

VALIDATION OF USES AND INTERPRETATIONS OF STATE ASSESSMENTS

ROBERT L. LINN

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING (CRESST)

UNIVERSITY OF COLORADO AT BOULDER

*Paper prepared for the validity subcommittee of the
Technical Issues in Large-Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS) of the
Council of Chief State School Officers (CCSSO)*



THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Technical Issues in Large-Scale Assessment State Collaborative on Assessment and Student Standards

The Council's State Collaborative on Assessment and Student Standards strives to provide leadership, advocacy and service in creating and supporting effective collaborative partnerships through the collective experience and knowledge of state education personnel to develop and implement high standards and valid assessment systems that maximize educational achievement for all children.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Rick Melmer (South Dakota), President
Elizabeth Burmaster (Wisconsin), Past President
T. Kenneth James (Arkansas), President-Elect

Gene Wilhoit, Executive Director

John Tanner, Director Center for Innovative Measures
Duncan MacQuarrie and Douglas Rindone, Co-Coordiators, TILSA SCASS

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

Copyright © 2008 by the Council of Chief State School Officers, Washington, DC

All rights reserved.

TABLE OF CONTENTS

INTRODUCTION.....	1
STATE ASSESSMENT SYSTEMS	2
<i>Standards-Based Assessments</i>	2
<i>NCLB Requirements</i>	3
<i>End-of-Course Assessments</i>	3
<i>Multiple Purposes and Uses</i>	4
<i>Alternate Assessments</i>	4
PROFESSIONAL VALIDITY STANDARDS	4
<i>Perspectives of Validity Theorists</i>	6
CONSEQUENCES OF USES OF NCLB ASSESSMENTS.....	8
<i>Regular State Assessments</i>	8
<i>Alternate Assessments</i>	12
CONSEQUENCES OF END-OF-COURSE ASSESSMENTS.....	12
VALIDITY OF SCHOOL QUALITY INFERENCES FROM STUDENT ASSESSMENTS.....	12
CONCLUSION	13
REFERENCES	15

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

VALIDATION OF USES AND INTERPRETATIONS OF STATE ASSESSMENTS

State assessments of student achievement are used for a variety of purposes and the results are subject to a wide range of interpretations. Assessment results play a central role in both state and federally mandated school accountability systems. They also are used in making both high- (e.g., grade-to-grade promotion, high school graduation, school quality) and low-stakes (e.g., guidance to teachers on topics to emphasize in instruction) decisions. Justification of the uses and interpretations of assessment results requires an evaluation of the evidence supporting or refuting each particular use or interpretation. The process of marshalling and evaluating the evidence relevant to particular uses and interpretations of assessment results is called validation.

Introduction

Validity is the most fundamental consideration in the evaluation of the appropriateness of claims about uses, and interpretations of educational assessment results. Although casual discussions often refer to the validity of an assessment, it should be understood that it is the uses, interpretations, and claims about assessment results that are validated. Evidence may support the conclusion that a particular use of assessment results has good validity. That same assessment may produce results, however, that have little or no validity when interpreted or used in a different way. For example, an assessment may provide a good indication of what students know and can do in a specified content area and provide information that is useful in instructional planning, but have inadequate validity for making high-stakes decisions about individual students such as the award of a high school diploma.

Validity is a matter of degree rather than an all or none characteristic. One use of assessment results may have a high level of validity while another use has modest validity and a third use has little or no validity. The variation of the degree of validity across different uses and interpretations is an important consideration for state assessments because the results of those assessments are used in such a variety of ways and subject to such a wide range of interpretations. The degree of validity that is sufficient to justify a specified use or interpretation of assessment results is a matter of judgment. Not surprisingly, different observers may come to different conclusions about the degree of validity needed in a particular instance. The task of validation is to present the relevant evidence in the form of a coherent argument that forms the basis for a conclusion about the adequacy of the validity for the specified purpose.

Validity

- Most important consideration in the evaluation of the appropriateness of uses and interpretations of assessment results
- Matter of degree rather than all or none
- Degree of validity varies from one use or interpretation to another
- Validation involves the use of evidence in arguments that evaluate each use or interpretation of assessment results

A comprehensive validation program for state assessments would require a systematic analysis of the myriad uses, interpretations, and claims that are made. Evidence relevant to particular uses, interpretations, and claims would need to be accumulated and organized into relevant validity arguments (Kane, 2006).

The focus of this paper is on areas of validity that states have not already directly addressed as part of the NCLB peer review process. Primary attention is given to evidence regarding the consequences of uses and interpretations of state assessments. As Brennan (2006) has noted, the role of consequences in validity is a contentious topic, but as Brennan goes on to argue, consequences are necessarily a part of validity because it is “almost universally agreed that validity has to do with the proposed interpretations and uses of test scores” (p. 8). Consequences are a suitable focus for this paper, in part, because of the contentious nature of the need to consider consequences as part of validation and, in part, because validation efforts such as those mounted by states either for their own purposes or to meet the

NCLB peer review requirements, have rarely attended to consequences. Furthermore, the Standards and Assessments Peer Review Guidance (U.S. Department of Education, 2004, p. 33) explicitly requires states to consider the consequences of the state’s interpretation and use of assessments as part of the validation process.

