance relative to a norm group? Do you want this information at the student, program, school, district, and/or state level(s)? The system might be designed to yield information about student status relative to performance standards, student strengths and weaknesses in particular content areas, and school and district progress in educating students. For example:

- If the system is to yield information about student status relative to performance standards, there must be a system of content standards, assessments, and performance standards. These system components must be aligned in terms of breadth, depth, and cognitive demand. *All* students must be provided with opportunities to learn the material included in the standards and fair opportunities to demonstrate their learning.

- If information for measuring progress is desired, a decision must be made whether progress will be measured for individual students or for groups of students. If progress will be measured for groups, then a decision must be made whether the tracking will be on a **longitudinal** basis—the exact same students are followed over time—versus a **cross-sectional** basis, whereby different groups of students at the same grade level are compared from one year to the next (e.g., fifth-grade students in one school year are compared to the performance of different fifth-grade students in a subsequent year). Due to student transience and issues involved in calibrating tests administered at different grade levels, it is often unfeasible to track individual students over time.

  Most states compare groups of students from different years on the same test to gauge the extent to which schools or districts improve relative to student achievement. Doing so is more cost effective but is less useful for tracking individual student gains. This may be one reason why comparing individual student performance to a pre-determined standard of proficiency is an increasingly attractive option.

  If the intent is to compare individual or group performance to a norm group, then a standardized, norm-referenced test must be included in the assessment system. Standards-based tests can be normed; however, it is highly unlikely that the norms will be national in scope. This is because it is difficult to obtain a national norming sample of students to be tested on a set of standards that are unique to another state. If a state is willing to make national comparisons at the state level only, and not at the student, school, or district levels, it could participate in the National Assessment of Educational Progress (NAEP), which is based on a framework developed through a national consensus process. Most states consider national standards, such as those developed by the National Council of Teachers of Mathematics (NCTM), when developing their state standards.

3. What accommodations are necessary to allow all students access to the assessment?

The characteristics of the student population must be taken into account when specifying the purposes and uses of the system. Factors to consider might include the following:

- Coverage of the full range of student knowledge and skills in the content areas tested. This will help ensure that students at all levels of learning have an opportunity to demonstrate their knowledge and skills.

- Opportunities for students to demonstrate their knowledge and skills in a variety of ways. Providing such opportunities increases the probability of capturing the breadth of students knowledge and skills that may not be demonstrated if assessment is limited to single methodologies or situations.

- Appropriate assessment techniques for students with limited English proficiency (LEP) or for students with disabilities. Unless the assessments are accessible to these students, they cannot provide accurate measures of student learning. Sometimes appropriately assessing students requires testing **accommodations** or, in a small percentage of cases, **alternate assessments.**

- Inclusion of content and contexts appropriate for the cultural and ethnic diversity in the student population. Determining the extent to which assessments and assessment items are appropriate for use with diverse populations requires quality reviews, including reviews for cultural and ethnic bias, during the test development phase.

4. Who will use the assessment results and how?

Answering this question requires consideration of the stakes attached to the assessments and the type of information needed to report results. For example, results might be used by different interest groups in different ways:

- by teachers, to inform classroom instruction
- by teachers, students, and parents, to evaluate individual student progress
- by school boards and administrators to evaluate the effectiveness of schools, districts, and programs
- by state officials to evaluate schools and districts
- by policymakers to inform local, statewide, and national decision making
- by the media to inform or influence public opinion

When the stakes associated with test performance are high, such as a student's graduation or a school's accreditation, the need for accurate and reliable scores that are free from bias is critical. The higher the stakes, the more attention must be paid to these issues. Most important, the test results must be valid for their intended use. For example, using a nationally normed test in determining a student's graduation from high school would be invalid because such tests are not curriculum-based.

When assessments will be used by teachers for guiding day-to-day classroom instruction, then the psychometric rigor necessary for high-stakes, large-scale assessments is less important because teachers have other ongoing information to help them interpret and use the results. What is most important is that the assessments are useful in determining where students require more or different instruction.

5. How will assessment results be reported?

Quite often, considering how results might be reported helps clarify and increase understanding of the intended purposes of the assessment system. Reporting is the linkage between purpose and assessment. Reporting considerations also help further define the content of the assessments and the types of instruments and tasks that should be included. A variety of reports might contain assessment results, including:

- student report cards showing performance related to content standards
- student score profiles for school and district use
- school and district performance reports showing overall performance and performance disaggregated by demographic groups (e.g., poverty level, race, ethnicity, gender), including both standards-based as norm-referenced results
- individual student reports that show performance on a number of assessment instruments
- "traditional" score reports, including student reports on a particular assessment and classroom, school, district, and state summary reports.

Summary/Conclusions

The purposes of an assessment system determine the kinds of test items to include in the system (e.g., multiple-choice, constructed-response, performance); the point of reference for interpreting test scores (e.g., norms, standards, prior performance); and the levels at which test results are reported (e.g., student, class, program, grade level, content area tested, school, district, state).

In general, state assessment systems have one or more purposes, including (1) system accountability, (2) student accountability, and (3) instructional improvement. Any of these purposes requires attention to the ultimate goals of the system, the kinds and levels of desired information, the fair and appropriate inclusion of *all* students, the purposes for which the results will be used, and how results will be reported.

Chapter III Glossary

Accommodations. (1) Changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To be appropriate, assessment accommodations must be those also made during instruction and must not alter the construct intended to be measured or the meaning of the resulting scores. (2) Specific changes in testing conditions, procedures and/or formatting that do not alter the validity or reliability of a state standard. Policies and procedures must ensure that the accommodations do not compromise the security of the test and are consistent with the student's Individualized educational Plan (IEP), 504, and/or Limited English Proficient (LEP) plan. Accommodations can be made available for use in both instruction and statewide assessments. These may include accommodations for scheduling, setting, equipment, presentations, and/or responses. Allowable accommodations for states' assessments are generally identified in State Education Agency (SEA) documentation. (3) Alteration in *how* a test is presented to the test taker or in how a test taker is allowed to respond; includes a variety of alterations in presentation format, response format, setting in which the test is taken, scheduling or timing and/or specialized equipment required by the student. The alterations do not substantially change level, content, or performance criteria. The changes are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate what is known. (4) Change in *how* a student accesses information and/or demonstrates learning; does not substantially change the content, instructional level, or performance expectations; provides for equal opportunity to demonstrate knowledge and skills.

Accountability. The systematic use of assessment data and other information to assure those inside and outside the educational system that schools are moving in desired directions. Commonly included elements are goals, indicators of progress toward meeting those goals, analyses of data, reporting procedures, and consequences or sanctions. Accountability often includes the use of assessment results and other data to determine program effectiveness and to make decisions about resources, rewards, and consequences.

Alternate assessments. An approach used in gathering information on the performance and progress of students whose disabilities preclude them from valid and reliable participation in typical state assessments as used with the majority of students who attend school. Under the re-authorized Individuals with Disabilities Education Act (IDEA, 1997), alternate assessments are to be used to measure the performance of a relatively small population of students who are unable to participate in the regular assessment system, even with **accommodations** or **modifications.**

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Cross-sectional studies. Comparison of different groups of individuals over time, e.g., the results obtained by a group of fifth-grade students on a standardized mathematics test in one year compared to the results obtained by a different group of fifth-grade students on the same test in another year. This kind of analysis is commonly used to track the progress of a school, district, state, or nation over time.

Longitudinal studies. Comparison of the same individuals' results over time. In such studies, care must be taken that the measures used are also reliable over time. Groups may be studied longitudinally, provided that the individuals within the group remain the same, i.e., there are no "dropouts" and there are no new members.

Modifications. (1) Changes made in the content and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. (2) Changes in the administration of an assessment that may cause the construct being measured to differ from the construct as measured under standard administration conditions, or produce a score that means something different from scores yielded by the standard administration. Unlike *accommodations*, modifications may directly or indirectly compromise either the validity or reliability of the State standard. Modifications may compromise test security and therefore are not recommended for statewide assessments. Modifications are more appropriate for instruction and classroom tests and include a much wider range of supports and instructional scaffolding than do accommodations. Modifications can be identified on the student's IEP, 504 and or LEP plan. Modifications can be effectively used in combination with accommodations in instructional and assessment situations when individualized to the student's strengths and needs. (3) Changes in what a student is expected to learn, such as changes in content, instructional level, and/or performance expectations. The intent of modifications is to allow for meaningful participation and enhanced learning.

Performance-based assessments or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrate the range of performance for each performance level; and (4) cut scores that separate one level of performance from another.

Chapter III References and Resources

Almond, P. (1999). *A single assessment system for all students including those with special challenges—disabilities, limited English fluency, and poverty: What will it take?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gribbons, B., & Winter, P. C. (1999, January). *Using multiple measures of student achievement.* Unpublished proposal submitted to the U.S. Department of Education.

Hansche, L., Stubits, T., & Winter, P., et al. (1998, May). *Using existing assessments for measuring student achievement: Guidelines and state resources.* Washington, DC: Council of Chief State School Officers.

Chapter IV. What Should Be Measured? When? To What Extent?

Background

Issues of **alignment** are at the heart of this chapter. Alignment refers to the similarity or match between and among the **content standards, performance standards, curriculum,** instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned.

What should be measured by an assessment system depends upon the purpose(s) of the system as discussed in Chapter II. The "what" of measurement generally refers to what knowledge and skills should be assessed. In standards-based systems, this should be driven by the content standards.

When to measure implies that tests should be administered at certain times of the year and/or at particular grade levels. When to assess is also determined by purpose.

The extent of measurement concerns the **breadth** and **depth** to which content and skills are assessed. Properly written content standards can guide these measurement decisions. Performance standards strengthen the system by providing explicit statements about performance expectations relative to the content standards as well as exemplars for the entire range of possible performances, e.g., from failing to advanced.

Key Issues

1. **Content**—What subject matter areas should be assessed?
2. **Grade Levels**—At what grade levels should the specified subject matter areas be assessed?
3. **Coverage**—How do we assure that the content standards reflect the intended **domain** of knowledge and skills and that the standards are being covered in terms of breadth, depth, **process,** and **accessibility?**
4. **Timing**—When should large-scale assessments be administered?
5. **Alignment**—How well do the assessments connect to the standards?

1. Content—What subject matter areas should be assessed?

The content to be assessed as part of a large-scale assessment system should be driven by the purpose and goals of the system as discussed in Chapter II. The Council of Chief State School Officers conducts an annual survey showing which states assess in which content areas. Federal legislation (Title I of the 1994 Elementary and Secondary Education Act [ESEA]) requires that students be assessed in at least the core areas of reading/language arts and mathematics (Section 1111(b)(3); § §200.1(b)(2) and 200.4). Many states additionally administer statewide assessments in science and social studies.

In addition to statewide assessments, additional assessments might be administered locally for accountability purposes and to provide information for program planning and refinement. In some instances, local school districts administer tests beyond those required by the state because it is important to their communities to have information in these other

chosen content areas. Sometimes the state provides funding support for additional testing; in other cases the costs are borne entirely by the local district.

When tests beyond those required by the state are administered in a school district, care must also be taken to ensure that such tests are **technically sound.** Whether tests are supplied by a commercial vendor or developed locally, the technical soundness of the tests must be demonstrated and documented. In all cases, test users are responsible for requiring test developers and vendors to provide information regarding the technical quality of their tests for the intended uses.

If consequences are to be applied to the test results, technical soundness becomes an even greater issue. Building technical soundness into tests and scoring procedures can have cost implications. Most importantly, whatever tests are used, if they are intended to reflect certain standards or expectations for what students should know and be able to do as a function of schooling, then the tests must be aligned with the content and performance standards.

2. Grade levels—At what grade levels should the specified subject matter areas be assessed?

Federal law is fairly clear here. For Title I funding purposes, at least one grade level must be assessed for each of three grade spans: 3-5, 6-9, 10-12. The annual CCSSO survey of state assessment programs, noted previously, indicates which states administer assessments at which grade levels.

As with issues of content, the underlying purpose and goals of the assessment system should be the key factor in determining which grade levels are assessed. For example, if passing the state standards-based test is one of several requirements for promotion to certain grade levels, then assessments must be administered in the grades to which the requirement applies. Tests may also be administered at other grades for purposes of providing useful planning and diagnostic information to administrators and teachers.

If large-scale assessments are being used as an indicator of school effectiveness only, then administering tests at the culminating grade level for primary, elementary, middle, and high schools may be adequate. However, motivational factors should be considered when tests are used in this fashion. For example, graduating seniors may see little need to take or perform well on tests of little personal consequence, thereby yielding skewed or invalid indicators of actual achievement.

In most instances, cost is a factor in deciding which grade levels to test. Some states choose to test only fundamental skills such as reading comprehension and mathematics concepts at the primary levels, writing only at certain levels such as the end of elementary school and again in high school, and other subjects such as science and history at still other grade levels.

Another consideration is the burden of testing on students. It is ideal to strive for the minimum of testing that can provide the information needed to meet the purposes and goals of the system. Sometimes states strive to reduce the burden of testing by staggering what is tested across grades. For example, reading may be first tested in third grade while writing is not tested until fourth grade. In such cases it is important to ensure that writing instruction does not suffer during the year that reading is tested, and vice versa.

Ultimately, there are as many combinations as there are purposes and the resources to support them. The scenario at the end of Chapter II provides an illustration of key issues in choosing which grade levels to test. In the end, the hypothetical state illustrated in the scenario (whose purpose for assessment is to enable each student to develop the skills necessary for success in school and in life and to hold students and schools accountable for demonstrating learning results) made the following decisions:

- Test in grades 3, 5, 8, and 11 and at the end of selected high school courses.
- In grade 3, fundamental skills of reading and mathematics are emphasized.

- In grades 5, 8, and 11, writing tests that require students to respond to a **writing prompt** are administered.
- In grades 5, 8, and 11 students are tested on technology skills.
- Students must pass end-of-course tests in English, mathematics, science, and social sciences in order to graduate.
- Every year schools must increase the proportion of students meeting the proficient level or higher and decrease the proportion scoring "below proficient" in order to be accredited.

Even though this state may have wanted to do more testing, cost considerations and purpose guided the state board's decisions. A number of school districts do additional testing at their own expense. Usually this is restricted to the administration of commercially available tests at the off-grades due to the expense involved in developing new standards-based tests. A caution, of course, is that off-the-shelf tests seldom reflect the same standards as the standards-based tests. School districts, like the state, must be clear about what they want their assessment systems to accomplish when making decisions about supplemental testing.

3. Coverage—How do we ensure that the assessments reflect the intended domain of knowledge and skills and that they adequately reflect the content standards in terms of breadth, depth, process, and accessibility?

Key to this issue is determining the domain of knowledge and skills—from the entire universe of knowledge and skills that should be taught, learned, and assessed. These domains are the cornerstones of content objectives. They represent the breadth of what is to be taught, learned, and assessed. When formulating content standards, consideration must be given to the abilities and potentials of *all* students, including students with disabilities, students of limited English proficiency, and students who are academically gifted. A number of good sources exist relative to designing assessment systems that are accessible to special needs students (e.g., Kopriva, 2000; Ysseldyke, Olsen, & Thurlow, 1997). The subject is also treated in Chapter XIII of this document.

Adequately defining the knowledge and skill domain to be encompassed by the content standards requires input from the community as well as from knowledgeable experts, including content specialists, specialists in child development and learning, and especially teachers experienced with the wide variety of students enrolled in a state's schools.

Initial drafts of **performance standards** are critical in defining the domain. The performance level descriptors and examples illustrate the range of performance necessary for scoring at any particular level. Examples of performance standards, their development, and application may be found in Hansche (1998).

Once the domain is defined, **test blueprints** or **specifications** must be developed. The test blueprints ensure that the breadth of knowledge and skills represented by the content standards is reflected in the tests. While more than one form of a test may be developed, all forms must reflect the standards-based blueprint. Prior to test development, the test blueprints should be reviewed by persons other than the test developers to assure that they match the scope of the entire standards domain.

Once the test blueprints are approved and test items are developed, the items and tests should be reviewed prior to **pilot testing.** The purpose of this activity is to review the items for accuracy and **bias,** and to ensure that the test, as a whole, matches the blueprint. Once the tests are piloted, the results can be used to validate freedom from bias, and adjustments can be made to items that are inappropriate, too easy, too difficult, inaccessible to certain students, or otherwise flawed.

In the case of **constructed-response items** or tests, scoring **rubrics** are needed to ensure that the content domain being assessed is tested at the desired depth of cognitive

demand. The scoring rubrics specify what students must do in order to receive varying levels of credit for their responses.

4. Timing—When should large-scale assessments be administered?

Purpose and instructional considerations are the prime drivers when it comes to timing the administration of large-scale assessments. Most states and large school districts try to avoid administering such exams during times when absenteeism might be high (e.g., flu season, inclement weather) or attention might be distracted (e.g., near major holidays).

Most nationally norm-referenced tests have both fall and spring norms. Hence, norm-referenced tests can generally be administered during particular windows of time in the fall or spring, providing that the appropriate norms are used for interpreting test results.

It is generally desirable to administer standards-based assessments as close to the end of a course of study as possible so that students have the benefit of instruction covering as much of the standards as possible. This presents challenges if the assessments involve writing or performance tasks that cannot be quickly scored. It also presents challenges for creating class rosters of test results so that the receiving teacher gets the results for his/her incoming students the following term. Test publishers can generally facilitate some of these logistics for a price.

If a purpose of the large-scale assessment system is to track individual students over time so that statements about individual student progress can be made, it will be necessary to pre and post test students or use a series of tests that have been linked to the same scale. It can be extremely difficult to accurately track this kind of progress from one grade to the next because the tests designed for different grade levels necessarily cover different content and skills. However, pre and post testing relative to the standards for a particular grade or grade-span can be useful if the tests are adequate for this kind of use. States and school districts usually track progress based on **cohorts** of students. For example they compare results for fifth-grade students in mathematics to the results for ensuing groups of fifth-grade students on the same test or a **parallel form** of the test.

An excellent treatise on these issues is provided by Webb (1997). Popham (1998) also speaks to the appropriate and inappropriate uses of tests in judging school effectiveness.

5. Alignment—How well do the assessments connect to the standards?

Issues of alignment permeate each of the issues discussed in this chapter. In order for assessments to be used as a measure of educational quality, they must match what we profess to be important for students to learn. As discussed previously, the assessments must match the underlying content standards in terms of breadth of coverage, and they must match the performance standards in terms of depth of coverage.

However, no amount of standards-assessment alignment will ensure adequate student learning unless the curriculum is also aligned with the standards and assessments. Instructional strategies must include classroom assessments that are aligned with the standards-aligned curriculum. The alignment relationship for a hypothetical set of content standards is illustrated in Table IV-1. Note that many content standards can be operationalized by means of performance standards. But not all performance standards can be assessed by methods typically used for large-scale assessments.

Also note that not all of the desired learning is, or can necessarily be, assessed by large-scale assessments. Standards that are not amenable to assessment via large-scale methods should be tracked at the local level. This is especially important in light of Title I legislation, which calls for reporting at the content area level. The locally obtained information will have to be combined with the large-scale assessment results for a school or district in such a way that progress, in terms of the entire span of the content standards, can be reported.

Clearly, it is not sufficient to "map" from the standards to the assessments only. It is also necessary to assure that the assessments cover the entire range of the standards (LaMarca, Redfield, & Winter, 2000).

Table IV-1. Alignment Relationship: Content Standards, Performance Standards, Large-Scale Assessment Methodology

Content Standards	Performance Standards	Can Be Assessed by Large-Scale Methods	Requires Local Methods
Standard A1	A1.1 A1.2 A1.3	A1.2	A1.1 A1.3
Standard A2	A2.1 A2.2 Etc.	A2.1 A2.2 Etc.	 A2.3
Standard A3	Etc.	Etc.	Etc.
Standard B1	Etc.	Etc.	Etc.
Standard B2	Etc.	B2.1 B2.4 B2.5	B2.2 B2.3 Etc.
Etc.	Etc.	Etc.	Etc.

Conclusions

The content and skills to be measured, the grade levels at which they should be measured, the specificity of the content standards, and the rigor of the performance standards should all be driven by the purpose and goals of the education system. The education system goals should give rise to content and performance standards, which then drive curriculum, instruction, and assessment.

The scenario presented at the end of Chapter II illustrates the kinds of decisions made relative to each of these issues, including compromises and cost implications. If the assessments will be used for accountability purposes, then it is critical that the standards, curriculum, instruction, and assessments be aligned. Anything less would be unethical and indefensible.

Chapter IV Glossary

Accessibility. The extent to which the content, format, and response mode options of an assessment make it possible for *all* students, including students who have disabilities or limited English proficiency, to participate in an assessment.

Alignment. The similarity or match between and among the content standards, performance standards, curriculum, instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. This standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Bias. In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

Breadth. The comprehensiveness of the content and skills embodied in the standards, curriculum, and assessments.

Cohorts. In educational research, generally, groups of students who cannot necessarily be compared to themselves over time. This is usually due to attrition such as moving away or dropping out of school. Examples of cohort studies include comparing groups of different students at the same grade level over time or comparing scores from the same group over time, even though some group members may change.

Constructed-response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Construct. Underlying theoretical concept or characteristic a test is designed to measure.

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Curriculum. What is taught.

Depth. The taxonomic level of cognitive processing required for success relative to the performance standards, e.g., recognition, recall, problem solving, analysis, synthesis, evaluation.

Domain. The portion of all knowledge and skill in a subject matter area that is selected for the content standards once consensus is reached that it represents what is important for teachers to teach and students to learn.

Field test. A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting and sometimes test form equating. A field test is generally more extensive than a **pilot test.**

Parallel tests. Also called alternate test forms; two or more versions of a test that are considered interchangeable in that they measure the same **constructs,** are intended for the same purposes, are administered using the same directions, and yield comparable scores.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrates the range of performance for each performance level; and (4) cut scores that separate one level of performance from another.

Pilot test. A test administered to a representative sample of test takers solely for the purpose of determining the properties of the test. See **field test.**

Rubric. Scoring guide for constructed-response questions or performance tasks. Scoring rubrics contain a description of the requirements for varying degrees of success in responding to the question or performing the task.

Teaching the test. Teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

Teaching to a test. Teaching the broad-based knowledge and skills represented by a test's underlying content standards. Compared to **teaching the test,** it is not cheating.

Technically sound. Defensible assessments; they are reliable (consistent in their measurement and in the application of scoring procedures), valid for the purposes for which the results will be used, and are fair and unbiased.

Test blueprints. Written documents, often in chart form, that detail the number of questions to be included on a test, the item formats, and the content and skills that each set of items will assess. In the case of standards-based tests, it is important for the test blueprints to consider the performance standards as well as the content standards so that items cover the intended depth as well as breadth of the standards. In addition to guiding test development, test blueprints can be useful in preparing to take an examination.

