DOCUMENT RESUME

ED 472 155 TM 034 710

AUTHOR	Baker, Eva L.; Linn Robert L.
TITLE	Validity Issues for Accountability Systems. CSE Technical Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-585
PUB DATE	2002-12-00
NOTE	34p.
CONTRACT	R305B960002
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Accountability; *Educational Assessment; Educational Change; Educational Testing; Elementary Secondary Education; Test Use; *Validity

ABSTRACT

This report analyzes the validity issues that arise in the context of educational accountability systems. The report addresses validity from three interlocking perspectives. The first explores the theory of action underlying accountability provisions. It considers problems emerging from the distance between aspirations for accountability in educational reform and the actual strength of the research base supporting sets of policies and procedures. A second component of the analysis concentrates on the role of testing in accountability systems, as it defines the characteristics and potential of many systems. The discussion is grounded strongly in the "Standards for Educational and Psychological Testing" (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education; 1999). The third perspective results in suggestions about an approach to improve the validity of accountability systems. (Contains 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 472 155

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

K. Hurst

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ☐ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Validity Issues for Accountability Systems

CSE Technical Report 585

Eva L. Baker
CRESST/University of California, Los Angeles

Robert L. Linn
CRESST/University of Colorado at Boulder

TM034710

UCLA Center for the Study of Evaluation

In Collaboration With:

University of Colorado at Boulder • Stanford University • The RAND Corporation
University of Southern California • Educational Testing Service
University of Pittsburgh • University of Cambridge

BEST COPY AVAILABLE

Validity Issues for Accountability Systems

CSE Technical Report 585

Eva L. Baker
CRESST/University of California, Los Angeles

Robert L. Linn
CRESST/University of Colorado at Boulder

December 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.2: Systems Design and Improvement: Ideal and Practical Models for Accountability and Assessment

Eva L. Baker, CRESST/UCLA, and Robert L. Linn, CRESST/University of Colorado at Boulder, Project Directors

Copyright © 2002 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions and policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

VALIDITY ISSUES FOR ACCOUNTABILITY SYSTEMS[1]

Eva L. Baker
CRESST/University of California, Los Angeles

Robert L. Linn
CRESST/University of Colorado at Boulder

Abstract

The purpose of this report is to provide an analysis of the validity issues that arise in the context of educational accountability systems. We will address validity from three interlocking perspectives. The first explores the theory of action underlying accountability provisions. Here, we will consider problems ensuing from the distance between aspirations for accountability in educational reform and the actual strength of the research base supporting sets of policies and procedures. A second component of our analysis will concentrate on the role of testing in accountability systems, as it defines the characteristics and potential of many systems. This discussion is grounded strongly in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 1999). The third set of issues will offer suggestions about an approach to improve the validity of accountability systems.

Theory of Action for Accountability Systems

The theory of action underlying the adoption of accountability systems derives from the adage "knowledge is power." It assumes that when people (or institutions) are given results of an endeavor, they will act to build on strengths and remedy or ameliorate weaknesses. Such positive actions depend on at least seven enabling conditions.

1. The results reported are accurate.
2. The results are validly interpreted.
3. The cognizant individuals are willing to act and can motivate action by team members.

[1] A version of this report will appear as a chapter in S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning Accountability Systems*, New York: Teachers College Press.

4. Alternative actions to improve the situation are known and available.

5. Cognizant individuals and team members possess the requisite knowledge to apply alternative methods.

6. The selected action is adequately implemented.

7. The action(s) selected will improve subsequent results.

The theory also assumes that barriers to improvement have lower strength than the desire to achieve goals and that there are clear and powerful incentives for positive actions.

Richard Elmore (in press) and Jennifer O'Day (in press) address aspects of a theory of action. In particular, they delve in some detail into questions of how accountability stakes motivate different actors' to alter educational practice. In this chapter, we focus on issues of accuracy of information and the validity of interpretation of results produced by accountability systems, but it is important to recognize that those aspects derive their importance within the broader theory of action just outlined.

Parsing educational reform in this framework raises numerous questions. The first two enabling conditions, accurate and validly interpreted results, depend upon the quality of measures available and the capacity of the user to understand and interpret information. The first of these concerns is extensively treated in the section below on assessment. In quick summary, it may be that some assessments are not sensitive to instructional remedies and therefore are unsuitable for the accountability purposes to which they have been put. The second concern, the ability of individuals to use systematically derived information, is a known problem in education. Research at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) on the development of the Quality School Portfolio (QSP) (Baker, Bewley, Herman, Lee, & Mitchell, 2001) has documented that lack of sophistication in data interpretation on the part of many members of the school community, from school board members to teachers. However, it has also suggested that there is a real appetite for learning about how results can lead to improvement. Such experience leads us to believe that the third condition—willingness to act—may be a reality, despite commentary to the contrary about cleaving to the status quo and exerting low energy. The fourth condition, knowledge of powerful alternatives, is unlikely, partly because there is little history of systematically documenting the effects of such alternatives in the first place, and making them

available to teachers in the second place. It is far more likely that the teacher or instructional leadership team draws upon a palette of limited value, one largely composed by happenstance. In part, this limitation relates to the woeful lack of systematic curriculum designed to help students achieve the knowledge and skills required by standards-based assessments used in accountability systems. It is obvious that the knowledge of a system or an approach is not equivalent to the knowledge of how to use the approach well, and educational literature is replete with discussions of the lack of background—sometimes, for instance, particular content knowledge—upon which pedagogical knowledge hinges. While there is some motion to replace teacher-generated instruction with more lock-step, scripted formats, teacher knowledge of the to-be-learned standards, of pedagogical strategies, and of the students themselves is required if any productive extemporizing is to occur. The sixth condition has led to a particular focus in the evaluation world on implementation, that is, the need to verify that any alternative has been implemented as intended. In fact, results from many experimental studies have been discounted because the treatment variations were not implemented as planned, resulting in great within-group differences in process by teachers. Finally, it is difficult to know *a priori* whether an instructional treatment, even if all previous conditions have been met, will be effective for the particular student group, standard, and context for which it has been implemented.