In addition to considering consequences of the uses and interpretations of state assessments, attention will also be given to a topic that was not called for in the Peer Review Guidance, but which represents a major use and interpretation of state assessment results that needs to be validated. That is the use of student assessment results to make inferences about school quality. NCLB requires states to use student assessment results to determine whether or not schools make adequate yearly progress (AYP) each year and schools that fail to meet AYP for two or more years in a row are subject to “corrective actions.” Schools that fail to make AYP are implicitly judged to be of lower quality or less effective than schools that make AYP. The validity of that interpretation needs to be evaluated.

State Assessment Systems

With the sole exception of Iowa, states have mandated student assessment systems and implemented them prior to the enactment of NCLB, and almost all schools in Iowa have participated for a relatively long time in the administration of the Iowa Test of Basic Skills (ITBS) each year on a voluntary basis. The assessment systems were far from uniform across states, however. They varied greatly in the grades and subjects tested, in the nature of the assessments (.e.g., the proportion of constructed-response vs. multiple-choice items, the degree of customization to be consistent with states content standards), and in the reporting of results (.e.g., the metric used and whether or not results were separately reported for subpopulations of students). States also had different policies on the inclusion of students with disabilities and with limited English proficiency and differed with respect to the tolerance they had when schools failed to assess some eligible students.

Standards-based Assessments

In the 1980s and early 1990s, many states used off-the shelf, norm-referenced achievement tests for the statewide testing programs. During the last ten or fifteen years, however, states have moved away from a reliance on norm-referenced tests and have embraced standards-based assessments. This shift

came about in response to the standards movement that was encouraged by the federal government. The Goals 2000: Educate America Act and the Improving America’s Schools Act of 1994 (the precursor to NCLB) encouraged states to adopt content standards that defined the material that teachers should teach and students should learn. States were also encouraged to develop assessments that were aligned with their content standards and to set performance standards that established levels of achievement needed for a student to be considered proficient or to be at some other performance level (e.g., basic or advanced).

Not surprisingly, the content standards that states adopted differed with regard to the specific content that was covered and the grades at which particular concepts were introduced. Consequently, the degree of alignment of a norm-referenced test with state content standards varied from state to state and was generally considered inadequate without some customization (e.g., augmenting the norm-referenced test with items to cover aspects of the content standards that were not adequately addressed by the test). Alternatively, states developed requests for proposals to develop assessments that were designed to be aligned with their content standards.

The reporting of results also shifted away from percentile ranks or grade equivalent scores used with

norm-referenced tests to standards-based reports. Unlike normative score reports where scores depend on a student's achievement in comparison to that of other students, standards-based scores are reported in absolute terms (e.g., below basic, basic, proficient, or advanced performance).

NCLB Requirements

Although there is still considerable between-state variability in the overall assessment systems, NCLB has greatly increased the commonalities at least for assessments of mathematics and reading or English language arts in grades 3 through 8. States were required to adopt content standards in mathematics and reading or English language arts and they had to develop assessments that were aligned with those standards¹. Those assessments had to be administered each year to students in grades 3 through 8 and one grade in high school. Making AYP not only requires that students in the school score above set targets in both mathematics and reading or English language arts, but at least 95% of the eligible students had to be assessed in each subject.

States had to adopt student academic achievement standards that would identify at least three levels of achievement (usually called basic, proficient, and advanced) and set intermediate performance targets (called annual measurable objectives) each year that would lead to all students performing at the proficient level or above by 2014. Exclusions of students from assessments that were once commonplace are not allowed on assessments used to meet the requirements of NCLB. NCLB assessment results must be reported not only for the student body as a whole but must be separately reported for racial/ethnic subpopulations, economically disadvantaged students, students with disabilities, and students with limited English proficiency. To make AYP schools must assess at least 95% of eligible students in each subgroup that is large enough to allow disaggregated reporting in both mathematics and reading or English language arts. Furthermore, the annual measurable objective must be met in each subject for every reportable subgroup for the school to make AYP.

End-of-Course Assessments

As was already indicated, in addition to the requirement to assess students each year in grades 3 through 8, NCLB also requires states to assess

students in mathematics and reading or English language arts in at least one grade in high school. The focus, however, has been mainly on elementary and middle schools. Assessing high school students poses greater challenges than assessing elementary and middle school students because of the often substantial differences in course taking patterns of high school students. An assessment that may pose an appropriate degree of challenge for students who have had a single year of algebra may seem relatively trivial to students who have completed algebra II or an advanced placement course in calculus. Recognizing the major differences in the content covered by the variety of courses taken by high school students, some states have moved to or are considering the use of end-of-course assessments rather than generic mathematics and reading or English arts assessments.

The American Diploma Project (<http://www.achieve.org/node/604>) illustrates both the efforts that are underway to develop end-of-course assessments and the rationale for these assessments. The American Diploma Project involves a network of 29 states that are working with Achieve, Inc. that is encouraging states to use end-of-course assessments. End-of-course assessments are seen as a means not only of assessing students, but as a mechanism for promoting rigor and greater uniformity in the high school curriculum. A fact sheet regarding the algebra II end-of-course exam being developed under the auspices of the America Diploma Project, for example, claims that "using an end-of-course test would help ensure a consistent level of content and rigor in classes" (Achieve, Inc., no date, p. 1). Improving curriculum and instruction and making curriculum and instruction more uniform from classroom to classroom is an intended consequence of end-of-course assessments. Validity evidence is needed to support the claim that these intended consequences are achieved.