Test specifications. Sometimes used interchangeably with **test blueprints.** Test specifica-

tions provide a framework that specifies the proportion of items that assess each content and process/skill area; as well as the format of items, responses, and scoring protocols and procedures. These frameworks additionally specify the desired psychometric properties of the test and test items, such as the distribution of item difficulty and discrimination indices.

Writing prompts. Phrases or sentences designed to elicit written responses. In the primary grades they may take the form of story-starters. In later grades they may ask students to write an essay on a particular topic, often specifying a particular mode (e.g., persuasive, descriptive).

Chapter IV References and Resources

Council of Chief State School Officers. (1999). *Trends in state student assessment programs: Fall 1997.* Washington, DC: Author.

Hansche, L. N., Winter, P., & Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington DC.

Kopriva, R. (2000). *Ensuring accuracy in testing for LEP students: A practical guide for assessment development.* Washington, DC: Council of Chief State School Officers.

LaMarca, P., Redfield, D. L., & Winter, P. (2000). *State standards and state assessment systems: A guide to alignment.* Washington, DC: Council of Chief State School Officers.

Popham, J. (1998). *Inappropriate uses of tests to judge school effectiveness.* Richmond, VA: Virginia Association of Test Directors.

Webb, N. (1997, January). *Determining alignment of expectations and assessments in mathematics and science education.* National Institute for Science Education, *1*(2).

Ysseldyke, J., Olsen, K., & Thurlow, M. (1997). *Issues and considerations in alternate assessments.* (Synthesis Report No. 27). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Chapter V: How Many Tests Make an Assessment System?

Introduction

A **test,** for purposes of this document, is one measure of one content **domain** at one grade level.

Whether or not they are **performance-based,** tests may be **norm-referenced** or **criterion-referenced.** An **assessment** is a set of tests or other measures designed to meet one or more purposes, e.g., a set of tests to measure achievement in grade four for all students in mathematics, reading, and writing. An **assessment system** comprises all of the assessments to which a student is exposed (K-12), in addition to those components of the educational system that establish the context in which the assessments occur. These components include **content standards, curriculum** and curriculum guides, instructional processes and materials, **test blueprints** or specifications, and **performance standards.**

Issues and Trade-Offs

At least three key issues influence the number of tests required for an adequate and valid assessment system: (1) the purpose for testing, the context within which testing will occur, and how the results will be used; (2) **alignment** between the standards and the assessments; and (3) to whom or what student achievement will be compared.

1. What is the purpose for testing and the context within which the results will be used?

The reasons for having an assessment system should always influence decisions about the number and kinds of tests and other measures to be used. For example, if an assessment will be used in the determination of school accreditation, measures of achievement in each core content area, as well as measures of attendance, retention, and discipline referrals, might be included. Or, if assessments will be used to determine whether students graduate from high school, providing repeated opportunities to take tests in the required content areas is warranted. It is also important that students have opportunities to use multiple response formats (e.g., multiple-choice, essay, performance) so all students have adequate opportunities to demonstrate their knowledge and skills, regardless of their affinity (or lack thereof) for particular formats.

So . . . how many tests are needed? The system must include a sufficient number and variety of tests to fulfill the purpose(s) for testing.

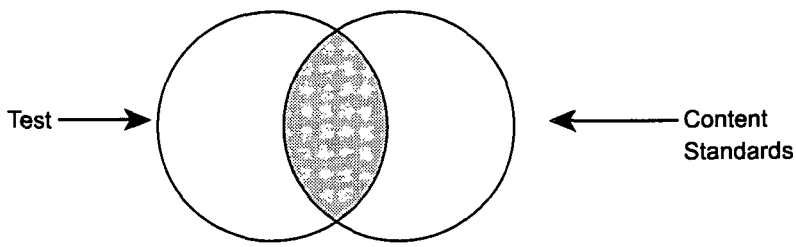
2. How many assessments are needed for an aligned system?

The answer to this question is not independent from the answer to question #1. An aligned assessment system is one that matches the depth and breadth of the tests within the system with the underlying **content standards** as well as standards of performance. As implied by Figure IV-1 in Chapter IV, a single test or other measure is inadequate for assessing the knowledge and skills represented by the content standards and the full range of proficiencies exemplified in the **performance standards.**

For example, it may be that a multiple-choice instrument can be used to efficiently assess students' editing abilities, but an examination that requires students to provide written responses is needed to assess the ability to compose. Together these instruments can more accurately assess students' written language achievement than either of the measures alone. And, if the intent of the standards is that students be proficient in English, then, at the least, both measures are required.

If the assessments in a system do not adequately represent the depth and breadth of the standards, then the system is not aligned. Assessments and content standards can be misaligned in several ways:

- The assessment or test may include only some items or tasks that can be directly aligned with the standards. The rest of the items or tasks cannot be logically aligned with the standards. This situation is depicted in Figure V-1.



Note: Perfect alignment would result in perfectly overlapping circles; the shaded portion represents the degree of alignment.

Figure V-1. Misalignment: Areas of Non-Overlap Between Test and Content Standards

In this situation, all of the instruction could be properly targeted to the content standards, yet students would not have had adequate opportunity to learn the knowledge and skills assessed by most of the test.

Such a misalignment could result in instruction being shifted away from the content standards to the knowledge and skills covered by the test, as teachers become aware of the mismatch between the test and the content standards.

- The assessment or test may include items that fully align with the content standards, but, in combination only cover a few of the parts covered by the content standards. This type of misalignment is depicted by Figure V-2.

In this situation all of the instruction could be properly targeted to the content standards, and the student would have had adequate opportunity to learn the knowledge and skills

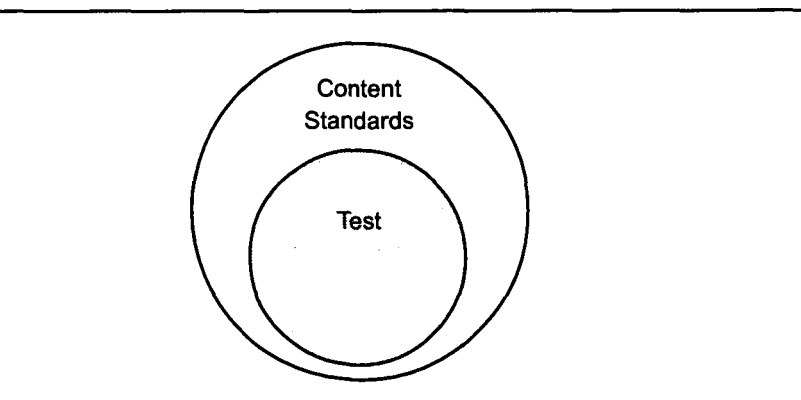


Figure V-2. Unbalanced Alignment: Test Covers Limited Portion of Content Standards

covered by the test. However, the test results would not accurately reflect the students' mastery of most of the domain covered in the content standards.

In the long run, this second situation would most likely result in a marked restriction of the breadth and/or depth of instruction, as teachers became aware of how narrowly focused the test was and targeted their efforts toward just those knowledge and skills they knew were covered on the test.

- Figure V-3 illustrates the situation wherein the content standards are much narrower than the knowledge and skills required to answer the test items. In this situation, even though the instruction is properly targeted toward the content standards, the expectations of the test may go beyond the demands of the instruction. For example, the test could require application, whereas the content standards do not mention application.

An assessment system designed to align with all of a state's content standards would most likely need to include several different types of assessments. For example, one component

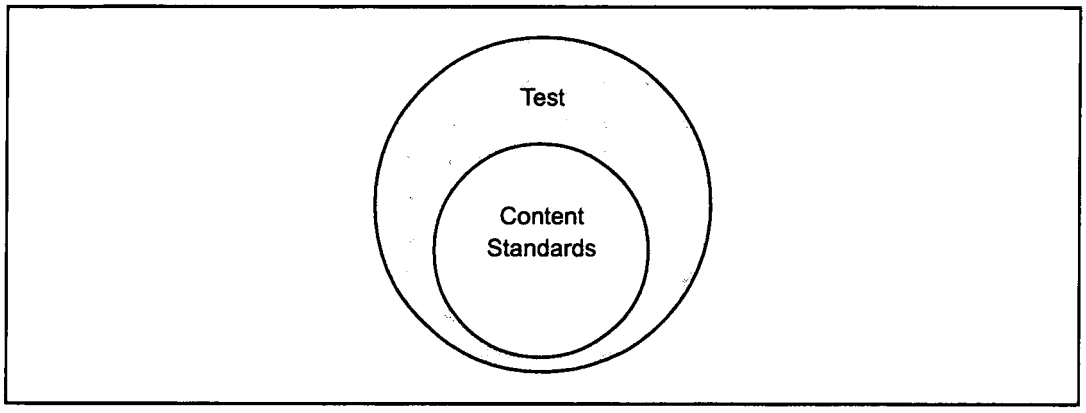


Figure V-3. Misalignment: Test Includes Material Not in the Content Standards

might be a mixture of **selected-response** and short **constructed-response** items. Part of this could include a norm-referenced test, depending on the purposes of the assessment. A second component might consist of extended constructed-response items such as essays.

A third component might be some sort of extended task, such as a project or **portfolio,** that would represent the students' abilities to draw together a number of examples of their work to demonstrate their mastery of a set of skills. Kentucky and Vermont provide examples of this type of approach on a large-scale basis. Gordon (1999) offers a discussion of scoring issues related to portfolio assessment.

Clearly, to be fair and defensible, there must be alignment among the standards, curriculum, and assessments. Otherwise, students may have inadequate opportunities to learn the content and skills that will be tested. In most instances, adequate alignment between standards and assessments requires several approaches or assessment measures that, together, comprise an adequate system. In general, as the stakes associated with test results increase, the need for attention to appropriate opportunities to learn the comprehensive scope of the standards-based material to the specified depth also increases.

So . . . how many tests are needed? The system must contain enough tests for alignment while minimizing the burden of testing on students.

3. To whom or what will student achievement be compared?

Here, we need to differentiate between individual student results and results that are **aggregated** for groups of students, such as students at a particular grade level or in a particular school, school district, or state.

Typically, the test scores of individual students are compared to a **norm** group and/or to a standard. If they are compared to a norm group, the scores will be expressed as **standard scores** such as percentile ranks or grade-equivalent scores.

If they are compared to a standard, the score is usually interpreted in terms of how the student's test performance compares to the criteria for proficiency. For example, a standard score of 400 may represent proficient performance while a standard score of 500 may represent advanced performance. However, neither 400 nor 500 reveal how well the student performed compared to other students.

Many states administer both kinds of tests, i.e., **norm-referenced** tests, which allow for comparisons to a national norm group, and **criterion-referenced** or **standards-based** tests, which allow for comparisons to criteria or a standard of proficiency.

When comparisons are made on the basis of group or aggregated results, they are most often compared to a norm group. Sometimes, however, comparisons are made on the basis of percentages of students meeting the "standard" in one group compared to another. For example: 70% of the students in School X obtained scores at the proficient level while 68% of the students in School Y obtained such scores.

As with individual scores, group averages may be compared to the norm group which took the same nationally normed, **standardized** test. However, there are some national and international tests, such as the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS) that do not allow for individual student comparisons. That is because these tests are administered on a matrix sampling basis where not all students are tested and those who are tested may take different items. Because the items are administered to large numbers of students, nationally or internationally, it is possible to make comparisons at the state-nation and national-international levels on the NAEP and TIMSS, respectively.

So . . . how many tests are needed? The number depends on the variety of measures needed to provide the kinds of interpretative comparisons desired. While criterion-referenced tests can be normed, it is seldom productive to do so because they are usually based upon a particular set of learning objectives or content standards that are not national in scope. Therefore, if it were desirable to compare individual students to a national norm group, a nationally normed test would need to be added to the mix. Many states use this kind of mixed testing model.

However, if a state simply wants to compare the state as a whole to national performance, an assessment such as the NAEP may suffice, if the state standards match those assessed by the NAEP. The NAEP is administered to samples of students in participating states at no cost to the state.

Summary/Conclusions

The number of tests included in an assessment system must be driven by purpose. Is the purpose to make norm-based comparisons? Comparisons to a standard? Individual comparisons? Group comparisons? In light of the responses to these questions, the balance between alignment and burden of testing on students must be weighed.

- The system must include a sufficient number and variety of tests to fulfill the purpose(s) for testing.
- The system must contain enough tests for alignment while minimizing the burden of testing on students.
- The number of tests needed depends on the variety of measures needed to provide the kinds of interpretative comparisons desired.

Chapter V Glossary

Aggregated scores. The total or combined performance for all individuals or groups on one test or subtest. For example, a state average usually represents the aggregation of scores for all students/groups of students who took the test.

Alignment. The similarity or match between and among the content standards, performance standards, curriculum, instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. This standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Assessment. Any systematic method of obtaining evidence from tests and other sources that is used to draw inferences about characteristics of people, objects, or programs for a specific purpose.

Assessment system. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. A standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Constructed response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Criterion-referenced. The reference point for interpreting test results using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to a norm or the performance of others.

Curriculum. What is taught.

Domain. The portion of all knowledge and skill in a subject matter area that is selected for the content standards once consensus is reached that it represents what is important for teachers to teach and students to learn.

Norm. Typical or average performance. The norm does not necessarily represent the most desirable performance.

Norm-referenced. Test interpretations whose scores are basesd on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Performance-based or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrates the range of

performance for each performance level; and (4) cut scores that separate one level of performance from another.

Portfolios/portfolio assessment. (1) Systematic collections of education or work products that are typically collected over time. (2) A collection of student-generated or student-focused products that provide the basis for judging student accomplishment. In school settings, portfolios may contain extended projects, drafts of student work, teacher comments and evaluations, assessment results, and self-evaluations. The products typically depict the range of skills the student has or reveal the improvement in a student's skill level over time. Salvia & Ysseldyke (1995) list six elements that typically are said to characterize portfolio assessment: (1) They target valued outcomes for assessment (generally those that require higher levels of understanding such as analysis, synthesis, and evaluation; those that require applying specific processes or strategies to reach answers; and those that are complex and challenging). (2) They use tasks that mirror work in the real world, i.e., that are *authentic*. (3) They encourage cooperation among learners and between teacher and student. (4) They use multiple dimensions to evaluate student work. (5) They encourage student reflections. (6) They integrate assessment and instruction.

Reference population. The population of test takers represented by a test's norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or other characteristics at the time of testing.

Selected response. Test tems that require students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Standard scores. A type of derived *score* such that the distribution of these *scores* for a specified population has convenient, known values for the mean and standard deviation.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Standards-based tests. A kind of criterion-referenced test. They consist of items that reflect a pre-established set of **content standards.** Results are then interpreted against a set of criteria or **performance standards.**

Systems of assessment. Consist of complementary components that, together, provide an accurate profile of student achievement.

Test. In contrast to an **assessment,** a test includes a number of measures that help create a more complete picture or profile of performance, is usually a single instrument or procedure such as a quiz, standardized measure, questionnaire, survey, observation, checklist, and the like. Thus, tests are typical components of aligned systems of assessment.

Test blueprints. Written documents, often in chart form, that detail the number of questions to be included on a test, the item formats, and the content and skills that each set of items will assess. In the case of standards-based tests, it is important for the test blueprints to consider the performance standards as well as the content standards so that items cover the intended depth as well as breadth of the standards. In addition to guiding test development, test blueprints can be useful in preparing to take an examination.

Chapter V References and Resources

Almond, P. (1999). *A single assessment system for all students including those with special challenges—disabilities, limited English fluency, and poverty: What will it take?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gordon, B. (1999). Score issues from a practitioner's perspective. In *Issues in scoring essays: Research and practice.* Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Chapter VI: Sampling

Background

The use of sampling procedures can provide cost-efficient information relative to the achievement of extensive bodies of information while simultaneously reducing the burden of testing on individual students. Classic, general references on sampling include Cochran (1953), Ebel (1980), Millman and Greene (1989), and Petersen, Kolen, and Hoover (1989).

A **sample** is a subset of the population of interest. The subset must be sufficiently large to represent that population. **Sampling** is the selection of the elements of a sample. When thinking about sampling, at least three methods can be considered:

- selecting students from a larger group of students to participate in testing
- selecting items from a larger pool of test items to be included on the test
- selecting a combination of students and items such that all students are tested, but they are tested with different items.

Issues and Trade-Offs

- Why use sampling, and what kind of sampling should be used?
- How does sampling affect test quality?

1. Why use sampling and what kind of sampling should be used?

When large numbers of students need to be tested relative to a large or complex **domain** of subject matter, using adequate sampling procedures can provide information relative to the achievement of the entire domain, reduce the burden of testing on individual students, and reduce costs.

To illustrate, consider that **standardized** achievement tests—whether they are **norm-referenced** or **criterion-referenced** and whether their format is **selected-response** or **constructed-response**—measure samples of what has been learned. When a student takes a vocabulary test, for example, the test does not measure all possible vocabulary, or the entire vocabulary that the student knows; it measures a subset of the universe of vocabulary. In educational settings, the ideal is for the subset of vocabulary that is measured to adequately represent vocabulary learning. For this reason, it is important for learning objectives, content standards, and test specifications to be carefully crafted on the basis of consensus among content specialists, learning and development experts, experienced education practitioners, and other vested stakeholders such as parents. Establishing such consensus requires time, money, patience, and a spirit of collaboration.

Further consider that even the set of knowledge and skills selected to represent the larger content domain is quite extensive and complex and that thousands of assessment items could be developed. Without an inordinate amount of testing time, no single student could complete all of these items. Hence, it is desirable to select items from the entire item pool.

The items might be selected at **random** or they might be selected to represent specific subtopics. The sampling method is a primary ingredient to be considered when generalizing results from a sample of items to the total item domain.

To monitor program effectiveness, we may need a measure of how well students enrolled in a particular program are doing overall, but we have no need for individual student scores. In such cases, students might be sampled, randomly or otherwise (e.g., in clusters according to selected variables such as socioeconomic status), to participate in testing. Since these students would be considered as representative of students enrolled in the program, test results could be used to draw inferences about the performance of the entire group of students enrolled in the program provided enough students participate in testing and that the technical properties of the test are robust enough to support such generalizations. As with the selection of items, the technique used in the sampling of students greatly influences the generalizability of results from the sample to the larger population.

To make efficient use of student test time, some states use a test design that is a combination of approaches. For example, schools are held accountable on the basis of how well students perform overall compared to the content and performance standards. All students are tested, but not all students take the same items. This allows for adequate coverage of the content domains the state wishes to assess without overburdening any one student with testing. This practice is called **matrix sampling.** Maryland, for example, uses matrix sampling.

The kind of sampling used should be determined by the purpose underlying the assessment program. Following are some considerations related to three aspects of sampling: **item sampling, student sampling,** and **matrix sampling.**

- Item sampling is useful when the pool of items representing the desired learning outcomes is large or would require an inordinate amount of testing time.

- Student sampling is less commonly used than item sampling. In part, this may be due to the political desirability of determining individual test scores. Parents often want to know results for their own children.

  Student sampling is desirable, however, when individual scores for each student are not required. It is particularly useful when the goal is to reduce the burden of testing on individual students while obtaining an adequate measure of how well the students in a program or system are performing as a whole.

- Matrix sampling is sometimes controversial because it is not well understood. Nonetheless, it can be quite useful when it is desirable to spread the burden of testing and when it is not necessary for every student to be tested on every item representing the same knowledge and skills within a content domain. Matrix sampling can provide broader coverage and, therefore, more detailed reports regarding the strengths and weaknesses in school curricula. Matrix sampling can also provide for the inclusion of performance tasks that are time consuming to administer due to their complexity. The National Assessment of Educational Progress is an example of an assessment based on matrix sampling.

  While it is commonly believed that matrix sampling procedures cannot provide individual scores, an assessment system can be designed to do so; however, individual student score interpretations may need to remain at a broad level. For example, it is possible to design a system that does not test every student on every mathematics item and that yields information about whether individual students are proficient in mathematics. At the school level, where information for students who took different items can be aggregated, it may be possible to say how many students are proficient in the various subdomains of mathematics. Kentucky is a state with such a system.

  Ultimately, when matrix sampling is used, it is important to be sure that the item pool is large enough to support the inferences that will be made on the basis of test results. Too few data points will result in too much error for valid interpretations.

2. How does sampling affect test quality?

Sampling can enhance or detract from the **validity** of the inferences that are made from test results.

- If the content standards or learning objectives upon which a test is based are narrow in scope, the generalizations that can be made about what students have or have not learned must be limited to that scope. For example, if learning objectives about writing are focused on the conventions of writing such as spelling, grammar, and punctuation, and the test of those learning objectives consists of editing tasks, it would be invalid to use the test results to draw conclusions about how well students can write.

- If the items sampled from the item pool inadequately represent the item pool in terms of the breadth or depth of the knowledge and skills they cover, the inferences drawn from the test results will be inaccurate. It is also important to sample enough items that, together, they provide a **reliable** measure. In general, the more items on a test, the more consistently it measures the intended content domain.

- The number of students who participate in testing must be large enough that valid inferences can be drawn about the larger group they represent. This is especially true if the data will be examined for different groups within the sample, such as differences in performance by grade level, gender, ethnicity, race, primary language, or socioeconomic status.

  In such cases, **stratified** sampling can be useful. Stratification can reduce errors of interpretation due to sampling if the persons selected are homogeneous (alike) within and heterogeneous (different) between strata.

Summary/Conclusions

Considerations in sampling include how much of the content domain needs to be tested, whether the testing process must provide valid scores for individual students, and the groups of students for which test results will be reported and used. Sampling plays an important role in the validity of test results.

Chapter VI Glossary

Constructed-response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Criterion-referenced. The reference point for interpreting test resulting using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to a norm or the performance of others.

Domain. The portion of all knowledge and skill in a subject matter area that is selected for the content standards once consensus is reached that it represents what is important for teachers to teach and students to learn.

Item samples. Subsets of a larger array of test items. Item samples must be sufficiently large to represent the full array of items.

Matrix sampling. A measurement technique whereby a large set of test items is organized

into a number of relatively short item sets. Each subset is then administered to a subsample of test takers, thereby avoiding the need to administer all items to all examinees, e.g., for program evaluation purposes.

Norm-referenced. Test interpretations whose scores are based on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Performance-based or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Performance tasks. A type of test item that is product- or behavior-based. They are designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Random sampling. The selection of a **sample** according to a random process, with the selection of each entity in no way dependent on the selection of other entities.