It should be clear that each of these conditions alone could substantially alter the likelihood of success of this theory of action. It defies reasonable expectation that these components will smoothly link together across teacher background, subject matter, student population, and educational setting. However, the components are supposed to work in this manner.

What is expected to focus the energy of the people in classrooms and schools to do now what they have been unwilling or unable to do before—that is, to systematically improve learning for students who have done poorly in the past? There is a belief that the power of incentives and sanctions will come into play and organize attention in the desired direction. Of concern to us as observers is that the rewards and sanctions may indeed focus attention on the bottom line, but not on needed steps or processes to get there. A lack of capacity (whether through selection, turnover, or inadequate professional development and resources) cannot be directly remedied by increased motivation to do well, especially over a short period. The central notion of the validity of accountability systems herein resides. Accountability

systems intending to promote real learning and improved effectiveness of educational services must themselves be analyzed to ensure that changes in performance (the proverbial bottom-line) are real, are due to quality instruction plus motivation, are sustainable, and can be attributed to the system itself. Before we further address how such accountability information could be obtained, let us turn our attention to the core of all educational accountability systems, the measures of student achievement.

Educational Testing and Assessment

Since testing is the key feature of systems currently under consideration at the federal level, as well as those that have been implemented by states in the last decade, a substantial portion of this chapter deals with the validity of uses and interpretations of tests. There are, however, broader validity issues for accountability systems, which go beyond those normally thought of in connection with tests, and we will also address some of those issues.

Our discussion will make frequent reference to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), which we will refer to as the Test Standards. The Test Standards are widely recognized as the most authoritative statement of professional consensus regarding expectations for tests on matters of validity, fairness, and other technical characteristics of tests. The Test Standards define validity as follows: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). The Test Standards go on to say that "validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 9).

As is clearly indicated in the Test Standards, validity is not a property of a test, but rather a property of the specific uses and interpretations that are made of test scores. Hence, it is not appropriate to make an unqualified statement that an assessment is valid. Rather, the assessment that has a high degree of validity for a particular use may have little or no validity if used in a quite different manner. For this reason, the Test Standards admonish the developers and users of assessments to start by providing a rationale "for each recommended interpretation and use" (AERA, APA, & NCME, 1999, p. 17).

Intended Uses and Interpretations of Tests

Tests are used for a wide array of purposes, ranging from low-stakes diagnosis for instructional purposes to high-stakes uses such as the award of high school

diplomas. At the institutional level, high stakes may mean the identification of schools that are failing or of schools where teachers are given substantial monetary reward for progress shown in the test results of students in the school. Although the uses of tests by teachers for day-to-day instructional purposes are among the most significant uses that are made of tests for improving instruction (Black & Wiliam, 1998a, 1998b), our focus is limited to the uses of tests for externally mandated accountability purposes; however, we expect that in the future, such differences in purpose may blur. For example, in the Los Angeles Unified School District, classroom administered assessments are intended to guide both instructional practice and provide information about effectiveness. Common models are used to guide the design of tests so that standards, cognitive demands, apt content and criteria are common to all purposes (Baker, 1997). In order to create assessments that provide a common framework for teacher practice, work is in process designing authoring systems for teachers to use to measure standards. Such systems may very well allow the use of teacher-developed tests to be aggregated to supplement externally mandated examinations (Baker & Niemi, 2001). The utility of a coherent system, in which assessments used at all levels, for internal purposes such as on the spot improvement of learning, for teacher planning, and for accountability, is obvious. One such vision in science has been proposed by Pellegrino, Chudowsky, and Glaser (2001). Even so, there will remain many variations in uses and interpretations of test results that deserve attention within the context of accountability systems, and those variations can have important implications for the evaluation of the validity of specific inferences drawn from test results and the decisions that are based on those results.

The following examples provide some indication of the range of uses and interpretations of test scores that are made within the context of accountability systems.

- Students who do not obtain a passing score on a test must attend summer school and pass an alternate form of the test to be promoted to the next grade.
- Students must score at the proficient level or higher on tests in four subject areas in order to receive a high school diploma.
- Teachers in schools that rank in the top 10% in terms of gains on the state's school accountability assessment will receive a bonus of $25,000.

- Parents of students attending schools found to be failing as defined by the test performance of their students may transfer their children to another public school.

- Schools with schoolwide Title I programs that fail to make adequate yearly progress in student test performance will be declared unsatisfactory and targeted for assistance.

- To be accredited by the state, schools must either have overall student achievement at or above a specified goal on the state assessment or meet targets for gains in student achievement.

Each of these examples of test use, as well as others that could be specified, has a number of validity questions associated with it. Each demands the identification of the most salient of those questions and the accumulation of evidence relevant to answering those salient questions. We will illustrate some of the issues that are linked most closely to specific uses and interpretations. There are, however, some issues that are general across the variety of uses of tests in accountability systems. We will begin with a discussion of those general issues. That discussion will be followed by a discussion of validity issues that are most relevant to three broad uses, beginning with the use of test scores for making high-stakes decisions about individual students. We will next consider uses of test results for making high-stakes decisions about schools. We will then turn to a discussion of the impact of accountability systems on instruction and learning. We will end with a brief summary and conclusion section.

Test Specifications

Educational achievement tests focus on content domains, such as reading, mathematics, and science. Such tests are intended to provide evidence of what a student knows and is able to do in a content domain without regard to an external criterion measure, such as subsequent performance in college or in the workplace. Hence, the content of an educational achievement test is an appropriate starting place for the validation process. The content of a test is critical to the creation of scores that support valid inferences about student achievement.