The uses made by states that have adopted end-of-course examinations differ in several respects. They differ in the number of courses for which examinations are used. North Carolina, for example, offers end-of-course tests in 10 subjects, while South Carolina has them for 5 subjects. End-of-course assessments are generally required for students taking the course or as part of high school graduation requirements. In some instances, such as the California Golden State exams, however, they are voluntary though results may be used to obtain an endorsed high school diploma. The Golden State Seal of Merit Diploma in California may be earned either by obtaining scores above a specified level on each the California Standards Tests in three subject areas or on Golden State exams in three content areas (<http://www.cde.ca.gov/ta/tg/sr/eligibility.asp>).

¹ Starting in 2007-08, states will have to assess students in science in at least one grade in each of three grade level spans (elementary, middle, and high school). So far, however, no use of the science assessments as part of the NCLB accountability system has been specified.

States use the results in different ways – as a percentage of the course grade, as an option for meeting graduation requirements, or in the determination of the status of a school in the state accountability system and/or for determining AYP. Proponents of end-of-course examinations argue that they help clarify what should be covered in specific courses and that they are more rigorous and challenging than generic subject matter tests that are not tied to specific high school courses.

Claims such as the one that end-of-course assessments will increase the uniformity and rigor of the high school curriculum or that they will lead to students being better prepared for college-level work need to be justified. In other words, they need to be supported by validity evidence and arguments.

Multiple Purposes and Uses

As was previously noted, assessment results are often used for more than one purpose. The state assessment results may not only be used for school accountability, but for making grade-to-grade promotion decisions. Sometimes the multiple uses are designated by the state. Florida, for example, uses assessment result for their state system of school accountability as well as for the NCLB accountability system. The Florida Comprehensive Assessment Test (FCAT) results also are used in making decisions to promote students from grade 3 to grade 4 and FCAT results enter into teacher evaluations.

In other instances uses are made of state assessments other than those designated by the state. Districts may use results in grade-to-grade promotion decisions, even though the state does not encourage that use. School principals may use student assessment results in teacher evaluation without any endorsement of that use by the state. Furthermore, uses are not limited to those made by educators. Real estate agents use results for schools to assist in selling homes in particular school districts and prospective teachers may use assessment results in choosing where they will submit job applications. Although we would not suggest that there is a need to validate claims made by real estate agents, the uses made by educators at the school and district level are worthy of attention in a validation program even though they are not uses that are promoted by the state.

Alternate Assessments

Prior to NCLB many students with disabilities were excluded from state assessments. Now, however, those students must be assessed. For the majority of students with disabilities the regular assessment is appropriate, though it may have to be administered with accommodations. The regular assessment, even with accommodations is inappropriate, however, for a small fraction of students with severe cognitive disabilities. NCLB requires that alternate assessments based on alternate standards be developed for use with the latter students. No more than 1% of students can be classified as proficient based on the alternate assessments.

The development of alternate assessments has posed a substantial challenge for states because of the small number of students involved and the great diversity of needs that those students. Nonetheless, states are expected to validate the uses and interpretations of their alternative assessments.

In addition to alternate assessments for students with the most severe cognitive disabilities, the Department of Education has also opened up the possibility that a state may, but does not have to, develop a second type of alternate assessment that is “based on modified academic achievement standards” (U.S. Department of Education, 2007, p. 20). An alternate assessment based on modified academic achievement standards must be based on the same content standards and same grade-level expectations as the regular assessment. Only students with a disability for whom the student’s individual educational program (IEP) team determines that the assessment is appropriate may take the alternate assessment based on modified academic achievement standards. No more than 2% of all students can be classified as proficient or above on the alternate assessment with modified achievement standards, or a total of no more than 3% may be classified as proficient on the basis of the two types of alternate assessments combined (U.S. Department of Education, 2007). States that choose to develop an alternate assessment based on modified academic achievement standards will need to develop validity evidence for that assessment.

Professional Validity Standards

Guidance on the conduct of a validation program is available to states from several sources. Two sources are particularly relevant in this regard. General guidance is provided by the Standards for Educational and Psychological Measurement (AERA,

APA, & NCME, 1999) jointly developed by the American Educational Research Association, the

American Psychological Association, and the National Council on Measurement in Education, hereafter referred to as the Test Standards². Guidance specific to the validation of assessments used for purposes of the No Child Left Behind ACT of 2001 (NCLB) accountability system is provided by the previously referenced Standards and Assessments Peer Review Guidance (U.S. Department of Education, 2004). The Peer Review Guidance identifies the types of evidence that states are required to provide to meet federal requirements under NCLB. Peer Review Guidance relies heavily on the Test Standards in specifying the validity evidence that states must provide for peer review..

It is generally agreed that the Test Standards (AERA, APA, & NCME, 1999) represent a broad professional consensus regarding a range of issues relevant to the evaluation of the uses and interpretations of assessments. As was true of Earlier editions, of the Test Standards (1954, 1966, 1974, 1985), the most recent edition has been widely referenced in legislation and court proceedings as an authoritative statement of technical expectations for assessments. Thus, the reliance of the Peer Review Guidance (U.S. Department of Education, 2004) on the Test Standards is not unique in that regard.