Reliable. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and, hence, are dependable and repeatable; the degree to which scores are free of *errors of measurement.*

Sample. A sample is a selection of a specified number of entities, called sampling units (test takers, items, etc.), from a larger specified set of possible entities, called the population.

Sampling. The selection of a **sample.**

Selected-response. Test items that require students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Stratified samples. Sets of samples, each of a specified size, from several different sets.

Student samples. Subsets of a larger population of students. Student samples must be adequate representations of the population they are meant to represent.

Validity. (1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Chapter VI References and Resources

Cochran, W. G. (1953). *Sampling techniques.* New York: John Wiley & Sons, Inc.

Ebel, R. L. (1980). *Practical problems in educational measurement.* Lexington, MA: D.C. Heath.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.). *Educational measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

Chapter VII: Norm-Referenced Versus Criterion-Referenced Test Results

Background

A dilemma faced by policymakers and designers of assessment programs is whether the state assessment program should report results as **norm-referenced** or **criterion-referenced** or both. A common misconception is that the terms "norm-referenced" and "criterion-referenced" refer to types of test items such as multiple-choice or constructed-response items. In fact, they do not. The terms refer to the manner in which test results may be interpreted.

The results of norm-referenced tests are interpreted by comparing them to the performance of a **norm group,** i.e., the group used to establish "typical" performance; whereas, the results of criterion-referenced tests are interpreted in terms of a pre-set criterion for success or proficiency. The criterion is usually a score that represents a certain level of achievement on specified subject matter. An example would be 30 items out of 40 on a test covering mathematics computation at the fifth-grade level.

Standards-based tests are a kind of criterion-referenced test. The criterion, however, is a performance standard. Like criterion-referenced tests, standards-based tests are based on **content standards** that specify the learning objectives or expectations for students. However, the results of standards-based tests are compared to a system of **performance standards** that includes (1) performance levels such as "proficient" or "advanced," (2) descriptions of the performance levels that indicate the kinds of performance assigned to each level, (3) the **cut score(s)** used to separate levels of performance, and (4) examples of student performances that demonstrate the acceptable range of performance within each level (Hansche et al., 1998).

Issues and Trade-Offs

1. Norms versus content and performance standards

While a norm can be developed for nearly any **reference group** by calculating the mean and **standard deviation** of the score distribution, most normed tests use a nationally representative sample of students at particular grade levels and in specific content areas such as spring or fall third-grade reading comprehension. This allows for comparisons that are national in scope. If tests are normed using a local, regional, or state reference group, then results can only be compared at local, regional, or state levels, respectively.

It is useful and appropriate to use nationally norm-referenced tests for the following purposes:

- To determine how well individuals or groups of students perform on broadly-defined content areas, e.g., reading or mathematics, relative to that test's norm group.
- To determine how well individual students perform relative to all other students who took this test.

- To determine how well groups of students, on average, compare to all other students who took the test, e.g., in reading or mathematics. These are referred to as "user" norms.
- To report results in **standard score** form such as **percentile** ranks, **grade-equivalent** scores, or other scaled scores such as those commonly used for college admissions tests (e.g., **SAT** scores or **ACT** scores).
- To report student performance on a measure that was *not* designed to measure achievement of local or state curricula. This can be desirable in the case of entry examinations when applicants come from diverse settings. For example, colleges and universities use tests such as the ACT and SAT in making acceptance decisions because they are designed to predict success in college rather than measure how well students have achieved relative to a particular curriculum.

It is useful and appropriate to use criterion-referenced or standards-based tests for the following purposes:

- to compare the performance of individual students to a criterion score or a performance standard representing a certain level of proficiency relative to a set of learning objectives or content standards.
- to determine the percentage of students in a school, district, or state who attain the criterion or standard or different levels of the standard.
- to compare the percentages of students at different schools who attain the criterion or standard or perform at different levels of the standard.
- to evaluate the extent to which students are achieving relative to a particular curriculum.

Norm-referenced, criterion-referenced, or standards-based tests may be used to track progress over time, provided the test forms are **equated.**

A combination of norm-referenced and criterion-referenced or standards-based assessments is most desirable in the design of a comprehensive assessment system. Depending on the purpose of the system, and as described in Chapter VI, this can be accomplished without testing every student on every item.

2. Test development, reporting, and interpretation

Test interpretation—in terms of norms, criteria, or standards—is discussed under issue #1. However, if such interpretations are to be **valid,** careful attention must be given to issues of test development and reporting.

- Nationally norm-referenced tests, and the accompanying norms, are typically purchased from a vendor. These vendors employ item writers to draft test items. Different vendors have different criteria for item writers, but in general item writers must have content expertise and knowledge of how students at different stages of development learn.

  Test content is most often based on national consensus about the knowledge and skills students should achieve at different grade levels. Test developers may be guided by the content standards developed by such professional organizations as the National Council of Teachers of Mathematics (NCTM), the National Council of Teachers of English (NCTE), or the National Academy of Sciences. They may also consider the national consensus frameworks used for the National Assessment of Educational Progress (NAEP). Test content also may be based on common elements found in curricula across states and textbooks used by a number of states. These practices may result in only partial overlap between the test and local curricula.

  In best practice, draft test items are reviewed for technical soundness, accuracy, and bias. They are then **pilot tested** using a relatively small sample of students. Based on results of

the pilot study, the items are revised as warranted and then **field-tested** to provide data to evaluate the items and test forms and to ensure that the testing instructions and procedures result in **standardized** and effective test administration. Numerous studies are conducted to document the **reliability** and **content validity** of the tests and to ensure that the tests contain items that represent the entire range of possible test performances, from low to high.

- The procedures used to develop criterion-referenced test items are similar to best practices in developing norm-referenced tests. This is because the primary difference between norm- and criterion-referenced tests is the manner in which the scores are interpreted. The items can, in fact, look very much the same, but the tests may cover different content or emphasize different skills.

  Although norm-referenced tests are usually developed nationally, it is more common for criterion-referenced tests to be "home-grown." This is because criterion-referenced tests tend to be developed around state standards or local curriculum. Some criterion-referenced tests are developed entirely without the use of vendors, some are developed entirely by vendors, and some are developed by contracting with a vendor who works with teachers and others to develop appropriate items.

  With norm-referenced tests, the test developers determine the content to be tested. With criterion-referenced tests, the content is determined by the learning objectives established at the state or district level.

  The development of standards-based tests requires the development of content standards. It additionally requires the development of a system of performance standards. This is because the performance standards define the criterion that will be used to document the extent to which students achieve the content standards. An excellent document on the development of performance standards is Hansche (1998).

- The results of norm-referenced tests are reported in ways that allow interpretation relative to a norm group. This is done by using **scaled scores**—statistical transformations of **raw scores**—putting results from different testings on the same scale. Commonly used scaled scores include percentile ranks and grade-equivalent scores. Whatever scores are reported, their meaning should be readily understood by those who will receive and use the reports. This means clear explanations must accompany reports.

- The results of criterion-referenced and standards-based tests are reported in terms of whether students met the criterion or standard for proficiency, excellence, etc. At the student level, these results are often reported on a pass/fail basis or by the performance level at which they scored, e.g., basic, proficient, or advanced in NAEP. Criterion-referenced and standards-based test results can also be reported in terms of scaled scores (e.g., 250), since the performance levels are determined by points along the scale score distribution. When students' scores are **aggregated** or combined into groups by class, grade level, school, district, or state, they are typically reported in terms of the percentage of students who obtained scores within the criterion or standard ranges, e.g., pass, fail, novice, proficient, etc.

- Whether score reports contain norm-referenced, criterion-referenced, or standards-based information, a major consideration must be the level of aggregation for the reported scores. While individual score reports for students are often desirable, it is never appropriate to publicly report test results for individual students because of privacy issues. Reporting results at the class level can be useful for instructional planning by teachers and program planning by principals. Public reporting of results at the school level is increasingly common. Parents and communities are interested in knowing how well their schools are doing.

3. Norms, baselines, and benchmarks

We have discussed norms as the range of performance demonstrated by a norm group. When scores are compared to a national norm, they are essentially compared to a national distribution.

- **Baselines** can be above or below the average; they can fall at any point along the score continuum because they represent the test score against which change will be gauged. For example, if a state decides to adopt a new norm-referenced test and the state average score falls at the 53rd percentile the first time the test is administered, the 53rd percentile constitutes the state's baseline. Similarly, if a state uses a standards-based test, a school having 62% of its students scoring at the proficient level or above, for example, the first time the test is administered, could use 62% proficient as the baseline against which to measure progress. In sum, results from subsequent test administrations can be compared to the baseline performance to document progress.

- In the world of business, **benchmarks** represent top levels of performance. They are identified and characterized so that they may be emulated. In education, benchmarks generally refer to the performance standards that are meant to be achieved at particular grade levels. They may include examples of student work that illustrate different levels of student performance. The grades tested are often referred to as "benchmarking" grades.

  Test results for the benchmarking grades are often used to make generalizations about the condition of schooling and learning. For example, states may have content standards at all grades but test only at grades 3, 5, 8, and 11. The results of the benchmark testing are interpreted as being dependent upon learning up to the time of testing. This practice raises issues for states with school-level accountability programs.

  Sometimes tests administered in grade three, for example, are perceived as being a measure of third-grade learning instead of a test of cumulative learning that is administered in grade three. On the other hand, results of tests covering specialized content introduced in a specific grade (e.g., state history), while somewhat dependent on cumulative learning, are generally more accurately interpretable as representing grade-level learning.

4. System components

- Components of an assessment system should be purpose driven. Systems that are based on local goals are usually standards-referenced systems. However, such systems often raise the political issue of how the system compares to others. In these cases, an external referent, such as a nationally normed test, is often added to the system. The point is that a comprehensive system likely requires different kinds of information. Different aspects of the system may carry different weight in decision-making processes. There are also ways to incorporate information that draw upon existing procedures, e.g., using state-NAEP data to gauge how well a state is doing compared to other states or the nation, or designing a test that can serve multiple purposes.

- A system that is primarily criterion-referenced or standards-based would, ideally, include the following components: learning objectives or content standards; a curriculum that is **aligned** to the objectives in terms of subject matter, skills, and cognitive demands; aligned instruction; aligned assessments, including both state and classroom assessments; a criterion or performance standard against which individual student performance on the state assessment can be compared; and score reports that tie test results to the learning objectives or content standards.

- It is difficult to conceptualize a system that consists solely of nationally norm-referenced assessments, unless the system's purpose does not include the assessment of student

attainment of particular standards. However, norm-referenced tests can play an important role in systems when comparisons to national performances are desired.

Conclusions

The key to selecting or developing tests that are valid for their intended purposes is clarity of purpose for the assessment system and the intended use of results. This chapter highlights the trade-offs involved in selecting norm-referenced tests, criterion-referenced tests, and standards-based tests.

Criteria and standards are hallmarks of an aligned system. Norm-referenced testing can provide external reference points. Ultimately, a system that meets more than one purpose will require more than one type of assessment. And even single-purpose systems will require more than one measure, especially if the results of the assessment will be used for decision making or accountability purposes.

Chapter VII Glossary

ACT (American College Tests). Tests administered by the American College Testing service. Results of the ACT are used by numerous colleges and universities in making decisions about student admission.

Aggregated scores. The total or combined performance for all individuals or groups on one test or subtest. For example, a state average usually represents the aggregation of scores for all students/groups of students who took the test.

Alignment. The similarity or match between and among the content standards, performance standards, curriculum, instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. The standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Baseline data. The initial measures of performance against which future measures will be compared.

Benchmarks. Specific statements of knowledge and skills to be demonstrated at the end of a specified range of grades. For example, benchmark content standards may be set at the end of grades 4, 8, and 12 to specify standards to be met by the end of primary, middle, and high school grade ranges. Benchmarks are located on a performance continuum and are used as checkpoints to monitor progress from one level to the next.

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Content validity evidence. Data that illuminate the extent to which (1) the knowledge, skills, and cognitive demands of the learning objectives underlying an assessment are accurately reflected in the assessment; and (2) the assessment adequately covers the **domain** of knowledge, skills, and cognitive demands represented in the learning objectives.

Criterion-referenced. The reference point for interpreting test results using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished

from those interpretations that are made in relation to a norm or the performance of others.

Cut score. A specified point on a score scale at which scores above that point are interpreted differently from scores below that point. Sometimes there is only one cut score, dividing the range of possible scores into "passing" and "failing" or "mastery" and "nonmastery." Sometimes two or more cut scores may be used to define three or more score categories, as in establishing performance standards.

Equated. Two or more forms of a test that yield equivalent or parallel scores for specified groups of test takers. Equating involves converting the score scale of one form of test to the score scale of another form so that the scores are equivalent or parallel.

Field test. A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a **pilot test.**

Grade-equivalent score. Represents a performance level that is typical of students in a particular grade at a particular time of year. In a statistical sense, it is the school grade level for which a given score is the real or estimated median or mean.

Norm group. The group used to establish "typical" or average performance on a particular test. Typical performance is not necessarily ideal performance.

Norm-referenced. Test interpretations whose scores are based on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Percentile rank. The percentage of scores in a specified distribution that fall below the point at which a given score lies.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrates the range of performance for each performance level; and (4) cut scores that separate one level of performance from another.

Pilot test. A test administered to a representative sample of test takers solely for the purpose of determining the properties of the test. See **field test.**

Raw score. The number of items correct.

Reference group. The group of test takers to which a particular test score will be compared.

Reference population. The population of test takers represented by a test's norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or other characteristics at the time of testing.

Scaled scores or derived scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Standard deviation. The average amount that scores in a distribution of scores deviate (differ) on either side of the mean.

Standard score. A type of derived *score* such that the distribution of *scores* for a specified population has convenient, known values for the mean and standard deviation.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Standards-based tests. Test items reflect a pre-established set of **content standards** that specify the knowledge and skills students are expected to acquire as a function of schooling. Results are then interpreted against a set of criteria or **performance standards** that define student performance relative to the content standards represented by the test items.

SAT (Scholastic Assessment Tests). Tests developed by Educational Testing Service and administered by the College Entrance Examination Board. Results of the SAT are used by numerous colleges and universities in making decisions about student admission.

Valid. The accuracy with which a test measures what it purports to measure. See **Validity.**

Validity. (1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Chapter VII References and Resources

Hansche, L. N., Winter, P., Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington, DC.

Chapter VIII: Test Formats

Background

Test formats fall into three basic categories: **selected response, constructed-response,** and **performance.** Often tests include items in more than one format.

Selected-response items require students to select an answer to the item from a list of given options. Common selected-response formats include multiple-choice, matching, and true-false items (Wesman, 1971).

Constructed-response items require students to create their own responses or products rather than choose a response from an enumerated set. Examples include short-answer items, extended-response items, essays, performances, projects, and written compositions.

Performance items include a wide range of assessment tasks that are product- or behavior-based. They are designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises, open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time (e.g., National Education Association, 1993; Paulson, Paulson, & Meyer, 1991; Priestley, 1982).

Issues and Trade-Offs

Issues that should be considered when deciding on types of test formats to use include:

- The purposes and intended uses of the test
- The amount of time available for testing
- The amount of time available for scoring and report preparation
- Cost

1. The purposes and intended uses of the test

A common belief is that selected-response formats, such as multiple-choice, are most appropriate for norm-referenced tests while constructed-response items are most appropriate for criterion-referenced and standards-based tests. This may or may not be true. The point of reference for test interpretation is a separate, but related, issue from test format.

Another common belief is that selected-response item formats cannot assess complex knowledge and skills. Clearly, this is not the case. Consider the Medical College Admissions Tests, the Law School Admissions Test, the Graduate Record Exam, Advanced Placement Tests, and others that include selected-response items designed to assess complex in-depth knowledge and complex skills.

A third widely held belief is that selected-response items are standardized while constructed-response items or performance-based assessments are not. This, too, is a misconception. When tests are standardized, it means that they are uniformly administered and scored across students, locations, and situations. This suggests that any assessment used as part of a large-scale assessment program, especially if accountability is involved, should be

standardized. It is often easier to standardize the administration and scoring of items that have one best response such as selected-response formats.

Given these facts: (1) the reference point for interpreting test results does not need to drive format, and (2) selected-response items can assess complex knowledge and skills, how should item formats be determined? The answer must be, "purpose!"

If a purpose of the assessment system is to assess knowledge and skills that can be accurately assessed via selected-response items, then selected-response items can be an efficient and cost-effective approach.

If a purpose is to assess knowledge and skills that cannot be reliably or accurately assessed through selected-response items, then constructed-response formats might be used. What is gained is validity to purpose. What is lost may be the costs associated with developing and implementing scoring criteria and the time required to turn around test results.

If a purpose is to assess complex performances such as writing a substantiated persuasive argument or applying science knowledge to independently run a lab experiment, then an on-demand or over-time performance tasks should provide more valid measures as well as valuable professional development opportunities for educators. The trade-offs include testing time, training for test administrators, time required for scoring, and the time and expense of developing reliable scoring **rubrics** and **test security.**

If a purpose is to change or influence instruction, then the assessments must mirror the knowledge and skills desired to be taught, as well as the manner in which they are desired to be taught. For example, how writing is tested influences how writing is taught. If teachers know that the test focuses on writing conventions such as spelling, grammar, and punctuation over the organization of ideas and supporting details, they will likely teach accordingly.

In most cases, a combination of test formats is required to accurately assess the extent to which students are achieving desired learning expectations.

2. The amount of time available for testing

In general, selected-response formats require less testing time. For example, about one minute is allowed for each multiple-choice type item, whereas some short-answer, constructed-response items are allowed 10-12 minutes. More complex constructed-response items would, of course, take longer.

Despite their efficiency, selected-response items may not be a good match to testing purpose if the underlying standards call for performances such as "display," "write," or "demonstrate." As discussed in Chapter VI, the time required for testing can be reduced by using sampling procedures. However, without careful test construction, such procedures are unlikely to produce individual test scores that accurately reflect the intended domain of knowledge and skills. Sampling procedures that allow for individual student scores will result in tests that are longer than those that produce group-level scores only.

3. The amount of time available for scoring and report preparation

Since selected-response items can be machine scored, scoring can generally occur more quickly than for performance items requiring scoring by humans. To illustrate, consider the amount of time it will take to train an adequate number of scorers to adequate levels of reliability. Next, assume that 20 scorers have been trained to acceptable levels of reliability. How many papers can each scorer read in a day and still remain reliable? Will each paper be read by more than one scorer, especially if **high-stakes** decisions will be made on the basis of test results? How much time is available for scoring, i.e., between the time of test administration and the time that score reports are required? Answers to questions such as these can help in making decisions about a reasonable number of constructed-response items to include in an assessment, balanced against the kinds of performance required by the underlying content standards.

4. Cost

The costs associated with testing varies, primarily depending on who develops the test and how it is developed. In general, the costs for **off-the-shelf tests** are less than for tests that need to be newly developed. Most off-the-shelf tests that are designed to survey achievement on a large-scale basis (e.g., Stanford Achievement Test, TerraNova, Iowa Test of Basic Skills) consist primarily of selected-response items.

Costs associated with tests requiring new development—for example, tests that are based on a state's content standards—may be more costly even if they consist solely of selected-response items. However, it is usually the case that standards-based tests include more constructed-response and performance items than off-the-shelf tests. While constructed-response items can be no more expensive to develop than selected-response items, the costs may be increased to the extent that constituents, such as teachers, need to travel and be paid to participate in the item development, review, and revision processes as well as the development of scoring rubrics and the scoring of the tests. If teachers will be involved in item writing, test administration, and/or scoring of constructed-response and performance tests, training will be required. Training adds costs; however, it can also provide valuable staff development opportunities.

Regardless of who scores the tests, the scoring of constructed-response and performance tasks is more costly than the scoring of selected-response tests. It involves the costs of training and the convening of scoring panels. However, the costs can be well worth it if the procedure contributes to the validity of the assessments, information to improve instruction and learning, and professional development.

As with the other issues discussed in this chapter and other chapters, the ultimate goal of the system should figure into decisions about test item formats. In many cases, a combination of formats provides the most valid assessment of desired learning outcomes.

Summary/Conclusions

Decisions about what testing formats to use should be primarily influenced by the purposes and intended uses of the test. Other considerations include the amount of time available for testing, the amount of time available for scoring and report preparation, and costs.

Consider the trade-offs associated with decisions in favor of selected-response formats over constructed-response and performance formats as well as vice versa. In most cases, the purposes, intended uses, and validity requirements of the system will lead to the inclusion of multiple formats.

Chapter VIII Glossary

Constructed response items. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

High-stakes. Tests whose results have important, direct, or lasting consequences for examinees, programs, or institutions.

Off-the-shelf tests. "Ready made," commercially available tests that can be purchased "as is" from a test publisher or vendor.

Performance task. A type of test item that is product- or behavior-based. They are designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a

laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Rubrics. Scoring guides for constructed-response questions or performance tasks. Scoring rubrics contain a description of the requirements for varying degrees of success in responding to the question or performing the task.

Selected-response. Test items that require students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Test security. The need to keep tests safeguarded so all students have equal exposure to the test materials and equal opportunities for success. If test security is violated, then some students can be placed at an unfair advantage or disadvantage. When this happens, the validity of tests is violated.

Chapter VIII References and Resources

National Education Association. (1993). *Student portfolios*. Washington, DC: NEA Professional Library.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership, 48*(5): 60-63.

Priestley, M. (1982). *Performance assessment in education and training: Alternative techniques*. Englewood Cliffs, NJ: Educational Technology Publications.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Chapter IX: Test Identification and Development

Background

Who should identify or develop the tests that will be used in an assessment system? In most cases, identifying an existing test is a policy-level decision informed by persons with testing and content expertise. Test development is usually guided by test development professionals. However, the ultimate answer to the question depends on the underlying goals of the system. The issues addressed in previous chapters concerning testing purposes, uses, formats, and the ability to make accurate inferences from test results should all influence the answer. Practical issues concerning the availability of resources, including time, money, manpower, and expertise, should also influence the answer.

Considerations and options include the following:

- "homegrown" assessments
- off-the-shelf assessments
- customized assessments
- the role of teachers in test development processes

Issues and Trade-Offs

1. Homegrown assessments

"Homegrown" assessments are assessments that are designed and developed by the state or other group that will use them. They are usually designed to be criterion-referenced or standards-based tests in that they are designed to assess a particular set of learning objectives or content standards.

Particularly if the assessment results will be used for decision making about students, programs, or schools, careful attention must be given to the technical soundness of the assessment items and instruments. They must be **reliable, fair,** free from **bias,** and **valid** for their intended uses.