Two questions are central in the evaluation of content aspects of validity. Is the definition of the content domain to be assessed adequate and appropriate? Does the test provide an adequate representation of the content domain the test is intended to measure? The first of these questions focuses on the content standards that states have developed to specify the content that teachers are expected to teach and

students are expected to learn. The content standards also specify the domain that a state test is expected to measure. The adequacy of the content standards for specifying the domain the test is intended to measure will generally depend upon the specificity and concreteness of the content standards. Given the breadth of most content standards, there is usually a need to create a table of test specifications that serves to map content standards into detailed prescriptions for the makeup of tests. Tables of specifications usually provide the basis for mapping test items according to specific content (e.g., addition, subtraction, multiplication) and process (e.g., factual knowledge, conceptual understanding, problem solving) categories, represented in different item formats (e.g., multiple choice or short answer). There are additional levels of specificity that might well be desirable to create a full descriptive system of test content (Baker, 2000), including a finer grain analysis of cognitive demands (see, for example, Anderson & Krathwohl, 2000) and linguistic characteristics of items (Abedi, 2001; Bailey, 2000; Butler, Stevens, & Castellon-Wellington, 1999; Stevens, Butler, & Castellon-Wellington, 2000). The adequacy of content aspects of validity is judged in terms of the definition of the content domain identified by the test specifications and the representativeness of the coverage of that domain by the test.

Whatever the breadth and depth of coverage or emphases of the content standards, it is generally intended that the assessment will be well enough aligned with the content standards so that student performance on the assessment can be used as the basis for making inferences about the degree to which a student has mastered the domain of content defined by the standards. Detailed analyses of the relationship between the content domain of the content standards and the specific content of the assessment are needed to support such inferences. Confirmation of alignment of the test items and content standards by independent judges provides one type of evidence. This may be accomplished by having judges assign assessment tasks to the content standards they believe the tasks measure and comparing those assignments with the assignments of the developers of the assessment tasks. The Test Standards are explicit about the need to relate the content of the test to that of the content standards, which are referred to as curriculum standards.

> **Standard 13.3** When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domains should be described

in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent. (AERA, APA, & NCME, 1999, p. 145.)

It is noted in the comment following standard 13.3 that tests are unlikely to cover the full domain of content covered by content standards. Hence, it is important to make it clear which aspects of the content standards are left uncovered by the test, which are covered only lightly, and which receive the greatest emphasis. Such an analysis provides a basis for judging the degree to which generalizations from the assessment to the broader domain of the content standards are defensible. Messick (1989) referred to the threat to validity of inadequate coverage of the domain as *construct under-representation*. Construct under-representation is a major concern in large-scale assessment because of the potential effect that only aspects of the domain that are relatively easy to measure will be assessed, which, in turn, can lead to a narrowing and distortion of instructional priorities.

In addition to identifying the content that students are expected to learn, content standards adopted by states generally also specify the cognitive processes that students are expected to be able to use (e.g., reasoning, conceptual understanding, problem solving). Hence, judgments of the alignment of the test with content standards need to attend to cognitive processes that students need to use to answer test items, as well as the content. This validity expectation is made clear in the Test Standards.

Standard 1.8 If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided. (AERA, APA, & NCME, 1999, p. 19.)

Use of Tests to Make High-Stakes Decisions About Individual Students

Evaluating the adequacy and appropriateness of test content and of the cognitive demands of a test provides only one link in a validity argument. Other links depend on evidence that can be used to judge the adequacy and appropriateness of the uses that are made of test results and the interpretations of the scores. The latter considerations clearly depend on the specific uses and interpretations that are made of test scores. Our discussion of the validity demands associated with specific uses is divided into three broad categories of use. We begin with uses for high-stakes decisions about individual students and then turn to uses for high-stakes decisions about schools.

Establishing Performance Standards

Using tests to make high-stakes decisions, such as for grade-to-grade promotion or high school graduation, involves the use of a passing score on the test. Performance standards are set, and cut scores on the test are identified that yield interpretations—for example, performance above the cut score implies that the student is proficient (passing), and performance below the cut score indicates that the student is not proficient (failing). The validity of these standards-based interpretations, also called criterion-referenced interpretations, depends on the appropriateness of the cut score. At a minimum, the interpretation needs to be supported by a rationale, as required by the Test Standards.

> **Standard 4.9** When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained. (AERA, APA, & NCME, 1999, p. 56.)

The rationale for a cut score to be used to define performance that is called *proficient*, for example, might include a description of the basis for the adoption of content standards and a description of the process used to identify judges and to obtain *judgments*; the definition of *proficient* used by the judges; and the process used to elicit judgments of performance on the test that was considered proficient. The rationale for a cut score used to determine grade-to-grade promotion might be similar to that provided for determining proficient performance, but might also include an analysis of the performance in the next grade for students whose scores are above and below the cut score.

Classification Errors

The use of performance standards to determine whether a student is proficient or not (passes or fails) reduces test scores to a dichotomy. Measurement error that is associated with any test score results in classification errors. That is, a student whose true level of achievement should lead to a passing score earns a score that is below the passing standard and vice versa. Valid inferences about student proficiency are undermined by measurement errors that result in misclassification of students. Hence, it is critical that the probability of misclassification is evaluated and the information is provided to users of the performance standards results. The precision of test scores can be enhanced by increasing test length. As Rogosa (1999a) has shown, however, even tests that have reliability coefficients normally considered to

be quite high (e.g., .90) result in substantial probabilities of misclassification. For example, if the passing standard is set at the 50th percentile for a test with a reliability of .90, the probability is .22 that a student whose true percentile rank is 60, and who therefore should pass, would score below the cut score and therefore fail on a given administration of the test. Even a student whose true percentile rank is 70, a full 20 points above the cut score, would have a probability of failing of .06 (Rogosa, 1999a).

> **Standard 13.14** In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores. (AERA, APA, & NCME, 1999, p. 148.)

Multiple Opportunities to Take Alternate Forms of the Test

Students should also be provided with a reasonable number of chances to take equivalent versions of the test before being retained in grade or denied a diploma, and with additional opportunity to learn between test administrations.

> **Standard 13.6** Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences. (AERA, APA, & NCME, 1999, p. 146.)