In addition to the Test Standards and the Peer Review Guidance, there are at least three other documents that are intended to provide guidance for the evaluation of the validity of assessment systems. These are the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices 2004), the Program Evaluation Standards: Second Edition, How to Assess Evaluations of Educational Programs prepared by the Joint Committee on Standards for Educational Evaluation (The Joint Committee on Standards for Educational Evaluation, 1994), and the Standards for Educational Accountability (Baker, Linn, Herman, & Koretz, 2002).

The Code of Fair Testing Practices is intended to be consistent with and reinforce the Test Standards by serving as a guide to professionals to help them ensure that tests they provide are fair to all test takers. The Program Evaluation Standards were developed in response to concerns that the 1974 version of the Test Standards did not address the uses of

assessments in program evaluation with sufficient specificity. One of four sections of the Program Evaluation Standards focuses on the accuracy of evaluations. The accuracy standards address issues of program documentation, context analysis, the analysis of quantitative and qualitative data, the validity and reliability of information, and the justification of conclusions. Finally, the Accountability Standards were a collaborative effort of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the Consortium for Policy Research in Education (CPRE) with input from professional organizations, educational institutions, and test publishers. The Accountability Standards build upon the Test Standards as well as “research findings on testing and accountability systems, and studies of best practices” (Baker, et al., 2002, p. 1).

² The three associations responsible for the *Test Standards* recently decided that it is time to revise the Test Standards once again and have appointed Barbara Plake and Lauress Wise as Co-Chairs of the revision committee. Although the revision will introduce changes and new ideas, the conceptualization of validity has evolved gradually from one edition of the *Test Standards* to the next. Hence, it seems unlikely that the revised *Test Standards* will make fundamental changes in the conceptualization of validity.

Documents Providing Guidance for the Evaluation of the Uses and Interpretations of Assessments

Standards for Educational and Psychological Testing. (1999)

Broadly recognized as an authoritative statement of professional standards for the evaluation of the uses and interpretation of Assessments.

Code of Fair Testing Practices. (2004)

The Code specifies fair practices for developers and users of test results as well as test takers and is intended to be consistent with the Test Standards

Program Evaluation Standards. (1994)

The Evaluation Standards Provide guidance for the evaluation of educational programs.

Accountability Standards. (2002)

The Accountability Standards extend the Test Standards into the area of educational accountability systems

Peer Review Guidance. (2004)

The Guidance specifies evidence that states must provide regarding their standards and assessments used to satisfy the requirements of NCLB.

For the most part the *Code of Fair Testing Practices*, the *Program Evaluation Standards*, and the *Accountability Standards* reinforce particular aspects of the Test Standards. With regard to a few issues of concern with the use and interpretation of state assessments, however, one of more of the three supplementary documents extends or clarifies the *Test Standards*. In those instances, mention will be made of these other sources, but in most instances the analysis will rely primarily on the Test Standards and the *Peer Review Guidance*.

The chapter on validity in the *Test Standards* begins with the following definition. "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the 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" (AERA, APA, & NCME, 1999, p. 9). Later in the validity chapter the *Test Standards* address the issue of consequences. After noting that assessments are "commonly administered in the expectation that some benefit will be realized from the intended use of the scores" the *Test Standards* go on to conclude that a "fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized" (1999, p. 16).

The *Peer Review Guidance* identifies several categories of validity evidence with reference to the *Test Standards*. In particular, the *Peer Review Guidance* requires states to collect evidence related to assessment content, evidence based on relationships

of assessments with other variables, evidence based on student response processes and evidence related to the internal structure of assessments. In addition, it calls for evidence regarding the consequences of uses and interpretations of state assessments. Although not discussed under the general validity heading, the *Peer Review Guidance* also requires states to submit evidence of the alignment of their assessments with their content standards and the alignment results clearly have a bearing on aspects of validity related to the content of the assessment.

States submitted multiple documents containing a considerable amount of evidence in response to the *Peer Review Guidance* and have received feedback from the U.S. Department of Education regarding their submissions. Although the submitted documentation adequately covered most of the validity issues that the U.S. Department of Education expected states to address, evidence regarding the consequences of the uses and interpretations of state assessments was generally absent (see, for example, Linn, in press).

Perspectives of Validity Theorists

As can be seen either by reviewing the various editions of the *Test Standards* or by an analysis of the writing of leading measurement experts and validity theorists in major publications such as the four editions of *Educational Measurement*, the concept of validity has evolved over time. Until the early 1950s validity was generally equated either with the

relationship of test scores to an external criterion measure or with the degree to which a test measured what it was intended to measure. Thus, Gulliksen (1950) stated that “validity of a test is the correlation with some criterion” (p. 88) and Cureton’s (1951) chapter on validity in the first edition of *Educational Measurement* equated validity with “how well a test does what it is supposed to do” (p. 621).