The process for test development includes the following steps:

- *Stakeholders reach agreement and clarity about the purposes and intended uses of the assessments.* The ultimate decision about purpose often results from legislation or action of the state board of education. Ideally, the decision-making process is informed by parents, educators, and the community at large.
- *Based on the agreed-upon purposes and intended uses, assessment design experts, in collaboration with subject matter and child development experts, develop an assessment system framework.*
- *Learning objectives or* ***content standards,*** *which specify the knowledge and skills students are expected to achieve as a function of schooling, are established.* These standards should represent consensus among experts in child development, subject matter content, and

pedagogy. The standards may be separated into content areas or integrated across content areas. They may be developed for each grade level or span several grade levels. Examples of states that have differentially approached the development of content standards are provided in Hansche (1998). The content standards should drive decisions about testing formats, as described in Chapter VIII.

- *Performance descriptors are agreed upon for describing acceptable test performance.*
- *Test specifications are developed to provide a framework (test blueprint) for the numbers and kinds of items that need to be developed in order to accurately assess the learning outcomes represented by the content standards.* It is usually desirable to develop more than one **test form** so that there is a form available for make-up purposes or for use in future years. The test blueprints for state assessments are available on a number of state Web sites.
- *Item writers are trained to develop items and scoring criteria or* ***rubrics*** *in accordance with the test specifications.* More items will need to be written than will be used for any one form of the test. This is due to the desirability of having more than one test form and because some of the items will not successfully make it through all of the development processes such as bias review, **pilot testing,** and **field-testing.**
- *Items and rubrics are drafted and submitted to review for content, technical soundness, accessibility, and freedom from bias.*
- *As a function of the item and rubric review, items and/or rubrics are revised or discarded as necessary.*
- *Items and rubrics are pilot tested with a limited sample of students.*
- *As a function of experience with the pilot testing, items and rubrics are revised as warranted.*
- *Resulting items are used to construct tests in accordance with the test specifications and field-tested. Field test results are analyzed to determine whether each item is technically sound, and the results are used to finalize test administration procedures.*
- *Test forms are developed according to test specifications, using items that have been shown to be sound.*
- *If the tests will be used for accountability purposes,* ***performance standards*** *are developed to define and depict the ranges of performance deemed as proficient, advanced, etc.* Processes for developing systems of performance standards are described in Hansche (1998).
- *Cut scores are established to distinguish between the different performance levels.*
- *Those responsible for test administration are trained to administer the assessments in a standardized manner.* They are also trained in the administration of tests requiring accommodations in accordance with state policy and students' Individual Education Plans (IEPs), 504 Plans, or Limited English Proficiency (LEP) Education Plans.
- *Those responsible for scoring the tests are trained to score them to an acceptable level of reliability.* The higher the stakes associated with test results, the higher the level of acceptable reliability should be. Some states develop their own assessments but contract with a vendor to do the scoring. Other states employ teachers to score the responses of students who are not in their classes because the activity is considered to be good professional development.
- *Tests are scored and results are analyzed and reported.* If the state has contracted with a vendor to score the tests, it may also contract to have the results analyzed and reported.

It is important to ensure that the reports can be easily understood and used by those who will receive them. It is typical (1) for students, parents, and teachers to receive individual student reports; (2) for teachers and principals to receive reports for schools and classes within schools; (3) for superintendents and school boards to receive reports for the district and schools within the district; and (4) for the community-at-large to receive **school report cards** that provide information about how well schools are doing, including information about student achievement overall. For example, a school report card may contain the average score for each grade level, or the percentage of students scoring at the proficient level in each content area tested, as well as additional information about the school, such as class size, teacher credentials, and attendance rates.

- *Assistance is provided in the interpretation and use of results.* This can involve professional development opportunities for teachers, administrators, and school boards. If a vendor has facilitated the scoring and report development process, it may also provide such technical assistance and training.
- *Development and revision continues* to ensure equivalent and valid testing over time. As test forms are used, new items must be developed and subjected to the same rigorous review as the original item sets. If comparisons are to be made over time, new test forms must be **equated** to the original test forms. Quality control and assurance procedures are incorporated throughout the processes of development, revision, and maintenance.

2. Off-the-shelf assessments

Off-the-shelf tests are "ready made," commercially available tests. Since they have already been developed, they may be purchased directly from vendors or test publishers, usually along with other services such as scoring, report preparation, and result interpretation. Many publishers of **norm-referenced** tests also provide student data management systems to schools and local districts. It is also common for publishers to have several equivalent forms of a particular test.

While the purchase of off-the-shelf tests can relieve states from the burden of test development, criteria should be established for test selection. The criteria will be driven by the purposes and intended uses of the test. For example, if a state is looking for a test to complement its criterion-referenced or standards-based assessments by providing a national norm comparison, it may or may not be interested in a significant amount of overlap in the content and skills assessed by the two assessments. The system's purpose will determine the desired range of overlap and the off-the-shelf test that most closely matches the state's intent should be the one selected. Most states use a review process that includes having teachers and other educators determine the match between off-the-shelf assessments and state content and performance standards.

Test reviews are available in Buros' *Mental Measurements Yearbook* and *Test Critiques.* They highlight the technical properties, appropriate uses, and misuses of many published tests. Test publishers also provide technical manuals describing the test development process, the technical properties of the test, and specimen test items.

3. Customized assessments

Customized assessments are often developed through contracts with test development vendors to meet the specific needs of a state assessment system. In such cases, it is common for contracts to be awarded on the basis of responses to Requests for Proposals (RFPs). Via RFP processes, states (or others) write and issue a RFP to develop a system that meets their needs. In some cases, meeting a state's needs requires the development of entirely new assessments. For example, the *Wyoming Comprehensive Assessment System* uses customized assessments developed by a contractor as the primary indicator of student achievement. In other cases, a partially customized assessment is adequate. For example, the ninth edition of

the *Stanford Achievement Test* was augmented for use in California to better address California's content standards.

The quality of the RFP is critical because it determines the quality of the contracted assessments. RFPs must be clear about

- the purposes of the desired system
- the amount of time students are expected to spend in testing
- the cost parameters
- any services beyond test development that will be required
- who will be tested as well as the content and skills in which they will be tested
- expectations regarding item formats, item equity review processes, item bias/sensitivity concerns, requirements for item writers, number of test forms needed, and data analyses
- specifications for technical analysis and special studies
- any potential consequences that might be applied on the basis of test results
- the levels of aggregation at which score reports will be needed
- reporting requirements

If the contract will be to develop a criterion-referenced or performance-based test, the contractor will use approximately the same process as previously described for the "home-grown" tests, except that the developers will be the contractors rather than teachers and others supervised by the state. The RFP, however, can and should specify the extent to which the contractors are expected to involve teachers or others in the development processes, including item review and pilot testing.

4. Role of teachers

In general, teachers are not trained to write test items that have the technical properties necessary for accountability testing. Hence, it may be inappropriate or too time-consuming to include them in the initial drafting of items, although some testing programs have used teachers in this capacity with a degree of success. Nevertheless, it is imperative for teachers to be involved in developing the assessment framework and in item review and pilot testing processes.

Teachers might also be involved in evaluating the alignment among the standards, test specifications, test items, and scoring rubrics. Regardless of the extent to which teachers are involved in such evaluation activities, test vendors should not be the sole evaluators.

When teachers are involved in item writing, extensive training is required. Teachers as item writers also raise issues of **test security,** especially relative to any consequences that will be applied on the basis of test results.

Summary/Conclusions

A number of options are available with regard to test adoption. They include developing your own, buying one that already exists, or hiring someone to develop one for you. The route you choose should be determined by considerations of the overall purpose for the system, the manner in which test results will be used, the amount of time available for adopting or developing a sound assessment, and costs, including human capital and expertise.

Whatever the route, consider who needs to be involved in the process—or at various points in the process—as well as implications for professional development, technical support, and public relations. It is also important to realize that test development is a

dynamic, long-term process, requiring periodic review and revision of the tests and the content and performance standards upon which they are based. The following list provides a summary of the critical considerations in test development.

- the goals, purposes, and intended uses of tests and test results
- availability of resources
- technical soundness of tests, scoring, and interpretation of results
- role of stakeholders
- spectrum of expertise embodied by test framers and developers
- degree of alignment with standards
- issues of bias and fairness
- number of test forms
- test security
- role of teachers in test development, administration, scoring, and interpretation of results
- training needs
- levels of reporting (e.g., student, school, state)
- mechanisms for periodic test revisions
- amount of time available for testing
- who will be tested
- what will be tested
- consequences associated with test results

Chapter IX Glossary

Bias. In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

Construct. The underlying theoretical concepts or characteristics a test is designed to measure.

Customized assessments. Assessments that are customized or tailor-built to meet a particular need. Usually they are developed to cover a particular set of content standards.

Errors of measurement. The differences between observed scores and the theoretical true score; the amount of uncertainty in reporting scores; the degree of imprecision that may result from the measurement process (e.g., test content, administration, scoring, or examinee conditions), thereby producing errors in the interpretation of student achievement.

Fair tests. Yield student scores that are not influenced by such irrelevant factors as native language, prior experience, gender or race.

Field testing. The administration of a test in order to check the adequacy of testing procedures and generally includes attention to test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a **pilot test.**

High-stakes. Tests whose results have important, direct consequences for examinees, programs, or institutions.

Norm-referenced. Test interpretations whose scores are based on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Off-the-shelf tests. "Ready made," commercially available tests that can be purchased "as is" from a test publisher or vendor.

Pilot test. A test administered to a representative sample of test takers solely for the purpose of determining the properties of the test. See **field test.**

Reference population. The population of test takers represented by a test's norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or other characteristics at the time of testing.

Reliable. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and hence are dependable, and repeatable; the degree to which scores are free of **errors of measurement**

School report cards. Reports that provide information about schools, as a whole, rather than about individual students. For example, they may include information about the number of students who score at the proficient level on state tests, information about the number of teachers teaching in their areas of primary training, as well as information about attendance, retention, and discipline referrals. In some cases, the data on school report cards are used to make programmatic decisions about schools or to determine whether they meet accreditation criteria.

Test forms. Parallel or alternate versions of a test that are considered interchangeable in that they measure the same **constructs,** are intended for the same purposes, and are administered using the same directions.

Test security. The need to keep tests safeguarded so that all students have equal exposure to the test materials and equal opportunities for success. If test security is violated, then some students can be placed at an unfair advantage or disadvantage. When this happens, the validity of tests is violated.

Valid. The degree to which a test measures what it purports to measure. See **Validity.**

Validity. (1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Chapter IX References and Resources

Hansche, L. N., Winter, P., Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington, DC.

Chapter X: Test Preparation

- Background
- Preparing Students to Perform Well on Tests
- Ethics
- Summary/Conclusions
- Chapter Glossary
- Chapter References and Resources

Background

Common sense and research show that being well grounded in the content that will be tested and having good test-taking skills can improve test performance. However, it is important to distinguish between **teaching to a test** versus **teaching the test.**

Teaching to a test means that students are taught the knowledge and skills embedded in the learning objectives or content standards upon which a test is based. Assuming that the content standards represent consensus about what is important for students to know and be able to do, then teaching students to be successful relative to the standards is desirable.

Teaching the test, however, is cheating on the part of the teacher. In addition, it cheats students of learning. Teaching the test means that students are exposed to actual, or very similar, test items prior to actual test administration. State and local education agencies have the responsibility of ensuring that teachers and other test users adhere to codes of fair and ethical testing practice.

This chapter deals with two sets of issues: (1) ways to help students improve test performance and (2) ethical issues involved in helping students improve test performance.

Preparing Students to Perform Well on Tests

A number of checklists offer items for consideration when preparing students to do well on tests (e.g., Pike, 1973; Sabers, 1975; Utah Department of Education, 1999a & 1999b; Wahlstrom, 1998). Examples are provided in the Appendix. Readers are cautioned to use such guidelines within the context for which they were developed. Ultimately, if test preparation practice does not teach any content other than what's on the test, it is inappropriate. The best test preparation consists of high quality teaching of content throughout the year.

The examples provided in the Appendix primarily apply to the use of **selected-response** tests with young children. Mehrens, Popham, and Ryan (1998) present other, more general suggestions pertaining to the K-12 spectrum, and relevant to assessments that are more performance-oriented:

Guideline 1

"Determine whether the interpretation to be drawn from the student's performance is related only to the specific task or whether an inference is to be made to a broader domain of performance tasks." It is critical to know whether the assessment tasks are designed to show that students (1) can do the assessment task; (2) can do tasks like the assessment task; or (3) have the knowledge and skills, including cognitive skills, required to do the task such that they can apply these skills to all such tasks. For example, is a writing task designed to determine how well a student can (1) write an essay to the specific prompt used, (2) write an essay in the same mode of discourse called for by the prompt, or (3) write essays in general?

Guideline 2

"When the inference is to the broader domain, one should not instruct in any fashion that would minimize the accuracy of the inference to the broader domain." In most cases, it would be unethical to spend any time teaching to the specific performance task on the assessment. Students may master a particular task but be unable to generalize beyond the task. For example, if a teacher knew that the extended-response mathematics task on the state assessment would focus on probability, it would be inappropriate to limit instruction in mathematical reasoning to contexts that relied solely on probability.

Guideline 3

"Make certain that the student is not surprised, and hence confused, by the performance assessment's format." Just as proper preparation for multiple-choice type tests can be appropriate in moderation, it is also appropriate for students to be familiar, in general, with performance assessment formats. Messick (1994) makes the point that some aspects of a performance assessment may require skills "having nothing to do with the focal constructs in question, so that deficiencies in the construct-irrelevant skills might prevent some students from demonstrating the focal competencies (p. 16)." If the goal is to eliminate measurement error due to factors besides the content being assessed, it is advisable for students to have practice with the assessment format. For example, if the assessment includes tasks requiring students to compare and contrast different versions of historical events (assuming also that this is part of the state's content standards), instruction should include reading and analyzing such documents.

Guideline 4

"Identify evaluative criteria in advance of instructional planning, and communicate these to students." Students who are aware of how their performances will be evaluated are better prepared to perform well. Discussions concerning the differences in benefits between specific versus generalized criteria are provided by Arter (1993). For example, teachers can use the state's scoring rubrics to score student responses to classroom assessments.

Guideline 5

"Stress transferability of the skills and knowledge assessed by performance tests." Teachers can do this by helping students see how the knowledge and skills they are learning do and do not apply in other situations. For example, if the assessments include tasks that require students to critique a science experiment, students can be shown how these skills can help them design experiments that follow standard scientific criteria and evaluate reports of scientific findings in the popular press.

Guideline 6

"Foster students' self-evaluation skills." This guideline is related to guideline #4. Once students have learned the criteria for good performance, they may apply it to their own work. For example, students can be asked to evaluate their own work using scoring rubrics and compare their evaluations to those of their teacher.

Mehrens, Popham, and Ryan (1998) emphasize that, just as with preparing students for taking multiple-choice tests, a balance is needed in preparing students for performance assessments. "Teachers should assess relevant content and should teach that content—but not in a manner that corrupts the inferences that individuals wish to draw from assessment scores" (p. 21).

Ethics

In a 1989 article, Mehrens and Kaminski drew some useful distinctions along the continuum of test preparation practices. Their conclusions most pertinent to **large-scale** educational testing are summarized below:

1. General instruction on objectives not determined by looking at the objectives measured on standardized tests is always ethical.

2. Practice or instruction on a published parallel form of the same test and practice or instruction on the same test is always unethical.

3. Teaching test-taking skills is typically considered to be ethical. According to Mehrens and Kaminski, "most measurement specialists believe that making students equally test-wise will increase validity" (p. 16).

4. The point marking the cross-over between legitimate practice and illegitimate practice can fall at any of the following three points, depending upon the inferences that will be drawn on the basis of test results:

   - Instruction on objectives generated by a commercial organization where the objectives may have been determined by looking at objectives measured by a variety of standardized tests.
   - Instruction based on objectives that specifically match those on the standardized test to be administered.
   - Instruction on specifically matched objectives where practice or instruction follows the same format as the test questions.

Mehrens and Kaminski (1989) also provide a review of several commercial test preparation products from the perspective of ethicalness for particular uses. For information on the effectiveness of test coaching, the reader is referred to Bangert-Drowns, Kulik, and Kulik (1983); Scruggs, White, and Bennion (1986); Samson (1985); Byrd (1987); Deaton, Halpin, and Alford (1987); Shepard (1987); and Kulik, Kulik, and Bangert (1984). While results are mixed, in general they point to the benefits of teaching students to be test-wise.

The Utah State Office of Education (1999b) has provided its teachers and administrators with lists of test preparation practices deemed as ethical or unethical. The lists are included here (see Table X-1), courtesy of the Utah State Office of Education, as a ready summary.

Table X-1. Ethical and Unethical Test Preparation Practices

Ethical—Not Considered Teaching the Test	Unethical—Teaching the Test
• General instruction on objectives that were not determined by looking at any set of published test objectives (e.g., the Core Curriculum).	• Instruction *limited* to objectives that specifically match those on the test to be taken (e.g., excluding parts of the Core Curriculum that are not covered by the CRTs).
• Regular classroom instruction dealing directly with the content of the test (e.g., teaching the Core Curriculum, familiarizing students with a variety of terminology, including what they are likely to encounter on the test), but employing classroom assessments that represent a variety of formats.	• Instruction based on objectives that specifically match those on the test to be taken following the same format (e.g., using the same content, scenarios, activities) as the test questions.
• Instruction covering test-taking skills relating to a variety of test formats incorporated throughout the year into regular instruction.	• Special instruction and practice based directly on a *current or previous* form of the test.
• Increasing motivation for improved performance through appeals to students, parents, and teachers about the importance of taking testing seriously.	• Using questions from *current or previous* forms of the test, or any practice questions that are parallel to those on the test, as practice tests before students actually take the test.
• Checking answer sheets to make sure that each has been properly completed (only to the extent that the test developer recommends it, or all units that are being compared engage in the same practice).	• Using commercially prepared score-boosting materials, or other activities aimed specifically at boosting scores.
	• Dismissing low-achieving students on testing day to boost scores artificially.

Finally, Fred Brown (1984) admonishes us to consider the *Standards for Educational and Psychological Testing* from the context of ethics, especially since the interpretation of test results can affect individuals' lives and livelihoods. In doing so, he draws upon the American Psychological Association's *Ethical Standards for Psychologists* (1997). These principles hold lessons for the developers and users of educational tests as well. The most salient (principles 1, 2, 5, 6, and 8) are summarized here, because they provide a context for policy making relative to the adoption, development, and use of tests.

Principle 1: Responsibility

This principle speaks to "objectivity and integrity" in the development of tests and testing procedures, as well as to the responsibilities of testing professionals to "accept responsibility for the consequences of their work and to ensure . . . that their services are used appropriately."

Principle 2: Competence

Testing professionals are responsible for maintaining "high standards of professional competence." . . . They "recognize the boundaries of their competence and the limitations of their techniques and only provide services, use techniques, or offer opinions as professionals that meet recognized standards." They "maintain knowledge of current scientific and professional information related to the services they render."

2d: Those "with the responsibility for decisions involving individuals or policies based on test results have an understanding of educational measurement, validation problems and other test research."

Principle 5: Confidentiality

"Safeguarding information about an individual . . . is a primary obligation" of the testing professional."

5b: Information or evaluative data concerning children or students "are discussed only for professional purposes and only with persons clearly concerned with the case. Written and oral reports should present only data germane to the purposes of the evaluation and every effort should be made to avoid undue invasion of privacy."

Principle 6: Welfare of the consumer

Testing professionals "respect the integrity and protect the welfare of the people and groups with whom they work. . . . [They] fully inform consumers as to the purpose and nature" of an evaluation or procedure and they inform students or other appropriate clients of their rights relative to participation.

Principle 8: Utilization of assessment techniques

"In the development, publication, and utilization of . . . assessment techniques," testing professionals observe relevant standards of their professional organizations such as the American Psychological Association. Persons tested "have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgments were based. Test users avoid imparting unnecessary information which would compromise test security, but they provide requested information that explains the basis for decisions that may adversely affect that person" or entity.

8b: "When a test is published or otherwise made available for operational use, it is accompanied by a manual (or other published or readily available information) that fully describes the development of the test, the rationale, and evidence of validity and reliability. The test manual explicitly states the purposes and applications for which the test is recommended and identifies special qualifications required to administer the test and interpret it properly. Test manuals provide complete information regarding the characteristics of the normative population."

8c: "In reporting test results, testing professionals indicate any reservations regarding validity or reliability resulting from test circumstances or inappropriateness of the test norms for the person tested. They strive to ensure that the test results and their interpretations are not misused by others."

8e: Testing professionals or organizations offering test scoring and interpretation services "are able to demonstrate that the validity of the programs and procedures used in arriving at interpretations are based on appropriate evidence . . . every effort is made to avoid misuse of test reports."

Summary/Conclusions

While there is no substitute for the teaching and learning of content, students' test scores can be improved via instruction in test taking skills. The extent to which such practices are ethical is dependent upon the "closeness of the match of the preparation materials to the tests and the inference one wishes to make from the test scores" (Mehrens & Kaminski, 1989).

Chapter X Glossary

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Equated. Two or more forms of a test that yield equivalent or parallel scores for specified groups of test takers. Equating involves converting the score scale of one form of test to the score scale of another form so that the scores are equivalent or parallel.

Large-scale. Assessment programs that test or assess relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale programs are in contrast to tests and other assessments administered on a smaller scale, for example, by classroom teachers for instructional purposes.

Selected-response. Test items that require students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Teaching the test. Teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

Teaching to a test. Teaching the broad-based knowledge and skills represented by a test's underlying content standards. Compared to **teaching the test,** it is not cheating.

Chapter X References and Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

American Psychological Association (1997). *Ethical standards for psychologists.* Washington, DC: Author.

Arter, J. (1993, April). *Designing scoring rubrics for performance assessments: The heart of the matter.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED 358 143).

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. *Review of Educational Research, 53,* 571-585.

Brown, F. G. (1984). *Guidelines for test use: A commentary on the standards for educational and psychological tests.* Washington, DC: National Council on Measurement in Education.

Byrd, M. (1987). *A comparison of the effectiveness of four test preparation programs* (Final Evaluation Report). Chicago: Chicago Public Schools, Department of Research and Evaluation.

Deaton, W. L., Halpin, G., & Alford, T. (1987). Coaching effects on California Achievement Test scores in elementary grades. *Journal of Educational Research, 80,* 149-155.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435-447.

Mehrens, W. A., & Kaminski, J. (1989, Spring). Methods of improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and practice,* 14-21.

Mehrens, W. A., Popham, J. W., & Ryan, J. M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*(1), 18-22.

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

Pike, E. O., Jr. (1973). *Influence of a test-taking skills instructional program on the Metropolitan Achievement Tests performance of children from low-income families.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Sabers, D. L. (1975). *Test-taking skills.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Samson, E. (1985). Effective training in test-taking skills on achievement test performance: A quantitative syntheses. *Journal of Educational Research, 78,* 261-266.