The importance of providing multiple opportunities to pass a test using alternate forms of a test when failure has high-stakes consequences can be illustrated by a simple example. Assume, for example, that the cut score has been set at a level corresponding to the 10th percentile. Rogosa's (1999a) analyses, show that if the test has a reliability of .90 that a student whose true performance was at the 20th percentile would have a probability of scoring below the cut score due to errors of measurement of .0633. If given a second opportunity to take an equivalent form of the test, however, the probability that the student would score below the cut score a second time would drop to .0040. Thus, while there would still be a non-zero probability that the 20th percentile student would fail twice in a row due to errors of measurement, the probability is substantially reduced by providing the second

opportunity and would, of course, be reduced still further by a third testing opportunity.

Multiple Ways of Demonstrating Specified Competencies

Since no test can provide a perfectly accurate or valid assessment of a student's mastery of a content domain, the Test Standards caution against over-reliance on a single test score when making high-stakes decisions about students. The Test Standards indicate that multiple sources of information should be considered when the addition of information other than a test score enhances the validity of the decision.

> **Standard 13.7** In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision. (AERA, APA, & NCME, 1999, p. 146.)

This statement is consistent with conclusions reached in a National Academy of Sciences report prepared by a committee formed in response to a Congressional mandate to review the use of tests for purposes of tracking, grade-to-grade promotion, and graduation (Heubert & Hauser, 1999). According to a recent decision by the U.S. District Court for the Western District of Texas (GI Forum et al. v. Texas Education Agency, 2000), the inclusion of other information in a decision that may have a major impact on students need not be done in a compensatory manner. The court ruled that Texas could require students to exceed a specified score on the Texas Assessment of Academic Skills (TAAS) test as well as pass certain required courses, thus allowing a conjunctive use of a test requirement.

In addition to using alternative indicators of student achievement to supplement test score information when making high-stakes decisions about students, it is often desirable to permit the substitution of alternate measures for test scores. Alternate indicators of achievement can be especially important in cases in which student performance on a test is likely to give a misleadingly low indication of the student's knowledge and understanding of the material because of debilitating test anxiety or student disabilities that call the validity of standardized test results into question.

Opportunity to Learn Material Tested

For tests used to determine grade-to-grade promotion or high school graduation, the Test Standards call for evidence regarding the opportunity students have to learn the material for which they are being held responsible.

> **Standard 13.5** When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills students have had an opportunity to learn. (AERA, APA, & NCME, 1999, p. 146.)

There are legal as well as moral and educational reasons for ensuring that students are provided with an adequate opportunity to learn the material on tests used for high-stakes decisions such as determining the award of high school diplomas. In Debra P. v. Turlington (1981), the court ruled that the graduation test must be a fair measure "of that which was taught" (406).

Remediation for Students With Repeated Failures

Accountability is most effective when it encourages shared responsibility for results. When individual students are held accountable for meeting established performance standards on a test, it is critical that teachers and the educational system also be held accountable for providing adequate opportunity for students to meet the established standards. When students continue to fail to meet the standards on the test after repeated attempts, it is critical that the educational system be held responsible for providing continued remediation.

Use of Test Results for High-Stakes Decisions About Schools

Many of the Test Standards are easily extrapolated to inferences drawn about schools. However, there are particular issues that might be considered for institutions. Consequences for poor performance may mean additional assistance, public identification, or consequent transfer of leadership or staff, either voluntarily or directed.

Subjects and Grades Tested

Accountability systems differ in who is tested and on what content. Some systems test every student in adjacent grades, allowing for an apparently longitudinal picture of growth. However, because the tests given at Grade 4 and Grade 5 will be different, interpretation of results may be confusing. For example, students scoring at the 50th percentile in the fourth grade who, in the following

year, score at the 50th percentile in the fifth grade did not stand still; they learned a considerable amount of new material. But often such results are used as evidence that the educational system is not making progress.

Consistent with recent law (Improving America's Schools Act [IASA], 1994), many systems focus only on particular grade levels (e.g., fourth-grade students) and on only a subset of subject matters (e.g., reading and mathematics). These emphases can have predictable results. The first, focusing on successive cohorts of students, assumes that changes from year to year in fourth-grade performance can be attributed to improvement, or lack thereof, in the instructional program. In fact, sources of error in the inference are many. They include changes in the student population from year to year. For example, in one California district, performance in mathematics and reading greatly improved in a single year. Although attributions to a talented principal were generally made, the finding was actually attributable to a business closing in a nearby wealthier county and the influx of a well-prepared student group to the target school (L. Burstein, personal communication, 1988).

The second source of error is the idea that the school itself is unchanged, and that the constancy of the building itself is equivalent to the constancy of the staff. We know, however, that urban turnover rates of teachers are high and that there is no evidence (and no request for it) to support that it is the same teacher, principal or team that accounts for performance year to year.

A third source of error is the emphasis on the subject matter taught. Warnings about "narrowed" curriculum, or glowing reports of "focused instruction" may amount to the same thing. If schools are to be responsible for services other than those measured in the accountability system, such as the arts, sciences, or community service, such efforts must find their way into the accountability system.

Characteristics of Students Attending Schools

Testing results are notoriously sensitive to student background characteristics. These characteristics include the economic and educational levels of parents, parents' expectations for student success, students' language backgrounds, the average length of time students attend a school, and the regularity with which students come to school. Even though there are notable variations in performance that can be found within these factors, it is clear that accountability systems must address these differences. We also know that students with at-risk backgrounds

often attend schools with fewer credentialed teachers and fewer resources, and in less well maintained facilities.

One approach is to focus on the absolute status of the school—how its students at targeted grade levels are doing in a cross-sectional view. In addition, some systems report growth, the change from year to year. Thus, the inferences about student growth, or the targets set for individual schools need to recognize these differences. Far different actions may in fact be inferred, however. Schools with children far behind may in fact need to have high growth targets if they are to catch up to more affluent students. Yet, such catching up is likely to be difficult for students who may not have acquired neither the desired prerequisite knowledge, skills, and cognitions, nor attitudes that support school achievement.