An expanded view of validity was articulated in the first edition of the *Test Standards* which were then called *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, 1954) where four types of validity (content, concurrent, predictive, and construct) were described. Concurrent validity was only a slight variation of the prevailing predictive validity which distinguished between correlations of a test with a criterion measure where the latter measure was obtained at nearly the same time rather than at some later point in time. Content validity also represented a perspective on validity that was familiar at that time, particularly with regard to educational achievement tests. Construct validity, on the other hand, was a relatively new idea when introduced in 1954 *Technical Recommendations*. Notions of construct validity that were introduced in the *Technical Recommendations* were elaborated more fully by the following year in a classic article by Cronbach and Meehl (1955), both of whom were members, with Cronbach as chair, of the committee that developed the *Technical Recommendations*.

The Cronbach and Meehl (1955) paper had a lasting impact on subsequent thinking about and discussions of validity. Although construct validity was initially thought to be more relevant to the assessment of psychological characteristics such as anxiety or extroversion than to the assessment of educational achievement, construct validity gradually came to be widely accepted as relevant to assessments of educational achievement. The acceptance of construct validity came about, in part, because leading theorists such as Cronbach (1971) and Messick (1975) continued to promote the concept of construct validity and, in part, because they illustrated how it could be made more practical and

how construct validity considerations necessarily entered into an evaluation of the uses and interpretations of any assessment. Cronbach’s (1971) emphasis on the validation of interpretations of assessment results laid the foundation for more of a unitary view of validity and the subordination of content-related and criterion-related evidence to a construct validity perspective.

By the time the fourth edition of the *Test Standards* (AERA, APA, & NCME, 1985) was published this view of validity had become dominant. “The concept [of validity] refers to the appropriateness, meaningfulness, and usefulness of the specific inferences that are made from test scores. Test validation is the process of accumulating evidence to support each inference” (AERA, APA, & NCME, 1985, p. 9). Shortly after the publication of the 1985 *Test Standards*, however, Messick (1989) argued that the *Test Standards* had not gone far enough in moving toward a unified view of validity in which notions of construct validation were paramount.

Messick’s 1989 chapter, which Shepard (1993) described as the “most cited authoritative reference on the topic” (p. 423), began with the following definition. “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the **adequacy** and **appropriateness of inferences and actions** based on test scores or other modes of assessment” (Messick, 1989, p. 13, emphasis in original). Messick elaborated his definition of validity in a two-by-two table corresponding to the adequacy/appropriateness and inferences/actions distinctions of the definition. The rows of the table distinguish two types of support for validity claims—the **evidential** basis and the **consequential** basis that are used to support claims of adequacy and appropriateness. The two columns of the table distinguish between **interpretations** of assessment results (e.g., 55% of the fourth grade students are proficient in mathematics) and **uses** of results (e.g., award of a high school diploma). Although uses generally related to interpretations, implicitly if not explicitly, the uses involve actions of some type whereas interpretations may not.

Views of Validation in the last three editions of *Educational Measurement*

“One validates, not a test, but an interpretation arising from a specified procedure” (Cronbach, 1971, p. 447).

“Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test score or other modes of assessment” (Messick, 1989, p. 13, emphasis in original).

“To validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation and use. The evidence needed necessarily depends on the claims being made. Therefore, validation requires a clear statement

of the proposed interpretations and uses” (Kane, 2006, p. 23

As was implied by Brennan’s (2006), the inclusion of consequences as part of validation is controversial (see, for example, Linn, 1997; Mehrens, 1997; Popham, 1997; Shepard, 1997). However, as Kane (2006, p. 54) has recently noted there is, in fact, nothing new about giving attention to consequences in investigations of validity. What is relatively new is the salience of the topic and the breadth of the reach that is no longer limited to immediate intended outcomes (e.g., students perform better in classes following the use of a placement test). The inclusion of broader social consequences and the inclusion of negative unintended as well as positive intended consequences led to objections by some measurement experts (see, for example, Green, 1990; Mehrens, 1997; Popham, 1997; Wiley, 1991).

The objections have more to do with the question of whether consequences should be a part of validity than they do with the question of whether or not consequences are relevant to an evaluation of a use or interpretation of assessment results. There is broad consensus regarding the importance of investigations of consequences as part of the overall evaluation of particular interpretations and uses of assessment results (Cronbach, 1980; 1988; Kane, 2006; Linn, 1994; 1997; Linn, Baker, & Dunbar, 1991; Shepard, 1993; 1997), but some authors have maintained that such an evaluation is outside the scope of validity.

Regardless, of whether consequences are considered a part of validity or as part of a broader evaluation of test uses and interpretations, the Peer Review Guidance makes it clear, as has already been noted, that states are expected to attend to consequences. “In validating an assessment, the State must ... consider the consequences of its interpretation and use. Messick (1989) points out that these are different functions and that the impact of an assessment can be traced either to an interpretation or to how it is used. Furthermore, as in all evaluative endeavors, States must attend not only to the intended outcomes, but also to unintended effects” (U.S. Department of Education, 2004, p. 33). The Department’s position with regard to responsibility for evaluating consequences of the uses and interpretations of assessments seems consistent with the prevailing view that validation should be a broadly conceived evaluative process.