Scruggs, T. E., White, K. R., & Bennion, K. (1986). Teaching test-taking skills to elementary-grade students: A meta-analysis. *The Elementary School Journal, 87,* 69-82.

Shepard, L. A. (1987). *A case study of the Texas Teacher Test: Technical report.* Los Angeles: University of California, Graduate School of Education, Center for the Study of Evaluation.

Utah State Office of Education. (1999a). *Test-taking tips and strategies: Student/parent pamphlet.* Salt Lake City, UT: Utah State Office of Education.

Utah State Office of Education. (1999b). *Test-taking tips and strategies: Teacher/administrator guidelines.* Salt Lake City, UT: Utah State Office of Education.

Wahlstrom, D. D. (1998). *Practical ideas for teaching and assessing the Virginia SOL.* Virginia Beach, VA: Successline, Inc.

Chapter XI: Scoring the Tests: Reliability, Rubrics, and Reality

Background

Reliability is the precision and consistency of test scores, and it is measured by looking at the degree of agreement between test scores that ought to be the same. For example, if a child's test paper is scored by two different scorers, the scores should be the same, or close enough to the same that any differences do not lead to different conclusions.

If we were able to give the same child the same test again, erasing the child's memory of the first test, the two scores should be the same. This indicates that the test consistently measures the child's knowledge and skills.

A similar example applies to tests that are machine scored. Hence, both scorer consistency and score reliability are important aspects of reliability.

The amount of error, typically expressed as the **standard error of measurement** (SEM), that may be associated with any particular score (i.e., differences between **true scores** and obtained scores) should be determined by the test developer and reported. This is why scores are sometimes reported in bands. For example, on average, if a student receives a score of 40 on a test that has an SEM of three, we can be 68% sure that the student's "true" score falls between 37 and 43.

A test with an SEM of six would mean that a student's observed score of 40 could be interpreted as a true score between 34 and 46. Thus, as the SEM increases, the precision with which we can interpret the results declines. Similarly, if we require a higher degree of confidence, the width of the score band will widen. For example, to achieve a 95% confidence level with an observed score of 40 and SEM of six, the score band doubles to 28-52.

The issue of reliability is complex enough when tests consist of items that have a single keyed response and can be machine scored. It is even more complex in the case of performance assessments that often accommodate a range of correct responses and are scored by humans.

This chapter focuses on factors that can affect the reliability of test scores. It also provides some fundamental information about types of test scores.

Types of Test Scores

A student who answers 35 of 50 items correctly would receive a **raw score** of 35. The percent of items correct would be 70%. By itself, a score of 35 has little meaning unless we have a point of reference such as the total number of items on the test and a good idea of the test content and how difficult the test is. The percentage of items correct, however, can be misleading. For example, 70% correct on a difficult test may require more proficiency than 90% on an easier test.

It is for reasons such as this that test developers report scores in a **"standard"** form, which allows for comparisons to the performances of students in a **norm** or **standard-setting group.** Commonly used standard scores include **percentiles, grade equivalents,**

stanines, and **scale scores** on **norm-referenced tests,** and **performance levels** and **scale scores** on **criterion-referenced tests.**

A percentile indicates the percentage of students in the norm group the student scored above. For example, a percentile score of 70 means that the student scored higher than 70% of the students in the norm group. It does not mean that the student correctly answered 70% of the questions. Teachers are sometimes asked why a student's raw score on a test is higher at the end of the school year than at the beginning of the year, but the student's percentile score is lower. It is because more students in the norm group scored higher on the end-of-year norming. An advantage of percentile scores is that they are perceived as being easily understood. A disadvantage is that they can be easily confused with percentage correct.

A grade equivalent score represents a performance level that is typical of students in a particular grade at a particular time of year. For example, a grade equivalent score of 5.5 means that the student scored about as well on the test as the average student in the norm group who is halfway through the fifth grade would score. Grade equivalent scores do not represent the lowest acceptable performance for a grade. Neither do they mean, for example, that a third-grade student who scored 5.5 should be transported to fifth grade. Rather, in this example, it means that the average fifth-grade student, halfway through the school year, would have answered the same number of questions correctly on the third-grade test. An advantage of grade-equivalent scores is that they can provide policymakers with a ready indicator of the proportion of students performing at grade level. A disadvantage is that they can be easily misinterpreted to mean that individual students are more or less proficient than they actually are.

Stanine scores range from 1-9, the term stanine meaning "standard nine." The entire range of scores is divided into parts and each is given a number. Stanine scores of 4, 5, and 6 are in the middle and, hence, are considered in the average range. Stanine scores of 1-3 are considered below average and stanine scores of 7-9 are considered to be above average. Stanine scores can give a rough indicator of where students or groups fall within a range of scores. This can be an advantage if all that is needed is a rough indicator of achievement or progress; however, if specific diagnosis of individual strengths and weaknesses relative to a particular course of study is desired, stanine scores are inadequate.

College admissions tests also report results in standard score units. The type of score reported depends on the test. Test publishers provide interpretive information along with the tests and score results.

When test results are compared to a criterion or standard of performance, results are typically reported according to (1) whether the student met the standard or (2) the level of the standard that was met, (e.g., proficient or advanced). Quite often, standard scores are also reported using a scale (e.g., 100-300) developed for the specific test.

Score Reliability and Generalizability

Given that reliability is the degree of consistency between test scores that ought to be the same, there are a number of ways to measure reliability. The crucial issue in measuring the reliability of an assessment instrument is to identify the major sources of error in test scores (American Educational Research Association et al., 1999), keeping in mind the intended purposes and actual uses of the tests and test results. Ideally, to determine the consistency of a test score, the entire assessment process would be replicated and results from the replication compared to the original results (American Educational Research Association et al., 1999). This is seldom practical, however, and in the real world, such a replication introduces additional sources of error such as student familiarity with the test.

In traditional measurement practice, three measures of reliability are typically considered: **test-retest,** alternate or **parallel forms,** and **internal consistency.** Each of these methods provides an estimate of the amount of error in a student's score versus how much of a

student's score reflects "true" levels of proficiency; however, the methods do not necessarily yield the same results for the same assessment. In general, the method chosen often is based on practical constraints and the structure of the assessment program. These three methods produce a correlation coefficient that is an estimate of "the degree to which scores are free of measurement error" (American Educational Research Association et al., 1999, p. 181). **Reliability coefficients** close to 1.00 are most desirable, but rarely achieved. In this chapter, we also consider **scorer reliability** because it is an important consideration in systems that include performance assessments. Information about the reliability of scores and scorer reliability, when applicable, should be provided in the technical materials furnished by the test publishers.

- **Test-retest** reliability coefficients are obtained by giving the same students the same assessment instrument twice, with a period of time between administrations. The reliability coefficient is the correlation between students' performances on the two occasions. It is usually impractical to use this technique as evidence of reliability for educational tests for several reasons. For example, if the period of time between the two testings is long enough for students to have forgotten the items and tasks on the assessment, it is probably long enough for them to have gained knowledge in the standards tested, making the scores non-equivalent for reasons other than error of measurement.

- **Alternate forms** reliability coefficients are obtained by giving students two parallel forms of the test. The reliability coefficient is the correlation between students' scores on the two forms. Alternate form reliability coefficients take into account both consistency over different times of testing and consistency over different samples of items and tasks.

- **Internal consistency** reliability coefficients are obtained by finding the correlations among items or sets of items on a single test. The average correlation reflects the extent to which the items and tasks on a test contribute toward measuring the same construct, based on the consistency of student responses to all the items (or sets of items) on the test.

While traditional reliability coefficients provide an estimate of the amount of error in test scores, **generalizability theory** allows test developers and users to estimate the error in student scores from different sources such as scorers, items, and occasions. This allows users to better interpret the scores and allows assessment developers to reduce error identified in the field-testing process, making the scores more reliable. Generalizability coefficients reflect the total error and are interpreted in the same way as traditional reliability coefficients.

Score reliability should be examined for both the test score as a whole and for any other score that is reported and used. For example, if subscores are reported and used for instructional purposes, it is important to consider the reliability of the subscores. If scores are used to categorize students by proficiency levels, the reliability of those categorizations should be reported and interpretations should be informed by the consistency of classification. If there is reason to believe that reliability of scores may vary for different subgroups of students (e.g., age, disability status), reliability information should be reported for these groups as practical, especially if scores for categories are reported and used for school and program evaluation and improvement (e.g., Title I requires that scores be disaggregated and reported according to several categories).

As noted earlier, the standard error of measurement is a useful metric for reporting the accuracy of scores. The SEM is directly related to the reliability coefficient and can be computed for the test as a whole, as in the illustration at the beginning of the chapter, or for any critical score on the test (e.g., the cut score between "advanced" and "proficient").

Scorer Reliability and Rubrics

When tests are scored by humans, rather than machines, it raises additional reliability issues. The humans who score tests must use a set of scoring rules and they must be applied consistently from one scorer to another.

The scoring guides used by human scorers include scoring rules that are often called **rubrics.** They contain a description of the requirements for varying degrees of success in responding to the question or performing the task. Scoring guides also contain examples of student responses at each score point illustrating the range of student responses eligible for each level on the rubric. Guidance in the development of scoring rubrics can be found in Harris and Carr (1996) and in Regional Educational Laboratories (1998).

Scorer reliability pertains to the accuracy with which scorers apply the scoring rubrics. A first criterion for selecting scorers is that they meet certain professional standards, (e.g., that persons considered for scoring English literature exams have training in English literature and experience in teaching English). Then, it is important to train selected scorers so that different scorers assign similar scores to the same assessment responses. Those who do not meet the criteria for scoring reliably would continue to receive training until meeting the criteria, or they would be released. The degree of agreement between scorers is called **inter-rater reliability.**

It is also important to check scorers throughout the actual scoring process and over time to ensure that fatigue or other factors do not influence scoring stability. This aspect of scorer reliability is called **intra-rater reliability.** Information about scorer reliability should accompany the technical materials provided by the developers or publishers of tests that are scored using scoring rubrics.

Other Issues

In addition to the issues raised above about the reliable application of scoring rubrics and the reliability of test scores in general, there are issues of

- Who should score the tests
- Test security
- Cost

1. Who should score the tests?

Selected-response tests such as multiple-choice are machine scored. Sometimes the state does the scoring, but more often the state contracts with a test publisher or another vendor to score the tests and prepare score reports.

Constructed response and **performance tests** may be scored by contractors or by education professionals within the state, such as teachers. In either case, the scorers must be trained to acceptable levels of reliability, and the scoring must be monitored to ensure that individual scorers remain consistent over time.

Training teachers to score the tests can be an excellent and effective staff development tool. However, it may not be possible to release enough teachers for training and to score the tests fast enough to meet state needs. It can also raise issues of test security, which can affect the validity of test results and any accompanying decision making about students, programs, or schools.

2. Test security

Test security is ensuring that test results are valid by restricting access to the test items. This may seem counterproductive in the case of achievement tests where the goal is to assess the extent to which students have mastered the content being tested.

However, the purpose of keeping tests secure is not to prohibit **teaching to a test.** It is to prohibit **teaching the test.**

Holding test content secure does not mean that there should be no communication about the tests, their format, or content. In fact, students should be prepared for tests in ways that help reduce anxiety and help them perform to the best of their abilities. States can release sample items and tasks or a portion of the items and tasks administered. Publishing the scoring rubrics and sample student responses can help teachers prepare students for the test.

In addition to jeopardizing the reliability and **validity** of test results, breaches of test security can be very expensive and may even require developing entirely new tests.

Some states, such as Texas, release the state achievement tests to the public after each test administration. Adopting such a practice has implications for the kinds of new tests that can be developed from year to year as well as the associated costs.

3. Cost

In general, the cost of scoring an off-the-shelf, machine-scorable test is the least expensive option. However, such tests rarely, if ever, are fully aligned to state or district standards. The costs associated with the scoring of constructed-response and performance tasks are comparatively high, but such tasks may provide the most valid and useful results, including the potential for professional development.

Summary/Conclusions

The reliability of test scores is important. Tests must be reliable for their intended purposes and uses, and scorers of performance tests must be trained to a criterion of reliability that holds up across scorers and over time.

Scoring rubrics used by reliable scorers must be true to the standards underlying the tests. They can also be valuable teaching aids.

Test security also has a significant impact on the interpretation of test results. If the test content is compromised, test results cannot be interpreted consistently or accurately, and the compromise will have a particularly detrimental effect on decisions about students, programs, or schools. Reliability is a necessary, but not sufficient, condition for validity.

The cost implications associated with scoring must be considered. If scoring is to be done by teachers as part of a professional development effort, the state will also need to invest in training teachers, purchasing teachers' time to do the scoring, and face challenges of getting scores in a timely fashion.

Chapter XI Glossary

Alternate forms reliability. "Alternate forms" is a generic terms referring to two or more versions of a test that are considered interchangeable in that they measure the same constructs, are intended for the same purposes, and are administered using the same directions. Alternate forms are reliable to the extent that the scores of every individual hold their ranks in a score distribution from one alternate form to another.

College admissions test scores. Scores yielded by tests used in college admissions decisions. Two commonly used college admissions tests are the ACT and the SAT. In 1998, the national average ACT scale score was 21. The possible range of ACT scale scores is

from 1 (low) to 36 (high). The standard scores for the SAT range from 200 to 800 and have a mean of roughly 500 and a standard deviation of roughly 100.

Construct. The underlying theoretical concept or characteristic a test is designed to measure.

Constructed response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Correlate. In the statistical sense, two sets of scores are perfectly **correlated** if every individual has the same score rank on one measure as on the other. For example, if the top-scoring individual on measure 1 also obtains the top score on measure 2, the second-best individual on measure is the second best on measure 2, etc., then the measures are perfectly correlated, i.e., the **correlation coefficient** is +1.00. To the extent that the individuals in the score distributions do not maintain their ranks, the correlation coefficient is reduced.

Correlated. When the relationship between the ranks of individuals in different score distributions are statistically examined and described via a correlation coefficient, they are said to be correlated. The correlation coefficient can range from a perfectly negative relationship (-1.00), meaning that the top scoring individual on one measure is the lowest scoring individual on the other measure, etc. to a perfectly positive relationship (1.00), meaning that the top scoring individual on one measure is the top scoring individual on the second measure, the individual with the second best score on one measure has the second best score on the second measure, etc.

Correlation coefficient. The statistical representation of the relationship between two, or more sets of scores. See **correlate** and **correlated.**

Criterion-referenced. The reference point for interpreting test results using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to a norm or the performance of others.

Derived scores or scaled scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Errors of measurement. The differences between observed scores and the theoretical true score. The amount of uncertainty in reporting scores; the degree of imprecision that may result from the measurement process (e.g., test content, administration, scoring, or examinee conditions), thereby producing errors in the interpretation of student achievement.

Generalizability theory. Contributes to reliability by allowing test developers to estimate the amount of error in students' test scores from different sources (e.g., raters, items, testing occasions).

Grade-equivalent score. Represents a performance level that is typical of students in a particular grade at a particular time of year. In a statistical sense, it is the school grade level for which a given score is the real or estimated median or mean.

Internal consistency. The degree to which the test items, on average, **correlate** with the entire test. It is a measure of the extent to which a group of items contribute to measuring the **construct** measured by the test.

Inter-rater reliability. The degree to which different scorers agree on the score to be assigned to a test response.

Intra-rater reliability. The degree to which an individual rater is consistent over time.

Norm group. The group used to establish "typical" or average performance on a particular test. Typical performance is not necessarily ideal performance.

Parallel forms reliability. Parallel forms are a type of **alternate form** that have equal raw score means, equal standard deviations, and equal **correlations** with other measures for any given population. Parallel forms are reliable to the degree to which scores of every individual hold their ranks in the score distribution from one parallel form to another.

Percentile. The score on a test below which a given percentage of scores fall.

Performance-based or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Raw score. The number of items correct.

Reliability. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and, hence, are dependable and repeatable; the degree to which scores are free of **errors of measurement.**

Reliability coefficient. A unit-free index that reflects the degree to which scores are free of **errors of measurement.**

Rubrics. Scoring guides for constructed-response questions or performance tasks. Scoring rubrics contain a description of the requirements for varying degrees of success in responding to the question or performing the task.

Scaled scores or derived scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Scorer reliability. The degree to which the scores assigned by raters are consistent over repeated applications of the scoring rubric. **Inter-rater reliability** and **intra-rater reliability** are types of scorer reliability.

Standard error of measurement. The average amount that scores in a distribution differ from the corresponding **true scores** for a specified group of test takers.

Standard scores. A type of derived *score* such that the distribution of these *scores* for a specified population has convenient, known values for the mean and standard deviation.

Standard-setting group. A group used to inform or establish desired or proficient levels of performance on a particular test. Typical performance is not necessarily ideal performance.

Stanine scores or "standard nine" scores. Derived scores that range from 1-9. Stanine scores of 4, 5, and 6 are in the middle and, hence, are considered in the average range. Stanine scores of 1-3 are generally considered below average and stanine scores of 7-9 are generally considered to be above average.

Teaching the test. Teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

Teaching to a test. Teaching the broad-based knowledge and skills represented by a test's underlying content standards. Compared to **teaching the test,** it is not cheating.

Test-retest reliability. The extent to which the scores of every individual hold their ranks in the score distribution upon repeated administration of the same test to the same individuals.

True scores. In classical test theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In item response theory, the error-free value of test taker proficiency.

Validity. (1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Chapter XI References and Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

Harris, D. E., & Carr, J. F. (1996). *How to use standards in the classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

Regional Educational Laboratories (1998). *Improving classroom assessment: A toolkit for professional developers*. Portland, OR: Northwest Regional Educational Laboratory.

Chapter XII: Validity—Making Valid Inferences from Test Results

Introduction

If the message of this entire document could be expressed in a single word, that word would be "validity." Surely, nearly every issue discussed thus far (e.g., purpose, uses, scoring procedures, and reliability) affect validity. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999), **validity** is "the degree to which a certain inference from a test is appropriate or meaningful" (p. 94). Test scores are often misused. For example, the SAT, which was designed to predict first-year college grades, is often used to make inappropriate inferences about the quality of high schools.

Tests can be neither valid nor invalid per se and the degree to which they are valid can change. Validity is judged on the basis of evidence that the test's development, administration, scoring, reporting, and uses are in keeping with its purposes and the kinds of decisions that will be made on the basis of the test results. Hence, tests are more or less valid for particular purposes. Evidence may be provided in a number of ways. An excellent reference on the aspects of validity is provided by Messick (1989).

Kinds of Validity Evidence and the Issues They Raise

Validity evidence includes the following:

- Face validity
- Content-related validity
- Criterion-related validity
- Construct-related validity
- Curricular validity
- Consequential validity

1. **Face validity** refers to the degree to which tests look like they measure what they purport to measure. For example, a "writing" test that relies solely on multiple-choice questions about the conventions of writing such as grammar, punctuation, and spelling, is lacking in face validity. Although face validity is not always seen as an important piece of validity evidence, it is an important consideration for tests used in K-12 education, particularly when a purpose of the test is to focus attention on the content standards and model good assessment techniques. In addition, it is easier to communicate about tests to parents, policymakers, and publics, if they are face valid.

2. In the world of achievement testing, evidence of **content validity** is absolutely critical. Tests that have content validity have a good match to the underlying learning objectives or **content standards.** The test items will accurately reflect the knowledge and skills,

including cognitive skills, included in the standards. Content validity can be built into tests through the test development processes. These processes must include ensuring that (1) the content standards are reflected in the test and (2) the test reflects the entire scope of the content standards.

3. When used within the context of educational testing, a "criterion" is an indicator of an accepted or desired level of performance. It could be something like a certain grade point average or test score. Criteria usually serve as a standard against which test results are evaluated. **Criterion validity** is the extent to which there is evidence showing that scores on a test are related to a criterion measure. For example, if a test is intended to measure what is learned in a particular course of study, then the test scores and course grades should **correlate** strongly. Or, if the scores of college admissions tests truly predict college performance, the correlation between entrance exam scores and subsequent performance should be relatively high.

4. "Construct" is a term used by psychologists and developers of traditional tests to describe the hypothetical trait they wish to measure. Construct validity is the most important and encompassing type of validity. Constructs include traits such as intelligence, ability, aptitude, and achievement. **Construct validity** is the extent to which a test produces results that accurately reflect the construct they are designed to assess. For example, achievement tests are designed to assess what students have learned as a function of schooling. College admissions tests, on the other hand, are seldom based upon a particular set of learning objectives. Rather, they are designed to assess students' abilities to apply their knowledge and skills to new situations in order to predict their success in college.

   All types of validity evidence outlined in this list contribute to information about the construct validity of a test, but there are additional ways of collecting validity evidence that are not covered by the other categories. Although constructs are hypothetical or theoretical in nature, they are grounded in research and other empirical experience. In demonstrating construct-related validity, evidence must be provided to support the interpretation of test scores in terms of the construct. This can include evidence that (1) the assessment results are positively correlated with the results of assessments designed to measure the same or similar constructs (i.e., **convergent validity** evidence); (2) the results of the assessment do not correlate highly with the results of assessments designed to measure a different, but related, construct, (e.g., mathematics achievement versus reading ability) (i.e., **discriminant validity** evidence); (3) the assessment is sensitive to changes in the construct over time, (e.g., changes in science knowledge as a result of instruction); and (4) the items on the assessment are **internally consistent.**

5. **Curricular validity** is a relatively new term important to considering the validity with which the results of achievement tests can be interpreted. If an achievement test is based upon a particular set of learning objectives or standards, the test results will be valid only if students have had adequate opportunities to learn the curriculum being tested. This means that both the curriculum and the tests must align with the standards relative to knowledge, skills, and cognitive demand.

6. Like curricular validity, **consequential validity** is a relatively new term brought on by the movement toward educational accountability. It essentially asks the question: "Does the assessment system have the desired effects?" How does it affect students, teachers, administrators, the curriculum, and instruction? Great care should be taken during the design and development phases of system development to ensure that the standards, curriculum, assessments, resource allocation, and technical supports align and work cohesively toward the intended purposes and uses of the overall system. Such care will help ensure higher levels of consequential validity.

It is also important to continue to evaluate the assessment system and modify or renew it as needed to continue the bank of evidence for consequential validity. In education, this step often is neglected and programs are either unjustifiably scrapped or promoted.