Inclusion of Students With Limited English Proficiency

When we think of census testing, that is, testing all students in a school, we normally believe that every child will be included in test results. However, there are numerous examples in the past of testing programs that have tested all students, but systematically excluded results of a subset. To remedy this problem, many accountability systems require that the percent of students tested in the school be published. Further, some, such as the current California system, require that a stated percentage must be tested if the school is to receive a special monetary award.

How best to handle the inclusion of students with emerging English proficiency is not straightforward. Language experts and parent advocates argue that students should be tested in their native language until they have demonstrated a sufficient level of English language proficiency. Some states have prohibitions about testing in translated languages, or translate only one or two languages when students may represent 50 or 100 languages and dialects. Other systems provide various linguistic accommodations to assist students in testing. Common accommodations include longer testing periods, glossaries, or oral support. It is true that many of these accommodated test conditions are not subjected to validity studies to determine whether the construct or domain tested has been significantly altered. In part, this lack of empirical data results from restricted resources. Nonetheless, the major threat to accurate interpretation is that of construct-irrelevant variance (Messick, 1989), in which inferences drawn about the domain under examination may be contaminated by difficulty experienced by students in

deciphering the meaning of the language in which the test question(s) may be embedded. Standard 9.1 deals with this concern.

> **Standard 9.1** Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences. (AERA, APA, & NCME, 1999, p. 97.)

Inclusion of Students With Disabilities

Students with disabilities are also required to participate in accountability-focused testing programs. Depending upon the nature of the disability, students may be given accommodations involving more time, or sensory support, or may, in fact, be given an alternative assessment intended to address the particular goals identified for the learner. Such modifications or alternate assessments may themselves not be subject to empirical study, a fact recognized in the Test Standards.

> **Standard 10.3** Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications. (AERA, APA, & NCME, 1999, p. 106.)

Accounting for Performance of Identifiable Subgroups of Students

Many accountability systems and the proposed ESEA legislation require that student performance be disaggregated by identifiable subgroups. The logic is that schools should not try to meet their accountability targets by focusing on one or more groups to the exclusion of others. Many such systems require that progress by each subgroup must reach a particular threshold (e.g., 80% of the projected target). In some cases, even where the school as a whole has met a growth target, rewards are withheld if one or more subgroups had sub-par performance.

Classification of Schools

Accountability systems typically array schools into categories intended to reflect their level of performance and rate of progress in meeting explicit standards. These classifications may be wholly based on weighted averages of test score performance in some or all grades tested. There is a growing literature that addresses the problem of reliably classifying schools in categories such as *advanced, adequate,* or *needs improvement* (Kane & Staiger, in press; Linn, in press; Linn & Haug, in press, Rogosa, 1999b). The probability of misclassifying a school based on student test scores depends on a number of factors, including the number of categories used,

the number of students that enter into the calculation of the school's scores used in the classification, whether current performance or change in student performance is used, and whether subgroup performance as well as the total group performance is used to determine classification.

The research investigating the dependability of school-level results and classification error rates has shown that the probability of misclassification is substantial. Error rates increase as the number of students decreases. Thus the probability that a small school will be misclassified is greater than the probability of misclassification of a large school. Because of the relationship of misclassification probability to school size, it is common for the set of schools identified as the best performers, as well as the set identified as the worst performers to be disproportionately made up of small schools. Also as a consequence of the effects of size on misclassification probabilities, it is common to find that the schools that look best (or worst) based on their gains from year 1 to year 2 will generally not show up in the same category based on their gains from year 2 to year 3. As Linn and Haug (in press) have shown, there is a negative correlation between the gains schools make from year 1 to year 2 with the gains they make from year 2 to year 3.

Accountability systems that demand gains for schools not only for the total student population but also for all subgroups of students defined by the socioeconomic background or the racial/ethnic group to which students belong will also have higher rates of misclassification. As Kane and Steiger (in press) have shown, the disaggregation of scores by racial/ethnic group exacerbates the problems of volatility of school-level results.

Uncertainty is also greater when change scores are used as opposed to when status measures are used. It is well known that difference scores are less reliable than the scores that are used to compute the difference. This general result for individual student scores also applies to scores for schools.

The limited precision in estimates of school improvement based on comparisons of successive groups of students presents a major challenge for school accountability systems that rely on annual improvements in the performance of successive cohorts of students. There are several approaches that can be used to help ameliorate the problems of imprecision and the resulting high probabilities of misclassification errors. Accuracy can be improved by combining data across multiple grades, multiple subject areas, and/or multiple years. Combining across

either grades or years increases the precision of results by increasing the number of students used to estimate school results. Combining across grades has the added advantage of increasing the number of teachers who are teaching students whose performance directly contributes to the accountability results for the school, and thereby may increase the sense of shared responsibility of results. Combining across several years lengthens the accountability cycle, but produces results that are more trustworthy and therefore more likely to lead to real long-term improvements and to the identification of exemplary practices, as well as enhance fairness.

Whatever the level of precision of school-level results, the results for schools should be accompanied by information about the dependability of those results as required by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). This might best be done where schools are placed into graded performance categories by reporting information about the accuracy of classifications. Procedures for evaluating school-building misclassification probabilities are described by Rogosa (1999b) and by Hoffman and Wise (2000).

Impact of Accountability System on Instruction and Learning

The Test Standards require that validation of test use for high-stakes decisions about students include attention to evidence about the intended and unintended consequences of those uses. Test requirements for promotion or graduation clearly are intended to ensure that students have mastered specified content before they are allowed to move on to the next grade or graduate. There is also the implicit intent that students will learn more in the long run if they are held accountable for achieving at a specified level for the promotion or graduation decision. The Test Standards require that evidence be provided so that a reasonable evaluation can be made of the degree to which these intentions are realized by the promotion or graduation policy.

> **Standard 1.23** When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. (AERA, APA, & NCME, 1999, p. 23.)

In addition to providing evidence that the intended effects of the test requirements are met to a reasonable degree, the Test Standards require that

attention also be given to the collection of evidence relevant to plausible unintended negative consequences of the test use.