Kane (2006) outlined an approach to validation that builds upon the evaluative approach that was articulated by Messick (1989). He conceptualized validation as the process of developing two types of argument that he calls the interpretive argument and the validation argument. The “interpretive argument

specifies the proposed interpretations and uses of test results by laying out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances” (Kane, 2006, p. 23). The evaluation of the interpretive argument is called the validation argument. The validation argument brings evidence and logical analysis together for the purpose of evaluating the claims and propositions of the interpretive argument.

The November, 2007 issue of *Educational Researcher* featured an article by Lissitz and Samuelsen (2007) that proposed an alternative conceptualization of validity that departs in radical ways from the view espoused in the last two editions of the *Test Standards* (1985, 1999) and in the chapters on validity in the last three editions of *Educational Measurement* by Cronbach (1971), Messick (1989) and Kane (2006). Lissitz and Samuelsen argue that content considerations are the core of validity. Rejecting the idea that validity depends on the uses and interpretations of assessment results, they argue that validity is something that inheres in the assessment and that validity is determined by evidence related to internal characteristics of content, reliability, and latent processes. Relationships to other variables and considerations of the consequences of assessment uses and interpretations are seen as relevant to an evaluation of the assessment, but Lissitz and Samuelsen consider such “external” considerations outside the realm of validity.

The Lissitz and Samuelsen article is useful for calling attention to the importance of content considerations, but as is clear from the analyses by five well-known measurement specialists (Embretson, 2007; Gorin, 2007; Mislevy, 2007; Moss, 2007; and Sireci) that content considerations do not provide a sufficient basis for evaluating the validity of assessments. Embretson (2007) and Gorin (2007) both argue that considerations of content are not sufficient basis for evaluating validity. The meaning, and hence the interpretation, of assessment results requires a more unified approach to validity that includes the consideration of constructs and relationships to other variables in addition to consideration of content. In a similar vein, Moss (2007) and Sireci (2007) highlight the need to think of validity as dependent upon the uses and interpretations of assessment results rather than as a property of the assessment independent of the uses that made be made of the results or the ways in which assessment results may be interpreted.

Consequences of Uses of NCLB Assessments

Discussions of the consequences of assessments usually distinguish between intended positive consequences and unintended negative consequences. Although both types of consequences are relevant to the evaluation of the validity of an assessment program, what is a positive effect for one observer may be a negative effect for someone else (see, for example, Mehrens, 1998). Narrowing the curriculum provides an obvious example. Narrowing may be viewed as a positive outcome by those who want instruction to have a sharper focus on the material in the state content standards, but it may be viewed as a negative outcome for those that worry that important concepts and content areas (e.g., history or music) may be short changed.

Because there is not universal agreement about whether particular outcomes are positive or negative, the discussion below is not separated into intended positive and unintended negative consequences. Whether they are combined or treated separately conceptually, it clearly is undesirable to make a sharp distinction in data collection. Questionnaires are often used to collect evidence regarding the consequences of an assessment program. Questions that sharply distinguish positive and negative effects can be leading and should be avoided. It is better to have questions stated in a neutral fashion so that respondents are free to express either positive or negative opinions. Neutral statements also can reduce the confounding effects of social desirability.

Regular State Assessments

State assessments that are used for NCLB as well as for state defined purposes are intended to have a number of consequences. They are intended to focus instruction on the knowledge, skills, and understandings that are specified in the state content standards. They are intended to motivate greater effort on the part of students, teachers, and administrators and they are intended to lead to improved learning and to a closing of gaps in

achievement among subgroups of students (see, for example, Lane & Stone, 2002, for a discussion of these and other intended consequences of state assessments).

It is far easier to talk about the desirability of obtaining evidence that uses of an assessment have particular consequences than it actually is to be able to make a convincing case one way or the other. As Mehrens (1998) has noted “consequence implies a cause and effect relationship” (p. 5), but evidence can seldom, if ever, be gathered in a manner that unambiguously demonstrates a causal connection (see also, Reckase, 1998). Although it may be impossible to obtain evidence that demonstrates an irrefutable causal connection between the uses of assessments and the various intended positive outcomes, there are a variety of ways that evidence can be obtained that makes such causal links more or less plausible.

Lane and her colleagues (Lane & Stone, 2002; Lane, Parke, & Stone, 1998) provided a framework (see box below) for evaluating the consequences of assessments that builds on several years of research investigating the consequences of Maryland State Performance Assessment Program (MSPAP) that was reported in a series of technical reports (Lane, Parke, & Stone, 1999; Lane Parke, Stone Cerrillo, & Hansen, 2000a; 2000b, 2000c). Their approach is consistent with Kane’s (2006) conceptualization of validity as argument. They began by identifying a set of propositions that are central to the interpretive argument (e.g., “school administrators and teachers are motivated to adapt the instruction and curriculum to the standards” and “students are motivated to learn as well as to perform their best on the assessment”) (Lane & Stone, 2002 p. 27). For each proposition they then identified relevant types of consequential evidence and data sources (e.g., teacher, administrator, and student questionnaires; classroom artifacts such as classroom assessments; and assessment results) that can be used to obtain that evidence.