Threats to Validity

According to Mehrens (1984, p. 10) and Rudner, Conoley, and Plake (1989, p. 53), the only reasonable, direct inference you can make from a test score is the degree to which a student knows the content on the test. Thus, making accurate inferences requires attention to all factors that can affect the validity of such inferences. These factors include, but are not limited to, the following:

- the extent to which the test content and format match the learning objectives, instruction, curriculum, and scoring criteria
- methods for dealing with potential test **bias, fairness,** and the extent to which the tests are appropriate for all students
- test security and the manner in which all students are prepared to participate in the assessments
- reliability of the tests and/or scorers
- most importantly, the match between the purposes of the system and the use of the assessment results

Summary/Conclusions

Validity, or the extent to which inferences made from test scores are accurate, is the central issue in assessment. The greater the stakes associated with decisions involving the results of assessments, the greater the need for validity evidence.

Many factors can affect the validity of interpretations and decisions made on the basis of assessments. These include the **alignment** of the assessment system, including the alignment of the purpose underlying the system and the ways in which the assessment results will be used; the extent to which the system components are fair and free from bias; test security; test and scorer reliability; and the manner in which students are prepared to participate in the assessments.

The fundamental consideration in creating and implementing an assessment system with strong evidence for validity is the match between the system components and uses with its underlying purpose.

Chapter XII Glossary

Alignment. The similarity or match between and among the content standards, performance standards, curriculum, instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. This standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Bias. In a statistical context, systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

Consequential validity evidence. Data that illuminates the extent to which the assessment has the desired effects, e.g., on students, teachers, administrators, the curriculum, instruction and/or other entities..

Construct. The underlying theoretical concept or characteristic a test is designed to measure.

Construct validity evidence. Data that illuminates the extent to which a test produces results that accurately reflect the construct they are designed to assess.

Content standards. Statements of the knowledge and skills that schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Content validity evidence. Data that illuminate the extent to which (1) the knowledge, skills, and cognitive demands of the learning objectives underlying an assessment are accurately reflected in the assessment; and (2) the assessment adequately covers the **domain** of knowledge, skills, and cognitive demands represented in the learning objectives.

Convergent validity evidence. Data showing the degree to which the assessment results are positively **correlated** with the results of other measures designed to assess the same or similar **constructs.**

Correlate. In the statistical sense, two sets of scores are perfectly **correlated** if every individual has the same score rank on one measure as on the other. For example, if the top-scoring individual on measure 1 also obtains the top score on measure 2, the second-best individual on measure is the second-best on measure 2, etc., then the measures are perfectly correlated, i.e., the **correlation coefficient** is +1.00. To the extent that the individuals in the score distributions do not maintain their ranks, the correlation coefficient is reduced.

Correlated. When the relationship between the ranks of individuals in different score distributions are statistically examined and described via a correlation coefficient, they are said to be correlated. The correlation coefficient can range from a perfectly negative relationship (-1.00), meaning that the top scoring individual on one measure is the lowest scoring individual on the other measure, etc., to a perfectly positive relationship (1.00), meaning that the top scoring individual on one measure is the top scoring individual on the second measure, the individual with the second best score on one measure has the second best score on the second measure, etc.

Correlation coefficient. The statistical representation of the relationship between two or more sets of scores. See **correlate** and **correlated.**

Criterion validity evidence. The extent to which there is evidence showing that scores on a test are related to a criterion measure. For example, if a test is intended to measure what is learned in a particular course of study, then the test scores and course grades should **correlate.**

Curricular validity evidence. The extent to which there is evidence that students are taught a curriculum that aligns with the assessments and the learning objectives or content standards on which the assessments are based.

Discriminant validity evidence. Data showing the results of an assessment do not correlate highly with the results of assessments designed to measure a different, but related, construct, e.g., achievement versus ability.

Domain. The portion of all knowledge and skill in a subject matter area that is selected for the content standards once consensus is reached that it represents what is important for teachers to teach and students to learn.

Errors of measurement. The differences between observed scores and the theoretical true score. The amount of uncertainty in reporting scores; the degree of imprecision that may result from the measurement process (e.g., test content, administration, scoring, or examinee conditions), thereby producing errors in the interpretation of student achievement.

Face validity evidence. Tests that measure what they purport to measure. For example, a "writing" test that relies solely on multiple-choice questions about the conventions of writing such as grammar, punctuation, and spelling, is lacking in face validity.

Fair tests. Yield student scores that are not influenced by such things as native language, prior experience, gende, or race.

Internal consistency validity evidence. The degree to which the test items, on average, **correlate** with the entire test. It is a measure of the extent to which a group of items contribute to measuring the **construct** measured by the test.

Reliability. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and, hence, are dependable and repeatable; the degree to which scores are free of **errors of measurement.**

Validity. (1) an overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Chapter XII References and Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? *Educational Measurement: Issues and Practice, 3*(3), 9-15.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

Rudner, L. M., Conoley, J., & Plake, B. (Eds). (1989). *Understanding achievement tests: A guide for school administrators.* Washington, DC: American Institutes for Research.

Chapter XIII: Special Populations

Introduction

Today's students are more diverse in their characteristics than ever before, and many of them have "special needs." Included in these special populations are students with disabilities; students who are learning to speak English; and students whose families are migrant workers, moving from one area of the country to another.

While participating in **large-scale assessment** programs has numerous benefits for students with special needs, including them creates some unique challenges and issues for states to grapple with and resolve in order to maximize the usefulness of the assessment results.

This chapter clarifies the reasons it is important to include these students in large-scale assessments, defines the critical issues to consider (e.g., participation, accommodations, and reporting), and identifies key strategies for maximizing the inclusion of special populations in large-scale assessment programs. When relevant, distinctions will be drawn between the special needs subgroups of **students with disabilities** and **English language learners.**

Guiding Questions

Before addressing issues related to including special needs students in large-scale assessment systems, two underlying questions must be answered: (1) Who are students with special needs? and (2) Why is it important to include special populations in assessments? These two questions are addressed briefly here. Literally volumes are now being written on these questions, and some of these sources are cited in the discussion.

1. Who are students with special needs?

Special needs students reflect some of the diversity of students in schools today. Typically included among those students considered to be "special needs" are students with disabilities and students who are learning to speak English. Other groups of students, such as those whose families are migrant workers and those who are homeless, also often have special needs in instruction and assessment; these students are discussed here only as they are represented within the broader special needs groups of students with disabilities and English language learners.

Students with disabilities include youngsters who are eligible for special education services under the Individuals with Disabilities Education Act (IDEA), those who are eligible for accommodations under a "504 Plan," and those who may not receive any special services. It is the first two groups of students who have the greatest impact on assessment systems.

Students who receive special education services have Individualized Education Programs (IEPs) that specify, among other things, the goals of their instruction, whether they participate in state or district assessments or an **alternate assessment,** and the **accommodations** they must receive during instruction and assessment. Students with IEPs have one or more of 13 federally defined categories of disability (autism, deafness, deaf-blindness, hearing impairment, learning disability, mental retardation, multiple disability, orthopedic impairment, other health impairment, serious emotional disability, speech/language impairment,

traumatic brain injury, visual impairment). As of 1999, nearly 12 million youngsters were receiving services under IDEA (U.S. Department of Education, 2000). This number has risen steadily over time, and is expected to continue to increase.

Students with disabilities also are protected by Section 504 of the 1973 Rehabilitation Act. This civil rights legislation determined that students have the right to a free and appropriate education, regardless of the severity of a person's disability, and specifically that reasonable accommodations must be provided to ensure access to education, even if special education services are not needed. These rights are reinforced by the Americans with Disabilities Act (ADA).

Students who are protected by Section 504 also have individual accommodation plans. An important policy issue concerns the identification of appropriate accommodations in a student's plan. Accommodations should be identified on the basis of the student's instructional needs and the instructional practices regularly used with the student. Not all accommodations are appropriate for testing. If accommodations used in testing are not used in instruction, they may have detrimental effects on a student's test results.

Students who are learning English have been given many labels, including English language learners (ELLs), limited English proficient (LEP) students, non-English language background (NELB) students, and so on. Very often, the different terms have slightly different meanings. For discussion here, we will use the term English language learners.

English language learners are students whose first language is other than English and who are in the process of learning to speak and write in English. It is generally recognized that youngsters who are learning English first learn social language—that needed to converse with peers for example—and then learn academic English, the language used to impart the content of schooling. Although it varies by student and is complicated by whether the student learned academic content in his or her first language, it is estimated that it takes from three to five years to master social language and from five to seven (or more) years to master academic language (Collier, 1989).

Across the United States, the primary first language of English language learners is Spanish. In some locations, such as California and Texas, Spanish-speaking students comprise approximately one-fourth of the student population. In contrast, the most populous group of English language learners in some other locations, such as some urban areas in Minnesota, is Hmong and other Southeast Asian languages (e.g., Lao, Khmer).

The number of English language learners in American schools is difficult to estimate. However, the 1998 *Condition of Education* (U. S. Department of Education, 1998) indicates that the U.S. Hispanic school population will reach more than 20% by 2020, up from 9% in 1980. While not all of these students are English language learners, a significant percentage are, and they are added to students from more than 90 other language groups in U.S. schools today.

2. Why is it important to include special populations in assessments?

There are legal, philosophical, and practical reasons for including students with disabilities in state and district assessment systems. The legal reasons apply primarily to students with disabilities. They are embodied in the 1997 amendments to IDEA, where it specifies that states receiving federal funding for special education must include students with disabilities in their state and district assessments and report on both the number participating in the assessments and their performance. Also delineated in the amendments is the requirement that guidelines be developed for determining which assessment a student will participate in—the regular assessment or an alternate assessment developed for those students unable to participate in the regular assessment. IDEA also indicates that students with disabilities are to be provided accommodations during assessments, as appropriate, and that these accommodations are to be listed in each student's IEP.

Other laws also require that students with disabilities be provided needed accommodations in instruction and assessments. For example, Section 504 of the 1973 Rehabilitation Act

requires that accommodations be provided for students with disabilities even though they may not be receiving special education services. The Americans with Disabilities Act (ADA) also is commonly cited as a law that ensures individuals with disabilities the right to accommodations in assessment situations, including those encountered in educational settings, as well as the right to participate in assessments.

The legal basis for including students who are learning English is contained in legislation that provides Title I funds and in Title VI of the Civil Rights Act of 1964. The 1994 reauthorization of the Elementary and Secondary Education Act (ESEA), referred to as the Improving America's Schools Act, requires schools receiving Title I funds to include students with disabilities and students with limited English proficiency in their evaluations of performance and yearly progress. Furthermore, these evaluations are to be based on state assessments. Thus, by implication, these special needs students must be included in state assessments for schools to receive Title I funding.

The legal basis for including special needs students in large-scale assessments is supported by philosophical arguments about the importance of their inclusion. These arguments usually focus on the benefits derived from participation. Students who are included in assessments are included in instruction designed to help them meet the standards on which assessments are based. When they are not included, there is a tendency to not worry about whether they are learning the content included in testing. Furthermore, when education reforms are designed on the basis of how students perform on assessments, their exclusion from the assessments means that their needs are not being considered in all aspects of the reforms.

Practical reasons for including special populations in large-scale assessments focus on the consequences of their exclusion. Numerous unintended consequences occur when certain subsets of students are excluded from assessments. One of these is referrals to special education. When students with disabilities are not included in assessments, or their scores do not count, there is a tendency to refer to special education any child not expected to perform well. Similarly, if there is a policy of not including students who are learning English, then there will be a tendency to attribute poor performance in these students solely to their lack of English skills and therefore to exclude them from assessments. It is easy to continue to attribute lack of learning to inadequate English skills rather than to instructional deficiencies; and as long as these students are kept out of the assessment system, there is no need to face up to instructional issues.

The practical reasons often are the most convincing. As soon as there are ways for some students to be held out of the system, there is the opening for variability from one place to another. And, when this happens, comparisons become meaningless.

Critical Issues

Numerous questions arise as we consider the participation of special populations in assessment systems. Some of these questions, however, seem more critical than others. In this section, we address some of the more critical issues that arise as we consider the inclusion of special populations in assessment systems.

1. Participation of special populations in assessments

Special populations can be included in large-scale assessment systems in several ways. Typically, the options are organized into four alternatives: **standard assessment, partial assessment,** assessment with accommodations, and alternate assessment. The standard assessment option simply means that the assessment is administered to special populations in the same way as it is for all other students—without the use of accommodations or other adjustments (such as taking only parts of the test).

Partial assessment is another way special needs students might participate in assessments. This would be appropriate, perhaps, for a student who is just beginning to acquire English

skills but who is able to take a mathematics test that is language free (such as a test of computation). Partial assessment also might be an option for students with disabilities, particularly those with traumatic brain injury. For example, a student who has lost all numeracy skills as a result of a brain injury might still take all other parts of an assessment, but take an alternate assessment in the area of mathematics. While partial assessment is an option, it is probably needed by relatively few students.

Using accommodations while taking an assessment is another way special needs students can participate in assessments. Accommodations are changes in the way an assessment is administered, responded to, scheduled, or timed to allow the student's performance to better reflect his/her knowledge and skills, rather than the effects of the language barrier or disability. Students with disabilities and English language learners may use some of the same accommodations, as well as have separate sets of accommodations. (See next section for additional discussion of accommodations.)

A third way students can participate in assessment systems is to take an alternate assessment. What exactly the alternate assessment might consist of, as well as the characteristics of students who would participate, remains to be specified, as of 1999. Typically, this type of assessment is needed by only a small percentage—probably less that 2-3%—of the total school population in a state.

For students who are learning English, the alternate assessment might be a translated test; or it might be an assessment of English acquisition. The latter approach is highly controversial, however. Many educators consider it inappropriate to apparently put a hold on the student's acquisition of content while waiting for language skills to improve.

2. Accommodations for special populations in assessments

Examples of common accommodations for both groups of special needs students include allowing more time or extra breaks during testing, oral presentation of written directions, and individual or small group settings for test administration. For English language learners, additional accommodations include test translations and glossaries. For students with disabilities, additional accommodations can include those needed for individuals with sensory disabilities (e.g., Braille edition, signed directions, etc.) and others such as scheduling the test at a time best for the student, scribes to record responses, and answering in the test booklet rather than on a separate answer sheet. Examples of accommodations in use for English language learners and students with disabilities are presented in Table XIII-1.

The use of accommodations in general, and specific accommodations in particular, often creates controversy. Questions arise about the extent to which tests administered under accommodated conditions produce scores that are comparable to those produced under standard testing conditions. These questions are more emphatic for certain accommodations than for others. The most debated accommodations are those that seem to overlap with the

Table XIII-1. Examples of Accommodations for English Language Learners and Students with Disabilities

English Language Learners			
Setting	**Presentation**	**Response**	**Timing**
Separate room	Translation	Scribe	Extended time
Carrel	Directions read aloud	Point to response	Frequent breaks
Small group	Bilingual items	Oral answers	Multiple days
Students with Disabilities			
Setting	**Presentation**	**Response**	**Timing**
Separate room	Braille edition	Scribe	Extended time
Carrel	Directions read aloud	Point to response	Frequent breaks
Small group	Place markers	Mark in test booklet	Multiple days

constructs being tested. For example, reading a reading test to a student is rarely allowed in assessment systems, even if the purpose of the test is to evaluate the student's ability to understand written text. Similarly, having the student use a scribe to record the student's responses is allowed in some places but not others.

Accommodations are controversial, in part, because most tests were not standardized with them included. Determining the extent to which the use of accommodations changes the meaning of scores is very difficult. (See Thurlow, McGrew, Ysseldyke, Elliott, Thompson, & Phillips, 1999.) Several researchers now are conducting research on the effects of accommodations focused on this issue.

Deciding which assessment accommodations a student needs also appears to be difficult. To some extent, this difficulty may reflect the failure to make appropriate decisions about needed accommodations for a student during instruction. It is widely accepted that accommodations used during assessments should be those that are used during instruction, and that the accommodations are ones that the student needs. Determining student need may be difficult for some teachers and other educators. Researchers are developing procedures that may help teachers determine which accommodations students really need (Fuchs, Fuchs, Eaton, Hamlett, & Karns, in press).

The research currently underway on the effects of accommodations is just beginning to produce results. For example, Tindal and his colleagues (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998) have shown that there are no differences in performance when students use separate sheets to bubble in their answers, compared to marking on the test booklet. In contrast, the researchers found significant differences in the performance of students with disabilities when a math test was read to them by a trained teacher, compared to reading the test themselves. Analyses of existing state data also provide some evidence about the effects of accommodations. An important example here is the analysis of Kentucky's accommodation data which revealed that, for nearly all types of accommodations, the performance of students with disabilities remained below that of the total population (Trimble, 1998).

Similar work is underway on the effects of accommodations for English language learners. For example, Anderson, Jenkins, and Miller (1995) found that a bilingual mathematics exam in which students marked on just one version of each test question did not result in students using both versions of the test questions. In addition, a preliminary study by Liu, Anderson, Swierzbin, and Thurlow (1995) indicated that providing students with bilingual test items for an English reading passage did not necessarily affect their performance in a positive direction.

There is a clear need for research to continue, both on the effects of specific accommodations and on how to best make decisions about needed accommodations. These research efforts are underway.

3. Reporting the results from assessments of special needs students

In the past, states and districts rarely reported the performance of their English language learners or their students with disabilities. In fact, it was often the case that the scores of these students were systematically eliminated from the reports of assessment results. Sometimes the scores were aggregated and given to someone (e.g., the school building principals), but then discarded. An analysis of state reports in 1996 (Thurlow, Langenfeld, Nelson, Shin, & Coleman, 1998) produced only five states that had presented data for their students with disabilities; similar analyses for English language learners (Liu & Thurlow, 1999) indicated that eight states disaggregated their data from the data of all other students. In 1998, another analysis of state reports (Ysseldyke, Thurlow, Nelson, Teelucksingh, & Seyfarth, 1998) revealed that 13 states reported on the performance of students with disabilities in statewide assessments. While this number is a significant jump from just two years before, the federal government's intent was that this number would be close to 50 before July 1, 1999.

It is interesting that some states have suggested that reports of scores excluding students with disabilities or students who are English language learners are somehow more accurate than reports of scores that include those students. In fact, some states now report their scores both ways—with special needs students included and with them excluded.

4. Defining adequate performance: Single or double standard?

When high stakes consequences for students are implemented (as when the receipt of a standard diploma or passing from one grade to the next is based on performance on a test), some states have opted to provide adaptations for students with disabilities that seem to reflect a double standard. The most obvious double standard occurs when students with disabilities who are exempted from the test are still awarded standard diplomas or are allowed to move from one grade to the next. Some states have allowed students with disabilities to meet individually defined levels of performance as the criterion for receipt of a standard diploma.

These examples are different from the case where students with disabilities are allowed to take a test with accommodations, but must meet the same standards as other students. In this situation, the accommodation enables the student to show knowledge and skills without interference of the disability. To many, however, the possibility for crossing the line from an accommodation that does not change the meaning of a score to one that does is too great to be acceptable. It is because the effects of accommodations have been left to opinion, rather than research, that so much controversy continues. The critical issue of single or double standards usually returns one to the questions about the purpose of the test, the constructs being tested, and whether scores obtained in different ways can mean the same thing.

There are no easy solutions to the critical issue of what standards are being met when changes are made in the testing situation, at least regarding disabilities. On the other hand, most states and districts have determined that the distinction between the same standards and different standards is more clear for English language learners. There are many issues involved in acquiring language and in separating language skills from other skills. More research into assessment implications for those whose English skills are limited in one way or another needs to be undertaken, given the complexity of these issues.

Strategies for Maximizing the Inclusion of Special Populations in Assessments

The current literature about including students with disabilities in state and district assessments identifies several strategies that will maximize the participation of students with disabilities in assessments. A set of criteria for maximizing participation in assessments and accountability systems was developed by Elliott, Thurlow, and Ysseldyke (1996). Their criteria for participation, accommodations, and reporting are summarized in Table XIII-2.

While some of the same strategies apply to English language learners, different issues also arise. A comprehensive literature review of the needs of students with limited English skills produced several recommendations that promote greater participation of these students in large-scale assessments (Liu, Thurlow, Erickson, & Spicuzza, 1997). These are reproduced in Table XIII-3.

Table XIII-2. Criteria for Participation, Accommodations, and Reporting to Maximize Inclusion of Students with Disabilities in Assessments

Participation Criteria

1. Premise exists that all students, including all students with disabilities, are to participate in the district or state accountability system.
2. Decision about participation is made by a person (or group of people) who knows the student.
3. Form is used that lists the variables to consider in making participation decisions.
4. Reason(s) for exclusion are documented.
5. Student must participate in an assessment if the student receives *any* instruction on content assessed, regardless of where instruction occurs.
6. Decision about participation is *not* based on program setting, category of disability, or percent of time in the mainstream classroom.
7. Decision about participation allows for some students to participate in an alternate assessment or, when appropriate, in part of an assessment or assessment procedure.
8. Decision guidelines recognize that only a small percentage of students with disabilities need to participate in an alternate assessment (e.g., those with severe disabilities, about one to two percent of all students) or, when appropriate, to participate in a part of an assessment or assessment procedures.
9. Parents understand participation options and implications for their child **not** being included in an assessment or accountability system.
10. Decision about participation is documented on the student's IEP or on an additional form that is attached to the IEP.

Accommodations Criteria

1. Decision about accommodations is made by a person (or group of persons) who knows the student.
2. Decision about accommodations is based on the student's current level of functioning and learning characteristics.
3. Form is used that lists the variables to consider in making accommodations decisions and documents for each student the decision and reasons for it.
4. Accommodation guidelines require alignment of instructional accommodations and assessment accommodations.
5. Decision about accommodations is *not* based on program setting, category of disability, or percent of time in the mainstream classroom.
6. Decision about accommodations is documented on the student's IEP or on an additional form that is attached to the IEP.
7. Parents are informed about accommodation options and about the implications of their child (1) not being allowed to use needed accommodations, or (2) being excluded from the accountability system when certain accommodations are used.

Reporting Criteria

1. Written policy exists about who is included when calculating participation or exclusion rates.
2. Rates of exclusion that are specific to students with disabilities, and reasons for the exclusion, are reported when assessment results are reported.
3. Data reports include information from all test takers.
4. Records are kept so that data for students with disabilities could be reported separately, overall, or by other breakdowns.
5. Students keep records of the use of accommodations with disabilities, by type of accommodation, so that the information could be reported either by individual student or in aggregate.
6. Parents are informed about the reporting policy for their child's data.

Note: Criteria based on Elliott, Thurlow, & Ysseldyke (1996).