> **Standard 1.24** When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure to fully represent the intended construct. (AERA, APA, & NCME, 1999, p. 23.)

Improving Accountability Systems

We have embarked on a great national experiment, with various states attempting to meet desirable performance goals using their own systems of tests and accountability. How do we support the good intentions of policymakers in improving schools and simultaneously correct processes that may potentially mislead us? Furthermore, how can we create systems that will motivate students and educators to focus on high standards without sacrificing quality instruction and breadth of learning? The real answer is that no one knows for sure. Our proposal, however, is to promote a set of accountability standards to assist policymakers, the public, and the education community in understanding the quality of accountability systems in place. Standards for products, systems, and services give users and managers criteria to use to improve the quality of their efforts and outcomes. In education, standards have been promulgated for instructional products, for interoperability of software, for tests (AERA, APA, & NCME, 1999), for evaluations, and for desired goals and competencies of teachers and students.

In this era of educational accountability, standards for accountability system design, operation, and interpretation can assist educational policymakers, managers, teachers, the media, and parents in developing reasonable expectations and drawing appropriate conclusions from test results or other systematically collected educational information. The standards may help such systems avoid inadvertent negative effects and, instead, promote the interests of students and educational personnel who participate in accountability systems.

The standards offered below (Baker, Linn, Herman, Koretz, & Elmore, 2001) represent models of practice derived from three perspectives: (a) research knowledge, (b) practical experience, and (c) ethical considerations. The standards are intended to guide those interested in improving the validity and utility of accountability information. Because experience with accountability systems is still developing, the standards we propose are to help evaluate existing systems and to

guide the design of improved procedures. They should be thought of as targets for systems. It is not possible at this stage of the development of accountability systems to know in advance how every element of an accountability system will actually operate in practice or what effects it will produce, so we also suggest standards for the evaluation of impact.

To accommodate the differing maturity levels of accountability systems, we have devised standards that fall into two general categories: (a) those that should be applied to existing systems, and (b) those that specify necessary evaluation requirements for new systems. It should be understood that tests included in an accountability system should meet the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). What we have highlighted here are criteria that apply especially to accountability systems. It is likely also that additional standards will be subsequently developed based on evaluations of accountability system effects.

A. Standards on System Components

1. Accountability systems should employ different types of data from multiple sources.

> Comment: Although measures of student achievement may be of primary interest for accountability purposes, it is important to also obtain information about student and teacher characteristics to provide context for interpreting student achievement. It also is important to consider other student outcome data such as attendance, mobility, and rates of retention in grade, dropout and graduation. Moreover, it is important to obtain data on instructional resources and curriculum materials, and on the degree to which students are provided with adequate opportunity to learn the content specified in content standards and curriculum materials.

2. The weighting of elements in the system, different test content, and different information sources should be made explicit.

> Comment: Making sense of overall accountability indices requires an understanding not only of the elements that go into the index, but of the weights that are assigned to each element. It is informative to provide not only the weights that are assigned to the different elements by policy, but also information about how each element relates to the overall index. The relationship of an element to a weighted accountability index depends on the variability of the element across institutions as well as the weight assigned to the element according to policy.

3. Accountability systems should include data elements that allow for interpretations of student, institution, and administrative performance.

Comment: Students, teachers, administrators, and policymakers have a shared responsibility for achieving the results expected by accountability systems. The system needs to provide the information for each of these parties to know what actions need to be taken.

4. Accountability expectations should be made public and understandable for all participants in the system.

Comment: Explicit information about expectations is a prerequisite for participants to perceive the accountability system as fair. It is also needed for participants to act in ways that will allow them to meet expectations and to monitor their progress.

5. Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.

Comment: Previous practices that excluded many students from testing, due to absence on the day of test administration, or because of limited English proficiency, or because of student disabilities, gave a distorted and usually exaggerated view of overall performance. It also meant that there was no accountability for the performance of excluded students. Legal requirements as well as ethical considerations demand that all students be included in the accountability system. Many students who would have been excluded in the past can be included without any alterations in the test or administration conditions. Some accommodations in administration conditions will be required for other students, and for some students the test will need to be modified, or alternative assessments used, in order for the students to be included in the accountability system. No student should be left out of the system, however.

B. Testing Standards

6. Decisions about individual students should not be made on the basis of a single test.

Comment: No test is perfectly valid or perfectly reliable. There is always a degree of uncertainty associated with any test score. That uncertainty needs to be taken into account when making decisions about individual students. This can be done by looking for other information that will either support or disconfirm the information provided by a single test score. The importance of obtaining other information to confirm or disconfirm the information provided by a single test score increases as the importance of the decision and the stakes associated with it increases.

7. Multiple test forms should be used when there are repeated administrations of an assessment.

Comment: The items contained on a test form are only a sample of the domain that the test is intended to measure. Learning the answers to the items on a single form by focusing exclusively on those items is not the same as learning the material for the domain of content the test is intended to measure. Consequently, it is important to evaluate the generalizability of performance by administering a different form when a test is administered for a second or third time.

8. The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.

Comment: Validity is dependent upon the specific uses and interpretations of test scores. It is inappropriate to assume that a test that is valid when used for one purpose will also be valid for other uses or interpretations. Hence, validity needs to be specifically evaluated and documented for each purpose.

9. If tests are to help improve system performance, data should be provided illustrating that the results are modifiable by quality instruction and student effort.

Comment: Tests need to be sensitive to differences in instructional quality and student effort in order to be useful as tools in improving system performance. Sensitivity to instruction and to student effort is also a prerequisite for fairness if educators and students are to be held accountable for results.

10. If test data are used as a basis of rewards or sanctions, evidence of technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published.

Comment: Because tests are fallible measures, classification errors are inevitable when tests are used to classify students or institutions into categories associated with rewards or sanctions. In order to judge whether the risk of errors is acceptably low, it is essential that information be provided about the probability of misclassifications of various kinds.