Framework used by Lane and her colleagues to evaluate the consequences of the uses and interpretations of assessments

- Identification of a set of propositions about consequences that is central to an interpretive argument
 - (e.g., School administrators and teachers are motivated to adapt instruction and curriculum to the content standards)
 - (e.g., students are motivated to learn as well as to perform their best on the assessment)
- Teacher and student questionnaires and interviews regarding motivation and instructional practices
- Collection of multiple indicators of student achievement

For example, using teacher and principal questionnaires, Lane and Stone (2002) suggested that teacher and principal familiarity with content standards and state assessments can be evaluated. The questionnaires can also be used to probe teacher beliefs and attitudes toward standards and assessments, and to assess teacher and principal morale. The match of instruction and curriculum to content standards can be assessed both by the use of questionnaires and by the collection of classroom

artifacts such as classroom assessment tasks and test preparation materials. Changes in instruction reported by teachers and students can be related to changes in performance on state assessments (Lane & Stone, 2002, p. 27). Lane and Stone (2002) summarized illustrative propositions about assessment consequences the types of evidence that can be used to evaluate those propositions and the sources of data to be collected in a table. An adaptation of that table is presented in Table 1.

Table 1
Validity Evidence to Support Stated Propositions About Assessment Consequences*

Proposition and Stakeholder	Evidence of Consequences	Data Source
Administrators and teachers are motivated to adapt instruction and curriculum to the standards	Familiarity with standards and assessments	Teacher and principal questionnaires
Professional development support is being provided and resources are available	Nature of professional development support	Teacher and principal questionnaires
	Amount of professional development support	
Instruction and curriculum will be adapted to the standards	Degree to which instruction and classroom assessments reflect state standards	Teacher and principal questionnaires
	Degree to which instruction and classroom assessment strategies reflect the standards	Classroom artifacts such as instruction and classroom assessment tasks, scoring rubrics, and test preparation activities
Students are motivated to learn as well as to perform their best on the assessment	Beliefs of students toward the standards, instruction, and assessment	Teacher and student questionnaires
	Effort students put forth on the assessment	
Improved performance is related to changes in instruction	Relationship between school and class gains with the motivation, instruction and classroom assessment and professional development variables	Teacher, principal and student questionnaires over time
		Class and school performance over time

* Adapted from Lane and Stone (2002)

A few examples of questions used by Lane and her colleagues are shown in Table 2. Although the full questionnaires are not available on the web, they may be obtained from Suzanne Lane at the University of Pittsburgh. As can be seen from the illustrative questions in Table 2, the statements allow respondents to indicate that they believe an assessment has had either a positive or a negative effect.

Table 2				
Illustrative Questions Asked of Teachers and Principals in Studies of MSPAP by Lane and her Colleagues*				
(Teacher Question) To what extent has MSPAP had a positive or negative impact on teacher morale in your school?				
Very Negative Impact	Somewhat Negative Impact	Somewhat Positive Impact	Very Positive Impact	
(Teacher Question) MSPAP is a useful tool for helping me make positive changes in my instruction.				
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	
(Teacher Question) How well do you think the mathematics tasks on MSPAP allow your students to show what they know?				
Not at all	A little	Moderately	Very	
(Principal Question) MSPAP is a useful tool for helping teachers in my school make positive changes in their instruction.				
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	
(Principal Question) Some schools have found ways to raise MSPAP scores without really improving instruction.				
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	
* MSPAP refers to the Maryland School Performance Program which was the assessment that was investigated by Lane and her colleagues in their research on the effects of the uses of a state assessment. Questions are quoted from the questionnaires provided by Suzanne Lane (personal communication).				

Although they were designed and are generally used for purposes other than the evaluation of the consequences of assessments, the Surveys of the Enacted Curriculum (SEC) conducted under the auspices of the Council of Chief State School Officers in collaboration with states, the American Institutes for Research, and the Wisconsin Center for Education Research (Smithson and Blank, 2007) provide a tool that could have substantial value in evaluations of the consequences of state assessments. The SEC use a procedure developed by Porter and Smithson (2001; Porter, 2002) to study the alignment

of content standards, instruction, and assessments. The SEC provide information about instructional practice and the influence of educational policies on practice. The measurement of alignment of instructional practice with both content standards and state assessments is relevant to an evaluation of the effects of state assessments.

The SEC Instructional practice scales include Demonstrate Understanding, Make Connections, and Analyze Information as well as a scale that assesses the Influence of Standards on Practice. Information

from scales such as these can be linked to state assessments and used to monitor changes from year-to-year (Smithson & Blank, 2007). The SEC provide a way to collect data on what teachers' instructional practice that can be related directly to a state's content standards and assessments (Smithson & Blank, 2007). See [a;sphttp://www.ccsso.org/projects/Surveys_of_Enacted_Curriculum](http://www.ccsso.org/projects/Surveys_of_Enacted_Curriculum).

In his American Educational Research Association Division D, Vice Presidential Address, Mehrens (1998) reviewed the available evidence regarding the consequences of assessments. Although not presented as a framework for new data collection, he summarized the evidence in terms of five broad categories that might be used as a framework for posing questions about consequences of assessments and collecting relevant evidence. The five categories derived from Mehrens' review are (1) curricular and instructional reform, (2) teacher motivation and stress, (3) student motivation and self concept, (4) changes in student achievement, and (5) public awareness of student achievement.