Table XIII-3. Ways to Promote Greater Participation of English Language Learners in Large-Scale Assessments

1. Use more than one source of data to make inclusion/exclusion decisions. Specify how decisions should be made and create a clear decision-making tree.
2. Include LEP students in assessments for accountability even when there is doubt about the student's ability to take them.
3. Collect data on excluded students and periodically reassess their eligibility to participate based on the data. There should be a time limit on exemption; a student should not be exempted indefinitely.
4. Use an alternative method to monitor exempted students' academic progress.
5. Avoid using accommodations for students with disabilities as the standard of comparison for the types of accommodations offered to LEP students.
6. Develop a range of allowable accommodations for students with differing proficiency levels. Consider accommodations received in the mainstream classroom.
7. Evaluate the effects of accommodations for students with limited English proficiency.
8. Develop scoring procedures that recognize the influence of English language learning and the tendency toward code switching (i.e., the use of two languages in the same response).
9. Disaggregate data by LEP status. If possible, also evaluate performance as a function of former LEP status and accommodations used.

Note: From an extensive literature review by Liu, Thurlow, Erickson, & Spicuzza (1997).

Summary/Conclusions

Tremendous change has occurred during the past five years regarding the recognition of issues surrounding the exclusion of students with disabilities, English language learners, and other special needs students from state and district assessment programs. The recognition of the unintended consequences of excluding these students has, in large part, pushed national education data collection programs (e.g., the National Assessment of Educational Progress) to reconsider their policies about the participation of such students in their assessments. These same forces have pushed lawmakers to require the inclusion of special needs students in assessments for the evaluation of Title I programs and for special education funding to states.

The unintended consequences of exclusion include increases in referrals to special education (even at grades where referrals typically are extremely low, such as during high school) and the retention of students in grades other than those in which tests are administered. In addition it has been documented that instruction is often minimized for students not included in assessments, especially at times of preparation for testing. And the design of reforms based on assessment results may not address some students' needs if they were not included when assessments were administered.

For all of these reasons a concerted national effort is now underway to increase the participation of students with special needs in assessments. Despite the compelling reasons for their inclusion, however, there are still many critical issues to address. Many of these issues have been highlighted in this chapter.

Chapter XIII Glossary

Accommodations. (1) Changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To be appropriate, assessment accommodations must be those also made during instruction and must not alter the construct intended to be measured or the meaning of the resulting scores. (2) Specific changes in testing conditions, procedures and/or formatting that do not alter the validity or reliability of a state standard. Policies and procedures must ensure that the accommodations do not compromise the security of the test and are consistent with the student's Individualized Educational Program (IEP), 504, and/or Limited English Proficient (LEP) plan. Accommodations can be made available for use in both instruction and statewide assessments. These may include accommodations for scheduling, setting, equipment, presentations, and/or responses. Allowable accommodations for states' assessments are generally identified in State Education Agency (SEA) documentation. (3) Alteration in *how* a test is presented to the test taker or in how a test taker is allowed to respond; includes a variety of alterations in presentation format, response format, setting in which the test is taken, scheduling or timing, and/or specialized equipment required by the student. The alterations do not substantially change level, content, or performance criteria. The changes are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate what is known. (4) Change in *how* a student accesses information and/or demonstrates learning; does not substantially change the content, instructional level, or performance expectations; provides for equal opportunity to demonstrate knowledge and skills.

Alternate assessments. An approach used in gathering information on the performance and progress of students whose disabilities preclude them from valid and reliable participation in typical state assessments as used with the majority of students who attend school. Under the re-authorized Individuals with Disabilities Education Act (IDEA), alternate assessments are to be used to measure the performance of a relatively small population of students who are unable to participate in the regular assessment system, even with **accommodations** or **modifications.**

English language learners. Students whose first language is other than English and who are in the process of learning to speak and write in English.

Large-scale. Assessments or programs that test or assess relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale programs are in contrast to tests and other assessments administered on a smaller scale, for example, by classroom teachers for instructional purposes.

Modifications. (1) Changes made in the content and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. (2) Changes in the administration of an assessment that may cause the construct being measured to differ from the construct as measured under standard administration conditions, or produce a score that means something different from scores yielded by the standard administration. Unlike *accommodations*, modifications may directly or indirectly compromise either the validity or reliability of the state standard. Modifications may compromise test security and therefore are not recommended for statewide assessments. Modifications are more appropriate for instruction and classroom tests and include a much wider range of supports and instructional scaffolding than do accommodations. Modifications can be identified on the student's IEP, 504, and/or LEP plan. Modifications can be effectively used in combination with accommodations in instructional and assessment situations when individualized to the student's strengths and needs. (3) Changes in what a student is expected to learn, such as changes in content, instructional level, and/or performance expectations. The intent of modifications is to allow for meaningful participation and enhanced learning.

Partial assessment. The administration of certain parts of an assessment. It is one way that special needs students might participate in assessments. This would be appropriate, perhaps, for a student who is just beginning to acquire English skills but who is able to take a mathematics test that is language free (such as a test of computation). Partial assessment also might be an option for students with disabilities, particularly those with traumatic brain injury. For example, a student who has lost all numeracy skills as a result of a brain injury might still take all other parts of an assessment, yet take an alternate assessment in the area of mathematics. While partial assessment is an option, it is probably needed by relatively few students.

Performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing of a musical composition, and portfolios showing samples of work over time.

Standard assessment. The administration of an assessment in the prescribed, standard way, without the use of **accommodations** or **modifications.**

Students with disabilities. Students with physical, sensory, cognitive, behavioral, or learning limitations. Although many students with disabilities receive special education services via an Individualized Education Program (IEP), some need only a 504 accommodation plan to access instruction and assessments. Some students with disabilities do not need either an IEP or an accommodation plan.

Chapter XIII References and Resources

Anderson, N., Jenkins, F., Miller, K. (1995). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL quarterly, 23*(3), pp. 509-525.

Elliott, J., Thurlow, M., & Ysseldyke, J. (1996). *Assessment guidelines that maximize the participation of students with disabilities in large-scale assessments: Characteristics and considerations.* (Synthesis Report 25). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Fuchs, L., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of test accommodations with objective data sources. *School Psychology Review, 29,* 65-85.

Liu, K. K., Anderson, M. E., Swierzbin, B., & Thurlow, M. L. (1999, April). *Bilingual accommodations for LEP students on statewide reading tests.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Liu, K. K., & Thurlow, M. L. (1999). *What state accountability reports are saying about students with limited English proficiency.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Liu, K. K., Thurlow, M. L., Erickson, R., & Spicuzza, R. (1997). *A review of the literature on students with limited English proficiency and assessment* (State Assessment Series Minnesota Report No. 11). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Langenfeld, K. H., Nelson, J. R., Shin, H., & Coleman, J. E. (1998). *State accountability reports: What are states saying about students with disabilities?* (Synthesis Report No. 20). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., McGrew, K., Ysseldyke, J., Elliott, J., Thompson, S., & Phillips, S. (1999). *Accommodations research: Design and analysis considerations.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities in large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment: Students with disabilities in the KIRIS on-demand assessments from 1992-93 through 1995-96* (State Assessment Series Maryland/Kentucky Report No. 3). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U. S. Department of Education. (1998). *Condition of education 1998.* Washington, DC: National Center for Education Statistics.

U. S. Department of Education. (2000). *State ESEA Title I participation information for 1996-97: Summary report.* Washington, DC: Office of the Under Secretary, Office of Elementary and Secondary Education (Doc. #2000-01).

Ysseldyke, J., Thurlow, M., Nelson, R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report No. 23). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Appendix
Test Preparation

Test Preparation

In an evaluation study involving primary-aged children, Pike (1973) identified the following skills as improving the performance of students below third grade.

1. Filling in an oval shaped space with a pencil, i.e., "bubbling."
2. Filling in only one such space per test item.
3. Filling an oval shaped space under the picture of an object.
4. Filling in ("marking") the space under only one object when presented with three objects per test item.
5. Marking the space beside the word that identifies the picture of an object.
6. Marking the space beside the word that identifies a pictured object given four response options (words).
7. Marking the space beside a dictated word given four response options but no pictured object.
8. Drawing a ring around several pictured objects.
9. Writing a number in a box (square).
10. Writing the number of several pictured objects in a box.
11. Following the left-to-right and top-to-bottom test item sequencing format.
12. Listening to and following directions.
13. Finding identifiers and keeping in sequence with the teacher during dictated portions of the tests.
14. Working independently on a timed task for the duration of the time required.
15. Treating test items independently of personal experience when the keyed answer is dependent only on presented information.
16. Completing the task in the allotted time.
17. Marking an answer even when not certain it is correct.

Sabers (1975) offers the following tips for discussing "test wiseness" with children:

1. The instructional program should never be sacrificed to teach test-taking skills.
2. Discuss with students how to approach particular problems. For example, some students perform better when they read the reading comprehension questions before they read the passage. This strategy can be especially useful when the problem involves maps and graphs as there is usually much more information than is needed for answering the questions.
3. Throughout the course of day-to-day instruction, discuss why wrong answers are wrong.
4. If students will be taking multiple-choice tests, provide some practice items that have one clearly incorrect answer, one possible answer, and two answers that may seem correct, but one is more correct than the other. This can help students see how they can make answer choices by eliminating some of the options.
5. Help students see tests as a natural extension of learning. Encourage them to make up their own tests.
6. The research on changing answers is inconclusive. It appears that when children are told not to change answers but do it anyway, their changes are for the better. However, when told to change answers, they make indiscriminate changes which lower their scores. The best advice appears to be "change only when you have a good reason to change."

7. To provide students with practice in a "testing atmosphere" environment, some teachers have a quiet time each day. The rationale is that it makes little sense to have it noisy in the room every day but test day.
8. Students must gradually increase the amount of time they are able to work independently up to the amount of time they will be expected to work independently on the test.
9. If the test will be administered by someone other than the teacher, it can be helpful to have an "outsider" administer a practice test prior to the big test day.
10. Children should be taught strategies for "sitting quietly" while other students finish their tests.
11. Children need exposure to test items that are contrary to present life's experiences. An example might be a Thanksgiving story presented in July. A question such as "What holiday is Jessica preparing for?" requires the child to think "Thanksgiving" when the classroom looks like the Fourth of July. Practice items might present three realistic options, only one of which is correct.
12. If possible, read the instructions on using a test, the directions for administration, and study old tests to see what is being required of the students. Examine students' completed classroom tests to see what kinds of mistakes they are making.
13. When developing practice exercises, it can be instructive for teachers to try out exercises on each other and collaborate in the development of instructional strategies.
14. Separate answer sheets should be introduced to students long before they encounter them in the regular testing program.
15. Avoid creating a pattern of correct responses that may ultimately lower children's scores. For example, do not use the same option as the right answer consistently. According to Sabers, most teachers key "b" or the second response as the correct option; hence, students tend to mark that option. Students need to learn that what is keyed depends on the question being asked.
16. Children should be exposed to different kinds of practice exercises over a long period of time. The skills cannot be taught in one sitting.

The Utah State Office of Education provides the following test-taking tips and strategies to students, parents, teachers, and administrators (Utah State Office of Education, 1999a and 1999b).

Help Students Prepare for Tests (from *Understanding Tests—A Guide for Parents.* The Jordan School District)

1. Take an interest in standardized achievement tests, but don't be so concerned that you make your child nervous.
2. Talk about the tests as "opportunities to show what has been learned." Explain that the tests are not competitions or tests where students pass or fail.
3. Encourage your child to listen to ALL instructions and follow test directions carefully.
4. Remind your child that it is OK to ask questions if the instructions don't seem clear.
5. Be positive and express confidence that your child will be able to handle the tests well.
6. Urge your child to do the best work possible but to also keep in mind that test results are only one way to show how they are doing in school.
7. Show interest in your child's school work every day, not just on test days.

Helping Your Child Prepare for Test Day

The night before:

- Help your child get to bed on time. Research shows that being well-rested helps students do better.
- Help children resolve immediate arguments before going to bed.
- Keep your routine as normal as possible. Upsetting natural routines may make children feel insecure.
- Mention the test to show you're interested but don't dwell on it.
- Plan ahead to avoid conflicts the morning of the test.

The morning of test day:

- Get up early enough to avoid rushing. Be sure to have your child at school on time.
- Have your child eat a good breakfast but not a heavy one. Research shows that students do better if they have breakfast before they take tests.
- Have your child dress in something comfortable.
- Be positive about the test. Acknowledge that tests can be hard and that they're designed so that no one will know all the answers. Explain that doing your best is what counts. The important thing is to make your child comfortable and confident about the test.

After the test:

- Talk to your child about his or her feelings about the test, making sure you acknowledge the effort such a task requires.
- Discuss what was easy and what was hard; discuss what your child learned from the test.
- Discuss what changes your child would make if he or she were to retake the test.
- Explain that performance on a test is not a condition for you to love your child. You love your child just for the person he or she is.

Students: Preparing to Take a Multiple-Choice Test

General suggestions:

1. **Ask for help.** If you are ever taking a test and you don't understand the directions, ask the person giving the test for help. It is very important that you understand what the test-makers are asking you to do.
2. **Time yourself.** Don't spend too much time on any one answer. Do your best and then move on. Skip things you can't answer because you can always come back. Also, your mind might keep working on the problem you couldn't answer while you're doing other problems. (Remember, if wrong answers don't count against you, be sure to try to answer every question.)
3. **Do NOT change answers.** On multiple-choice tests, do not change your answers unless you are very uncertain about your first thought. Your first guess is usually right unless you are sure you have answered incorrectly.

4. **Know if wrong answers count against you.** Knowing whether or not you get points subtracted for wrong answers makes a difference in how you take the test.

   If the test counts wrong answers against you:
   - Leave blanks if you don't know the answer.
   - Don't guess.

   If the test does not count wrong answers against you:
   - Make the most intelligent guess you can.
   - Answer all the questions.

Tips for answering test questions

Multiple-choice questions ask a question and give you lots of choices about which is the best answer. If you're not used to multiple-choice questions, you may be asked to do things in ways you've never done them, and that could make taking the tests seem harder than it should be for you. Here are some suggestions to help.

Answering various types of questions

- **Always** read the entire question with all of the possible answers before choosing an answer.
- Identify key words or phrases in the question that will help you choose the correct answer—e.g., "what is the **best** answer," "select the answer that is most correct."
- Check yourself to make sure you understand what the question is asking—be sure you are responding to the question that is being asked.
- When there are several questions about a reading passage or chart, look for clues in other questions that will help you with those items about which you are unsure.
- If the test requires you to read passages and then answer questions about what you read, read the questions first. By doing this, you will know what you are looking for as you read. This also helps you go faster on the test.

Managing time wisely

Many times, multiple-choice tests give you a limited amount of time to complete each section of the test. It is important to know if the test you are taking has time limits so you can make the best possible use of your time. If the test does have a time limit, follow the steps below:

1. Glance through the entire section of the test to familiarize yourself with how it is laid out before beginning to answer the questions.
2. Next, go through the section, question by question. If you're **sure** you know the answer to a question, mark it immediately and go on. If you're **unsure** of an answer, skip the question. **Do not spend extra time on questions you can't answer.**
3. Once you have been through all of the questions once, go back and find questions you have some knowledge about and use your "partial knowledge" to eliminate one or two response choices. Then choose between the remaining responses. If you can eliminate two wrong answers, your chance of guessing the right answer is greater.
4. Now, if any time is remaining, spend it on those questions about which you know nothing or almost nothing. Use your logic and general knowledge to help you make the best choice. (**NOTE:** As you go back through, do not change answers unless you are

very uncertain about your first thought. Your first guess is usually right unless you are **sure** you have answered incorrectly.)

Ways of Overcoming Test Anxiety

- Learn and practice strategies such as those mentioned previously.
- Take advantage of the opportunity to take practice tests to familiarize yourself with the test format.
- When taking practice tests, simulate actual testing conditions as much as possible—e.g., follow the time limit, use an answer sheet, do only one section at a time, practice test-taking strategies.
- Learn and practice some basic relaxation techniques such as imagining yourself in a relaxing place you have been, a place where you feel unhurried, peaceful, and calm.
- Think about and write any other things you could do to relax when you're nervous or anxious during a test, then try them out while working on practice tests.

Appendix References

Pike, E. O., Jr. (1973). *Influence of a test-taking skills instructional program on the Metropolitan Achievement Tests performance of children from low-income families.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Sabers, D. L. (1975). *Test-taking skills.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Utah State Office of Education. (1999a). *Test-taking tips and strategies: Student/parent pamphlet.* Salt Lake City, UT: Utah State Office of Education.

Utah State Office of Education. (1999b). *Test-taking tips and strategies: Teacher/administrator guidelines.* Salt Lake City, UT: Utah State Office of Education.

Comprehensive Glossary

Accessibility. The extent to which the content, format, and response mode options of an assessment make it possible for *all* students, including students who have disabilities or limited English proficiency, to participate in an assessment.

Accommodations. (1) Changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To be appropriate, assessment accommodations must be those also made during instruction and must not alter the construct intended to be measured or the meaning of the resulting scores. (2) Specific changes in testing conditions, procedures and/or formatting that do not alter the validity or reliability of a state standard. Policies and procedures must ensure that the accommodations do not compromise the security of the test and are consistent with the student's Individualized Educational Program (IEP), 504, and/or Limited English Proficient (LEP) plan. Accommodations can be made available for use in both instruction and statewide assessments. These may include accommodations for scheduling, setting, equipment, presentations, and/or.responses. Allowable accommodations for states' assessments are generally identified in State Education Agency (SEA) documentation. (3) Alteration in *how* a test is presented to the test taker or in how a test taker is allowed to respond; includes a variety of alterations in presentation format, response format, setting in which the test is taken, scheduling or timing and/or specialized equipment required by the student. The alterations do not substantially change level, content, or performance criteria. The changes are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate what is known. (4) Change in *how* a student accesses information and/or demonstrates learning; does not substantially change the content, instructional level, or performance expectations; provides for equal opportunity to demonstrate knowledge and skills.

Accountability. The systematic use of assessment data and other information to assure those inside and outside of the educational system that schools are moving in desired directions. Commonly included elements are goals, indicators of progress toward meeting those goals, analysis of data, reporting procedures, and consequences or sanctions. Accountability often includes the use of assessment results and other data to determine program effectiveness and to make decisions about resources, rewards, and consequences.

Aggregated scores. The total or combined performance for all individuals or groups on one test or subtest. For example, a state average usually represents the aggregation of scores for all students/groups of students who took the test.

Alignment. The similarity or match between and among the content standards, performance standards, curriculum, instruction, and assessments in terms of knowledge and skill expectations. The inferences made on the basis of assessment results are valid only to the extent that the system components are aligned. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. This standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Alternate assessments. An approach used in gathering information on the performance and progress of students whose disabilities preclude them from valid and reliable participation in typical state assessments as used with the majority of students who attend school. Under the re-authorized Individuals with Disabilities Education Act (IDEA, 1997), alternate assessments are to be used to measure the performance of a relatively small population of students who are unable to participate in the regular assessment system, even with **accommodations** or **modifications.**

Alternate forms reliability. "Alternate forms" is a generic terms referring to two or more versions of a test that are considered interchangeable in that they measure the same constructs, are intended for the same purposes, and are administered using the same directions. Alternate forms are reliable to the extent that the scores of every individual hold their ranks in a score distribution from one alternate form to another.

Assessment. Any systematic method of obtaining evidence from tests and other sources that is used to draw inferences about characteristics of people, objects, or programs for a specific purpose.

Assessment system. An aligned assessment system is a series of assessments of student performance at different grade levels that are based on publicly adopted standards of what is to be taught, coupled with high expectations of student mastery. This standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards.

Baseline data. The initial measures of performance against which future measures will be compared.

Benchmarks. Specific statements of knowledge and skills to be demonstrated at the end of a specified range of grades. For example, benchmark content standards may be set at the end of grades 4, grade 8, and grade 12 to specify standards to be met by the end of primary, middle, and high school grade ranges. Benchmarks are located on a performance continuum and are used as checkpoints to monitor progress from one level to the next.

Bias. In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

Breadth. The comprehensiveness of the content and skills embodied in the standards, curriculum, and assessments.

Cohorts. In educational research, generally, groups of students who cannot necessarily be compared to themselves over time. This is usually due to attrition such as moving away or dropping out of school. Examples of cohort studies include comparing groups of different students at the same grade level over time or comparing scores from the same group over time, even though some group members may change.

College admissions test scores. Scores yielded by tests used in college admissions decisions. Two commonly used college admissions tests are the ACT and the SAT. In 1998, the national average ACT scale score was 21. The possible range of ACT scale scores is from 1 (low) to 36 (high). The standard scores for the SAT have a mean of roughly 500 and a standard deviation of roughly 100.

Consequential validity evidence. Data that illuminates the extent to which the assessment has the desired effects, e.g., on students, teachers, administrators, the curriculum, instruction and/or other entities.

Construct. The underlying theoretical concept or characteristic a test is designed to measure.

Construct validity evidence. Data that illuminate the extent to which a test produces results that accurately reflect the construct they are designed to assess.

Constructed-response. Items that require students to create their own responses or products rather than choose a response from an enumerated set.

Content standards. Statements of the knowledge and skills schools are expected to teach and students are expected to learn. They indicate what students should know and be able to do as a function of schooling.

Content validity evidence. Data that illuminate the extent to which (1) the knowledge, skills, and cognitive demands of the learning objectives underlying an assessment are accurately reflected in the assessment; and (2) the assessment adequately covers the **domain** of knowledge, skills, and cognitive demands represented in the learning objectives.

Convergent validity evidence. Data showing the degree to which the assessment results are positively **correlated** with the results of other measures designed to assess the same or similar **constructs.**

Correlate. In the statistical sense, two sets of scores are perfectly **correlated** if every individual has the same score rank on one measure as on the other. For example, if the top-scoring individual on measure 1 also obtains the top score on measure 2, the second-best individual on measure is the second-best on measure 2, etc., then the measures are perfectly correlated, i.e., the **correlation coefficient** is +1.00. To the extent that the individuals in the score distributions do not maintain their ranks, the correlation coefficient is reduced.

Correlated. When the relationship between the ranks of individuals in different score distributions are statistically examined and described via a correlation coefficient, they are said to be correlated. The correlation coefficient can range from a perfectly negative relationship (-1.00), meaning that the top scoring individual on one measure is the lowest scoring individual on the other measure, etc. to a perfectly positive relationship (1.00), meaning that the top scoring individual on one measure is the top scoring individual on the second measure, the individual with the second best score on one measure has the second best score on the second measure, etc.

Correlation coefficient. The statistical representation of the relationship between two, or more sets of scores. See **correlate** and **correlated.**

Criterion-referenced. The reference point for interpreting test results using a criterion that indicates a particular level of achievement. The criterion may be a predetermined number of correct responses or, in the case of performance tasks, a response that meets certain criteria for competent performance, e.g., the proper use of conventions and logical, supporting ideas for a point of view in writing. Criterion-referenced tests allow users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to a norm or the performance of others.