11. Evidence of test validity for students with different language backgrounds should be made publicly available.

Comment: Validity needs to be assessed separately for students with different language backgrounds. Whether a test is administered in English or in a student's primary language, validity of the test for students of different language backgrounds cannot be assumed from evidence based only on test results of students whose first language is English. Testing students in their primary language may be required for some students. However, translation and adaptation of tests to different languages is a complex

undertaking. There are many threats to validity of tests administered in different languages. Lack of consistency between the language of the test and the language of instruction is one of the major threats to validity that needs to be evaluated.

12. Evidence of test validity for children with disabilities should be made publicly available.

Comment: Accommodations may be needed for some students with disabilities to be able to participate in testing in a meaningful way. The goal of accommodations is to remove sources of difficulty that are irrelevant to the intent of the measurement. That is, an accommodation should make it possible for a student with disabilities to demonstrate their knowledge and skills in the content domain being tested so that the score reflects that knowledge and skill rather than the student's disability. The accommodation should level the playing field, but it is not intended to give the student with a disability an unfair advantage over other students. The validation task is to provide evidence that the test reflects the student's knowledge and skills but not the specific disability. For students with severe disabilities, assessments may need to be modified, or alternative assessments may need to be selected or developed, possibly designed to assess different learning goals than those of the assessments used for the majority of students. Evidence regarding the validity of interpretations made from modified or alternative assessments should be provided to the extent feasible.

13. If tests are claimed to measure content and performance standards, evidence of the relationship to particular standards or sets of standards should be provided.

Comment: The degree of alignment of a test with content standards may be evaluated, for example, by providing a mapping of the test specifications to the content standards. Such a mapping can reveal areas of the content standards that are not included in the test specifications as well as areas that are lightly or heavily sampled in the test specifications. The mapping may also reveal areas tested that are not part of the content standards. Performance standards generally provide verbal descriptions of levels of performance that are considered satisfactory or exemplary. The degree to which the descriptions map directly to the test items and the correspondence of the performance standards to the cut scores on the test need to be documented and evaluated.

C. Stakes

14. Stakes for accountability systems should apply to adults and students.

Comment: Asymmetry in stakes may have undesirable consequences, both perceived and real. For example, if teachers and administrators are held accountable for student achievement but students are not, then there are likely to be concerns about the degree to which students put forth their best effort in taking the tests. Conversely, it may be unfair to hold students accountable for performance on a test without having some assurance

that teachers and other adults are being held accountable for providing students with an adequate opportunity to learn the material that is tested.

15. Incentives and sanctions should be coordinated for adults and students to support system goals.

Comment: Incentives and sanctions that push in opposite directions for adults and for students can be counterproductive. They need to be consistent with each other and with the goals of the system.

16. Appeal procedures should be available to contest rewards and sanctions.

Comment: Extenuating circumstances may call the validity of results into question. For example, a disturbance during test administration may invalidate the results. Individuals may also have information that leads to conflicting conclusions about performance. Appeal procedures allow for such additional information to be brought to bear on the decision and thereby enhance its validity.

17. Stakes for results and their phase-in schedule should be made explicit at the outset of the implementation of the system.

Comment: Making plans for phasing in stakes for results is part of making accountability expectations explicit to participants. Explication of plans allows participants to make informed decisions about how best to achieve the ends expected by the accountability system.

18. Accountability systems should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions as the system aligns.

Comment: Starting with broad, diffuse stakes (e.g., public reporting of aggregate achievement results for schools) allows participants time to make the changes needed to meet expectations before being confronted with specific rewards or sanctions for performance (e.g., monetary rewards to schools or teachers, graduation requirements for students). Advance warning and phasing-in of stakes enhances both the perception of fairness and the actual fairness of the accountability system.

D. Public Reporting Formats

19. System results should be made broadly available to the press, with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results.

Comment: The press plays an important role in the interpretation of the results produced by accountability systems. Legitimate interpretations of results require an understanding of what goes into them and some of their technical characteristics. Those responsible for

the accountability system also have a responsibility to help ensure proper interpretation of the results and to minimize inappropriate interpretations to the extent possible. Efforts to assist the press in understanding the results, their strengths and limitations, and the legitimate and illegitimate interpretations can pay considerable dividends in improved coverage by the press and better understanding by the public.

20. Reports to districts and schools should promote appropriate interpretation and use of results by including multiple indicators of performance, error estimates and performance by subgroup.

Comment: Interpretations of results can be enriched by the reporting of consistencies and inconsistencies provided by multiple indicators of performance. Performance by subgroups needs to be considered to ensure that overall results do not conceal great disparities in subgroup performance. Understanding the degree of uncertainty in results can reduce the likelihood of misinterpretation and enhance the likelihood of appropriate use of results.

E. Evaluation

21. Longitudinal studies should be planned, implemented, and reported evaluating effects of the accountability program. Minimally, questions should determine the degree to which the system

a. builds capacity of staff;

b. affects resource allocation;

c. supports high-quality instruction;

d. promotes student equity access to education;

e. minimizes corruption;

f. affects teacher quality, recruitment, and retention; and

g. produces unanticipated outcomes.

Comment: The primary purpose of educational accountability systems is to improve instruction and student learning. The overarching evaluation question is the degree to which the intended benefits are realized and the costs in terms of unintended negative consequences are minimized. Listed items (a) through (d) reflect intended positive consequences, the realization of which is the focus of evaluation. Items (e) and (g) emphasize the needed evaluation of plausible unintended negative consequences. Item (f) requires the evaluation of both intended positive and unintended negative influences of the accountability system.

22. The validity of test-based inferences should be subject to ongoing evaluation. In particular, evaluation should address

a. aggregate gains in performance over time; and

b. impact on identifiable student and personnel groups.

Comment: Gains in performance may be spurious or real. Evaluation of the gains may be aided by investigations of the degree to which gains on the measures used by the accountability system are reflected in changes on alternative indicators of performance obtained from other tests, or more general indicators such as performance beyond school in college or the workplace. Differential effects on identifiable student or personnel groups may lead to different conclusions than those that are supported by the overall aggregate performance.

Application of the Accountability Standards

Standards abound in education—to guide content for students, to express expectations for performance, even to standardize software protocols. In the area of testing, the AERA-APA-NCME (1999) standards follow in a tradition of professional consensus and guide graduate training, testing practices, and legal interpretations. They are augmented by other efforts to summarize and highlight key concerns, including the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988/1994); *Responsible Test Use* (Eyde et al., 1993); *High Stakes: Testing for Tracking, Promotion, and Graduation,* (Heubert & Hauser, 1999); and *Testing, Teaching, and Learning: A Guide for States and School Districts* (Elmore & Rothman, 1999).

Summary and Conclusion

Educational accountability systems may not, by themselves, achieve the many goals held by their supporters. However, they are unlikely to do so unless their quality is improved. Through the adoption of these standards as achievable goals, state accountability systems themselves can become what they espouse—systems that learn from experience. To improve their quality and, as a result, the validity of inferences derived from their data, we suggest the following cycle. First, we need to understand the theories of action that support the development of one or another model, and address the implications of particular approaches. Second, without fail, the measures used to assess student and school performance should be grounded in and exemplify the best of the considerable research base associated with the technical quality of tests. Third, the public, parents, politicians, and educators should hold accountability systems to high technical standards. We must find a way

to support states and districts that attempt to reach accountability standards and to encourage the collection of evaluation data on them to assess the extent to which accountability systems or components help, are indifferent to, or undermine the goal of educational excellence.

References

Abedi, J. (2001). *Standardized achievement tests and English language learners: Psychometrics and linguistics issues.* Submitted for publication.

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

Anderson, L. W., & Krathwohl, D. R. (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives.* New York: Longman.

Bailey, A. (2000). Language analysis of standardized achievement tests: Consideration in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002, pp. 85-105). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*(4), 247-254.

Baker, E. L. (2000, November). Understanding educational quality: Where validity meets technology. In *William Angoff Memorial Lecture Series.* Princeton, NJ: Educational Testing Service.

Baker, E. L., Bewley, W. L., Herman, J. L., Lee, J. J., & Mitchell, D. S. (2001). *Upgrading America's use of information to improve student performance* (Proposal to the U.S. Secretary of Education). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Linn, R. L., Herman, J. L., Koretz, D., & Elmore, R. (2001, April). *Holding accountability systems accountable: Research-based standards.* Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Baker, E., & Niemi, D. (2001, June). *Assessments to support the transition to complex learning in science* (Proposal submitted to the Interagency Education Research Initiative [IERI], Program Solicitation NSF-01-92). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: King's College London, School of Education. (See also article with the same title, 1998, in Phi Delta Kappan, 80, pp. 139-148.)

Butler, F. A., Stevens, R., & Castellon-Wellington, M. (1999). *Academic language proficiency task development process* (Final Deliverable to OERI, Contract No. R305B60002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Debra P. v. Turlington, 664 f.2d 397, 6775 (5th Cir., 1981).

Elmore, R. F. (in press). Conclusion: The problem of stakes in performance-based accountability systems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems*. New York: Teachers College Press.

Elmore, R. F., & Rothman, R. (Eds.) (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Research Council.

Eyde, L. G., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., et al. (1993). *Responsible test use: Case studies for assessing human behavior*. Washington, DC: American Psychological Association.

GI Forum Image De Tejas v. Texas Education Agency, 87 F. Supp. 2d 667 (W.D. Tex. 2000).

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hoffman, R. G., & Wise, L. L. (2000). *School classification accuracy final analysis plan for the Commonwealth accountability and testing system*. Alexandria, VA: HumRRO.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Joint Committee on Testing Practices. (1988/1994). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Kane, T. J., & Staiger, D. O. (in press). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy*. Washington, DC: Brookings Institution.

Linn, R. L. (in press). Accountability models. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems*. New York: Teachers College Press.

Linn, R. L., & Haug, C. (in press). *Stability of school building accountability scores and gains* (CSE Report). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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

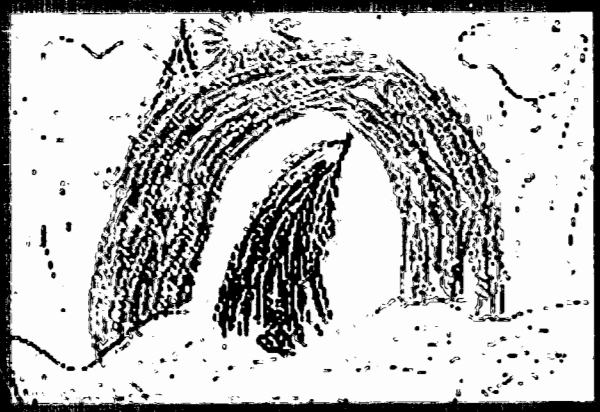
O'Day, J. (in press). Complexity, accountability, and school improvement. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems.* New York: Teachers College Press.

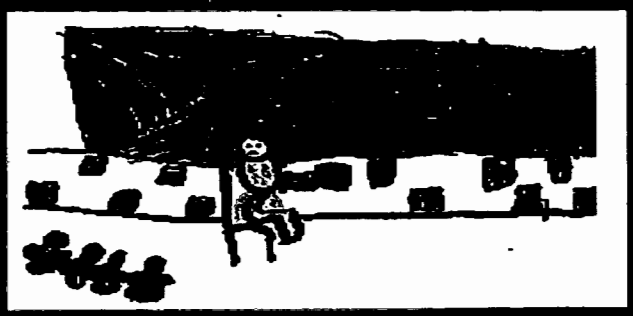
Pellegrino, J. P., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Rogosa, D. (1999a). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations* (CSE Tech. Rep. No. 509). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Rogosa, D. (1999b). *Reporting group summary scores in educational assessments: Properties of proportion at or above cut-off (PAC) constructed from Instruments with continuous scoring* (Draft deliverable). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.





BEST COPY AVAILABLE

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

[X] 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").