Criterion validity evidence. The extent to which there is evidence showing that scores on a test are related to a criterion measure. For example, if a test is intended to measure what is learned in a particular course of study, then the test scores and course grades should **correlate.**

Cross-sectional studies. Comparison of different groups of individuals over time, e.g., the results obtained by a group of fifth-grade students on a standardized mathematics test in one year compared to the results obtained by a different group of fifth-grade students on the same test in another year. This kind of analysis is commonly used to track the progress of a school, district, state, or nation over time.

Curricular validity evidence. The extent to which there is evidence that students are taught a curriculum that aligns with the assessments and the learning objectives or content standards on which the assessments are based.

Curriculum. What is taught.

Customized assessments. Assessments that are customized or tailor-built to meet a particular need. Usually they are developed to cover a particular set of content standards.

Cut score. A specified point on a score scale at which scores above that point are interpreted differently from scores below that point. Sometimes there is only one cut score, dividing the range of possible scores into "passing" and "failing" or "mastery" and "nonmastery." Sometimes two or more cut scores may be used to define three or more score categories, as in establishing performance standards.

Defensibility. The technical properties of an assessment that make its use for a particular purpose appropriate. Such properties include validity, reliability, fairness, and lack of bias.

Depth. The taxonomic level of cognitive processing required for success relative to the performance standards, e.g., recognition, recall, problem solving, analysis, synthesis, evaluation.

Derived scores or scaled scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

Discriminant validity evidence. Data that show the results of an assessment do not correlate highly with the results of assessments designed to measure a different, but related, construct, e.g., achievement versus ability.

Domain. The portion of all knowledge and skill in a subject matter area that is selected for the content standards once consensus is reached that it represents what is important for teachers to teach and students to learn.

Embargoed. Test results prohibited from being released until a specified date/time.

English language learners. Students whose first language is other than English and who are in the process of learning to speak and write in English.

Equated. Two or more forms of a test that yield equivalent or parallel scores for specified groups of test takers. Equating involves converting the score scale of one form of test to the score scale of another form so that the scores are equivalent or parallel.

Errors of measurement. The differences between observed scores and the theoretical true score; the amount of uncertainty in reporting scores; the degree of imprecision that may result from the measurement process (e.g., test content, administration, scoring, or examinee conditions), thereby producing errors in the interpretation of student achievement.

Face validity evidence. Tests that measure what they purport to measure. For example, a "writing" test that relies solely on multiple-choice questions about the conventions of writing such as grammar, punctuation, and spelling, is lacking in face validity.

Fair tests. Yield student scores that are not influenced by such irrelevant factors as native language, prior experience, gender, or race.

Field test. A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a **pilot test.**

Generalizability theory. Contributes to reliability by allowing test developers to estimate the amount of error in students' test scores from different sources (e.g., raters, items, testing occasions).

Grade-equivalent score. Represents a performance level that is typical of students in a particular grade at a particular time of year. In a statistical sense, it is the school grade level for which a given score is the real or estimated median or mean.

Guidelines. Information and the description of procedures that can be used by local school districts in implementing state board policies.

High-stakes. Tests whose results have important, direct, or lasting consequences for examinees, programs, or institutions.

Instruction. The teaching methods used to deliver the curriculum to students.

Internal consistency. The degree to which the test items, on average, **correlate** with the entire test. It is a measure of the extent to which a group of items contribute to measuring the **construct** measured by the test.

Inter-rater reliability. The degree to which different scorers agree on the score to be assigned to a test response.

Intra-rater reliability. The degree to which an individual rater is consistent over time.

Item samples. Subsets of a larger array of test items. Item samples must be sufficiently large to represent the full array of items.

Large-scale. Assessments or programs that test or assess relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale programs are in contrast to tests and other assessments administered on a smaller scale, for example, by classroom teachers for instructional purposes.

Laws. Legislative mandates that carry negative legal consequences when violated.

Legislation. The result of lawmaking activity; law.

Longitudinal studies. Comparison of the same individual's results over time. In such studies, care must be taken that the measures used are also reliable over time. Groups

may be studied longitudinally, provided that the individuals within the group remain the same, i.e., there are no "dropouts" and there are no new members.

Matrix sampling. A measurement technique whereby a large set of test items is organized into a number of relatively short item sets. Each subset is then administered to a subsample of test takers, thereby avoiding the need to administer all items to all examinees, e.g., for program evaluation purposes.

Modifications. (1) Changes made in the content and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. (2) Changes in the administration of an assessment that may cause the construct being measured to differ from the construct as measured under standard administration conditions, or produce a score that means something different from scores yielded by the standard administration. Unlike *accommodations*, modifications may directly or indirectly compromise either the validity or reliability of the state standard. Modifications may compromise test security and therefore are not recommended for statewide assessments. Modifications are more appropriate for instruction and classroom tests and include a much wider range of supports and instructional scaffolding than do accommodations. Modifications can be identified on the student's IEP, 504, and/or LEP plan. Modifications can be effectively used in combination with accommodations in instructional and assessment situations when individualized to the student's strengths and needs. (3) Changes in what a student is expected to learn, such as changes in content, instructional level, and/or performance expectations. The intent of modifications is to allow for meaningful participation and enhanced learning.

Norm. Typical or average performance. The norm does not necessarily represent the most desirable performance.

Norm group. The group used to establish "typical" or average performance on a particular test. Typical performance is not necessarily ideal performance.

Norm-referenced. Test interpretations whose scores are based on a comparison of a test taker's performance to the performance of other people in a specified **reference population.**

Off-the-shelf tests. "Ready made," commercially available tests that can be purchased "as is" from a test publisher or vendor.

Parallel forms reliability. Parallel forms are a type of **alternate form** that have equal raw score means, equal standard deviations, and equal **correlations** with other measures for any given population. Parallel forms are reliable to the degree to which scores of every individual hold their ranks in the score distribution from one parallel form to another.

Parallel tests. Also called alternate test forms; two or more versions of a test considered to be interchangeable in that they measure the same **constructs,** are intended for the same purposes, are administered using the same directions, and yield comparable scores.

Partial assessment. The administration of certain parts of an assessment. It is one way that special needs students might participate in assessments. This would be appropriate, perhaps, for a student who is just beginning to acquire English skills but who is able to take a mathematics test that is language free (such as a test of computation). Partial assessment also might be an option for students with disabilities, particularly those with traumatic brain injury. For example, a student who has lost all numeracy skills as a result of a brain injury might still take all other parts of an assessment, yet take an alternate assessment in the area of mathematics. While partial assessment is an option, it is probably needed by relatively few students.

Percentile. The score on a test below which a given percentage of scores fall.

Percentile rank. The percentage of scores in a specified distribution that fall below the point at which a given score lies.

Performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, constructed-response items such as mathematics

problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Performance-based or performance assessments. Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing of a musical composition, and portfolios showing samples of work over time.

Performance standards. Specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components: (1) performance levels that provide descriptive labels for student performance, e.g., advanced, proficient, basic; (2) descriptions of what students at each performance level must demonstrate relative to the test; (3) examples of student work that illustrate the range of performance for each performance level; and (4) cut scores that separate one level of performance from another.

Performance tasks. A type of test item that is product- or behavior-based. They are designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied. Examples of commonly used performance assessment formats include writing exercises such as essays, open-ended items such as mathematics problems that require students to show their work, demonstrations such as conducting a laboratory experiment or playing a musical composition, and portfolios showing samples of work over time.

Pilot test. A test administered of a test to a representative sample of test takers solely for the purpose of determining the properties of the test. See **field test.**

Policies. Procedures for implementing laws.

Portfolios/portfolio assessment. (1) Systematic collections of education or work products that are typically collected over time. (2) A collection of student-generated or student-focused products that provide the basis for judging student accomplishment. In school settings, portfolios may contain extended projects, drafts of student work, teacher comments and evaluations, assessment results, and self-evaluations. The products typically depict the range of skills the student has, or reveal the improvement in a student's skill level over time. Salvia & Ysseldyke (1995) list six elements that typically are said to characterize portfolio assessment: (1) They target valued outcomes for assessment (generally those that require higher levels of understanding such as analysis, synthesis, and evaluation; those that require applying specific processes or strategies to reach answers; and those that are complex and challenging). (2) They use tasks that mirror work in the real world, i.e., that are *authentic*. (3) They encourage cooperation among learners and between teacher and student. (4) They use multiple dimensions to evaluate student work. (5) They encourage student reflections. (6) They integrate assessment and instruction.

Random sampling. The selection of a **sample** according to a random process, with the selection of each entity in no way dependent on the selection of other entities.

Raw score. The number of items correct.

Reference group. The group of test takers to which a particular test score will be compared.

Reference population. The population of test takers represented by a test's norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or other characteristics at the time of testing.

Reliable. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and, hence, are dependable and repeatable; the degree to which scores are free of **errors of measurement**

Reliability. The degree to which the scores of every individual are consistent over repeated applications of a measurement procedure and, hence, are dependable and repeatable; the degree to which scores are free of **errors of measurement.**

Reliability coefficient. A unit-free index that reflects the degree to which scores are free of **errors of measurement.**

Rubrics. Scoring guides for constructed-response questions or performance tasks. Scoring rubrics contain a description of the requirements for varying degrees of success in responding to the question or performing the task.

Sample. A specified number of entities—called sampling units (test takers, items, etc.)—selected from a larger specified set of possible entities, called the population.

Sampling. The selection of a **sample.**

SAT (Scholastic Assessment Tests). Tests developed by Educational Testing Service and administered by the College Entrance Examination Board. Results of the SAT are used by numerous colleges and universities in making decisions about student admission.

Scaled scores or derived scores. Scores to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

School report cards. Reports that provide information about schools, as a whole, rather than about individual students. For example, they may include information about the number of students who score at the proficient level on state tests; information about the number of teachers teaching in their areas of primary training; as well as information about attendance, retention, and discipline referrals. In some cases, data on school report cards are used to make programmatic decisions about schools or to determine whether they meet accreditation criteria.

Scorer reliability. The degree to which the scores assigned by raters are consistent over repeated applications of the scoring rubric. **Inter-rater reliability** and **intra-rater reliability** are types of scorer reliability.

Selected-response. A test item that requires students to select an answer from a list of given options. A common selected-response format is the multiple-choice item.

Stakeholders. Persons holding a vested interest in the outcomes of the assessment program. These likely include parents, students, educators, and taxpayers.

Standard assessment. The administration of an assessment in the prescribed, standard way, without the use of **accommodations** or **modifications.**

Standard deviation. The average amount that scores in a distribution of scores deviate (differ) on either side of the mean.

Standard error of measurement. The average amount that scores in a distribution differ from the corresponding **true scores** for a specified group of test takers.

Standard scores. A type of derived *score* such that the distribution of these *scores* for a specified population has convenient, known values for the mean and standard deviation.

Standardized tests. Tests administered and scored in a uniform manner from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Standard-setting group. A standards-setting group is used to inform or establish desired or proficient levels of performance on a particular test. Typical performance is not necessarily ideal performance.

Standards-based systems of assessment. Include criterion-referenced test. In such systems, test items reflect a pre-established set of **content standards** that specify the knowledge and skills students are expected to acquire as a function of schooling. Results are then interpreted against a set of criteria or **performance standards** that define student performance relative to the content standards represented by the test items.

Stanine scores or "standard nine" scores. Derived scores that range from 1-9. Stanine scores of 4, 5, and 6 are in the middle and, hence, are considered in the average range. Stanine scores of 1-3 are generally considered below average and stanine scores of 7-9 are generally considered to be above average.

Students with disabilities. Students with physical, sensory, cognitive, behavioral, or learning limitations. Although many students with disabilities receive special education services via an Individualized Education Program (IEP), some need only a 504 accommodation plan to access insstruction and assessments. Some students with disabilities do not need either an IEP or an accommodation plan.

System of assessment. Consists of complementary components which, together, provide an accurate profile of student achievement.

Teaching the test. Teaching students the actual, or nearly identical, items that will appear on a test. Not only does such practice constitute cheating, it confines instruction to a mere sample of the knowledge and skill domain represented by the test.

Teaching to a test. Teaching the broad-based knowledge and skills represented by a test's underlying content standards. Compared to **teaching the test,** it is not cheating.

Technically sound. Defensible assessments; they are reliable (consistent in their measurement and in the application of scoring procedures), valid for the purposes for which the results will be used, and are fair and unbiased.

Test blueprints. Written documents, often in chart form, that detail the number of questions to be included on a test, the item formats, and the content and skills that each set of items will assess. In the case of standards-based tests, it is important for the test blueprints to consider the performance standards as well as the content standards so that items cover the intended depth as well as breadth of the standards. In addition to guiding test development, test blueprints can be useful in preparing to take an examination.

Test forms. Parallel or alternate versions of a test that are considered interchangeable in that they measure the same **constructs,** are intended for the same purposes, and are administered using the same directions.

Test. In contrast to **assessment**, a test that includes a number of measures that help create a more complete picture or profile of performance, is usually a single instrument or procedure such as a quiz, standardized measure, questionnaire, survey, observation, checklist, and the like. Thus, tests are typical components of aligned systems of assessment.

Test-retest reliability. The extent to which the scores of every individual hold their ranks in the score distribution upon repeated administrations of the same test to the same individuals.

Test security. The need to keep tests safeguarded so all students have equal exposure to the test materials and equal opportunities for success. If test security is violated, then some students can be placed at an unfair advantage or disadvantage. When this happens, the validity of tests is violated.

Test specifications. Sometimes used interchangeably with **test blueprints.** Test specifications provide a framework that specifies the proportion of items that assess each content and process/skill area; as well as the format of items, responses, and scoring protocols and procedures. These frameworks additionally specify the desired psychometric properties of the test and test items, such as the distribution of item difficulty and discrimination indices.

True scores. In classical test theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In item response theory, the error-free value of test taker proficiency.

Valid. The degree to which a test measures what it purports to measure. See **Validity.**

Validity. (1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. (2) The extent to which a test measures what its authors or users claim it measures. (3) The appropriateness of the inferences that can be made on the basis of test results.

Writing prompts. Phrases or sentences designed to elicit written responses. In the primary grades they make take the form of story-starters. In later grades they may ask students to write an essay on a particular topic, often specifying a particular mode (e.g., persuasive, descriptive).

Comprehensive List of References and Resources

Almond, P. (1999). *A single assessment system for all students including those with special challenges—disabilities, limited English fluency, and poverty: What will it take?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

American Psychological Association. (1997). *Ethical standards for psychologists.* Washington, DC: Author.

Anderson, N., Jenkins, F., & Miller, K. (1995). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

Arter, J. (1993, April). *Designing scoring rubrics for performance assessments: The heart of the matter.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED 358 143)

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. *Review of Educational Research, 53,* 571-585.

Blank, R., Manise, J., & Brathwaite, B. C. (1999). *State education indicators with a focus on Title I.* Washington, DC: Council of Chief State School Officers.

Brown, F. G. (1984). *Guidelines for test use: A commentary on the standards for educational and psychological tests.* Washington, DC: National Council on Measurement in Education.

Byrd, M. (1987). *A comparison of the effectiveness of four test preparation programs* (Final Evaluation Report). Chicago: Chicago Public Schools, Department of Research and Evaluation.

Cochran, W. G. (1953). *Sampling techniques.* New York: John Wiley & Sons, Inc.

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL quarterly, 23*(3), pp. 509-525.

Council of Chief State School Officers. (1999). *Trends in state student assessment programs: Fall 1997.* Washington, DC: Author.

Creech, J. (1998). *Annual benchmarks report.* Atlanta, GA: Southern Regional Education Board.

Deaton, W. L., Halpin, G., & Alford, T. (1987). Coaching effects on California Achievement Test scores in elementary grades. *Journal of Educational Research, 80,* 149-155.

Ebel, R. L. (1980). *Practical problems in educational measurement.* Lexington, MA: D.C. Heath.

Elliott, J., Thurlow, M., & Ysseldyke, J. (1996). *Assessment guidelines that maximize the participation of students with disabilities in large-scale assessments: Characteristics and considerations* (Synthesis Report No. 25). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Fuchs, L., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of test accommodations with objective data sources. *School Psychology Review, 29,* 65-85.

Gardner, D. P., Larsen, Y. W., Baker, W. O., & Campbell, A. (1983). *A nation at risk: The imperative for educational reform.* An open letter to the American people. A report to the nation and the Secretary of Education. Washington, DC: U.S. Department of Education.

Gordon, B. (1999). Score issues from a practitioner's perspective. In *Issues in scoring essays: Research and practice.* Symposium conducted at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Gribbons, B., & Winter, P. C. (1999, January) *Using multiple measures of student achievement.* Unpublished proposal submitted to the U.S. Department of Education.

Hansche, L. N., Winter, P., Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington, DC.

Hansche, L., Stubits, T., Winter, P., et al. (1998, May). *Using existing assessments for measuring student achievement: Guidelines and state resources.* Washington, DC: Council of Chief State School Officers.

Harris, D. E., & Carr, J. F. (1996). *How to use standards in the classroom.* Alexandria, VA: Association for Supervision and Curriculum Development.

Jaeger, R. M., & Tucker, C. G. (1997). *Analyzing, disaggregating, reporting, and interpreting students' achievement test results: A guide to practice for Title I and beyond.* Washington, DC: Council of Chief State School Officers.

Kopriva, R. (2000). *Ensuring accuracy in testing for LEP students: A practical guide for assessment development.* Washington, DC: Council of Chief State School Officers.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435-447.

LaMarca, P., Redfield, D. L., & Winter, P. (2000). *State standards and state assessment systems: A guide to alignment.* Washington, DC: Council of Chief State School Officers.

Liu, K. K., Anderson, M. E., Swierzbin, B., & Thurlow, M. L. (1999, April). *Bilingual accommodations for LEP students on statewide reading tests.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Liu, K. K., & Thurlow, M. L. (1999). *What state accountability reports are saying about students with limited English proficiency.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Liu, K. K., Thurlow, M. L., Erickson, R., & Spicuzza, R. (1997). *A review of the literature on students with limited English proficiency and assessment* (State Assessment Series Minnesota Report No. 11). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? *Educational Measurement: Issues and Practice, 3*(3), 9-15.

Mehrens, W. A., & Kaminski, J. (1989, Spring). Methods of improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice,* 14-21.

Mehrens, W. A., Popham, J. W., & Ryan, J. M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*(1), 18-22.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

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

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

National Assessment of Educational Progress. (n.d.) http://nces.ed.gov/nationsreportcard/ (retrieved 1-28-01).

National Education Association. (1993). *Student portfolios.* Washington, DC: NEA Professional Library.

Office of Educational Research and Improvement. (1991). *Striving for excellence: The national education goals.* Washington, DC: U.S. Department of Education.

Olson, L. (1999). Quality Counts '99. *Education Week.* Washington, DC: Editorial Projects in Education.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership, 48*(5): 60-63.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: National Council on Measurement in Education and American Council on Education.

Pike, E. O., Jr. (1973). *Influence of a test-taking skills instructional program on the Metropolitan Achievement Tests performance of children from low-income families.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Popham, J. (1998). *Inappropriate uses of tests to judge school effectiveness.* Richmond, VA: Virginia Association of Test Directors.

Priestley, M. (1982). *Performance assessment in education and training: Alternative techniques.* Englewood Cliffs, NJ: Educational Technology Publications.

Ragosa, D. R. (1998, May). *Accuracy of individual scores and group summaries.* Professional development session for Council of Chief State School Officers; State Collaborative on Assessment and Student Standards, Durham, NC.

Regional Educational Laboratories (1998). *Improving classroom assessment: A toolkit for professional developers.* Portland, OR: Northwest Regional Educational Laboratory.

Rudner, L. M., Conoley, J., & Plake, B. (Eds). (1989). *Understanding achievement tests: A guide for school administrators.* Washington, DC: American Institutes for Research.

Sabers, D. L. (1975). *Test-taking skills.* Tucson: University of Arizona, Arizona Center for Educational Research and Development.

Samson, E. (1985). Effective training in test-taking skills on achievement test performance: A quantitative syntheses. *Journal of Educational Research, 78,* 261-266.

Scruggs, T. E., White, K. R., & Bennion, K. (1986). Teaching test-taking skills to elementary-grade students: A meta-analysis. *The Elementary School Journal, 87,* 69-82.

Shepard, L. A. (1987). *A case study of the Texas Teacher Test: Technical report.* Los Angeles: University of California, Graduate School of Education, Center for the Study of Evaluation.

Texas Education Agency. (1998). *Professional development and appraisal system implementation manual for appraisers and teachers.* Austin, TX: Texas Education Agency.

Thurlow, M. L., Langenfeld, K. H., Nelson, J. R., Shin, H., & Coleman, J. E. (1998). *State accountability reports: What are states saying about students with disabilities?* (Synthesis Report No. 20). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., McGrew, K., Ysseldyke, J., Elliott, J., Thompson, S., & Phillips, S. (1999). *Accommodations research: Design and analysis considerations.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities in large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment: Students with disabilities in the KIRIS on-demand assessments from 1992-93 through 1995-96* (State Assessment Series Maryland/Kentucky Report No. 3). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (1994). *Goals 2000: A world-class education for every child.* Washington, DC: Author.

U.S. Department of Education. (1998). *Condition of education 1998.* Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2000). *State ESEA Title I participation information for 1996-97: Summary report* (Doc. #2000-01). Washington, DC: Office of the Under Secretary, Office of Elementary and Secondary Education.

Utah State Office of Education. (1999a). *Test-taking tips and strategies: Student/parent pamphlet.* Salt Lake City, UT: Utah State Office of Education.

Utah State Office of Education. (1999b). *Test-taking tips and strategies: Teacher/administrator guidelines.* Salt Lake City, UT: Utah State Office of Education.

Virginia Department of Education (1995). *Standards of quality and standards of accreditation*. Richmond, VA: Author.

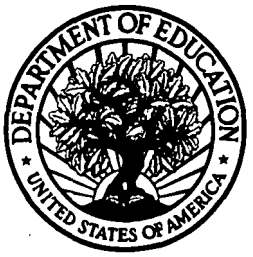
Wahlstrom, D. D. (1998). *Practical ideas for teaching and assessing the Virginia SOL*. Virginia Beach, VA: Successline, Inc.

Webb, N. (1997, January). *Determining alignment of expectations and assessments in mathematics and science education*. National Institute for Science Education, *1*(2).

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Ysseldyke, J., Olsen, K., & Thurlow, M. (1997). *Issues and considerations in alternate assessments*. (Synthesis Report No. 27). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., Thurlow, M., Nelson, R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report No. 23). Minneapolis: University of Minnesota, National Center on Educational Outcomes.



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").