Curricular and instructional reform. As suggested by Lane and Stone (2002), both teacher questionnaires and the collection of instructional artifacts such as classroom test items can provide relevant evidence regarding intended reforms in curriculum and instruction. Several of the studies reviewed by Mehrens (1998) (e.g., Koretz, Barron, Mitchell & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Rafferty, 1993; Madaus, West, Harmon, Lomax, & Viator, 1992; Miller, 1998; Stecher & Mitchell, 1996) used teacher questionnaires to obtain evidence regarding perceived effects of assessments on curriculum and instruction.

Other studies reviewed by Mehrens relied on interviews (e.g., Khattri, Kane, & Reeve, 1995), a combination of interviews and questionnaires (e.g., Smith, et al., 1997) or focus groups (e.g., Chudowsky & Behuniak, 1997). The collection of instructional artifacts as suggested by Lane and Stone (2002) appears to be less common according to the set of studies reviewed by Mehrens, but was used effectively in the work of Lane and her colleagues in the Maryland MSPAP studies that were conducted after the Mehrens (1998) review.

Teacher motivation and stress–student motivation and self concept. Questionnaire surveys of teachers and principals have been the most common way of investigating the perceived effects of assessments on teacher motivation and feelings of stress as well as student motivation and self concept (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Stecher, 1996; Lane & Parke, 1997; Miller, 1998). Qualitative studies have also been conducted, however (Kane, Khattri, Reeve,

& Adamson, 1997; Smith & Rottenberg, 1991). Sometimes teacher reports are relied upon to evaluate not only their perceptions of their own motivation, but their perceptions of the effects assessments have on student motivation. As Mehrens (1998) notes, however, teacher reports of effects of assessments on student motivation while possibly accurate reflections of their beliefs do not provide direct evidence of actual effects on students. Student questionnaires, student focus groups or student interviews can provide more direct evidence of student reactions to assessments.

Student achievement. Since assessment programs are expected to contribute to improved learning it is natural that trends in achievement provide one source of evidence that is relevant. Similarly, trends in the gaps in achievement among racial/ethnic groups are relevant to the proposition that the combination of standards, assessments, and the disaggregated reporting of results will reduce the magnitude of achievement gaps.

A recent study examined trends in achievement on state assessments for all 50 states (Center on Education Policy, 2007, available at <http://www.cep-dc.org/>). The Center on Education Policy (CEP) study also examined the trends in gaps in achievement between black and white students, between Hispanic and white students, and between low income and not low income groups. Although trend results were not available for all states due to recent changes in the assessment programs in some states, and the results were not uniform across states where trend results were available, more states showed increases in achievement since NCLB became law than had flat or declining achievement. For the states where gaps trends could be ascertained, there were also more states where the gaps decreased than where they increased or remained unchanged.

Although the overall results based on state assessments are positive, the report makes it clear that it is not possible to make a casual attribution of either the tendency for there to be gains in achievement or for the gaps to narrow to NCLB, or, for that matter, to any particular intervention such as the adoption of state content standards or the use of state assessments. As the 2007 CEP report acknowledges, there are a number of plausible explanations for the trends that were observed in state assessment results. The possibilities identified in the report include the possibility that there has been increased learning, that there has been a narrow teaching to the test, that the tests have become more lenient, or that the composition of the student population has changed (CEP, 2007, pp. 41-42).

One way to try to evaluate whether apparent improvements in state test results reflect increased learning or simply artificially inflated test scores is to

investigate the degree to which the results generalize to other indicators of student achievement. The trends on a state assessment can be compared to trends on other tests such as a norm-referenced test, a college admissions test, or the National Assessment of Educational Progress (NAEP) (see, for example, Hamilton, 2003; Koretz, 2005 for discussions of this issue).

NAEP trend results were compared to state assessment trends by CEP (2007). The comparisons showed that the gains found on the many state assessments were not reflected in the changes in NAEP results. NAEP and state assessment results often diverged and the correlation of the two sets of results across states were relatively low in both reading and mathematics at both grades 4 and 8. Although it was concluded that “NAEP results should not be used as a ‘gold standard’ to negate or invalidate state test results” (CEP, 2007, p. 61), the lack of agreement does raise questions about the reasons for the lack of generalizability of the state trends.

Alternate Assessments

One of the strongest rationales for the creation and use of alternate assessments for students with severe cognitive disabilities is that the inclusion of those students in the assessment and accountability system will encourage greater attention to the

academic learning of those students. The focus on the teaching of academic skills to students with severe cognitive disabilities that is encouraged by alternate assessments is presumed to be beneficial for the students in question and to provide them with academic knowledge and skills that they would not otherwise have. An investigation of the consequential aspects of validity of alternate assessments would start with an evaluation of the extent to which these expected positive effects of alternate assessments are realized in practice.

Focus groups involving teachers and parents as well as interviews of teachers and of parents could help identify the perceived changes in the education of students with disabilities that stem from the inclusion of these students in alternative assessments. Parents are a particularly relevant source of information for students with disabilities because of their active involvement as members of IEP teams.

Comparisons of alternate assessment results obtained over the course of several years could provide an indirect indication of change in curriculum and instruction for participating students. Tracking changes in assessment results could also be used as an indicator of the effects of alternate assessment on the achievement of this population of students.

Consequences of End-of-Course Assessments

Because they are more narrowly focused than, say, a general high school mathematics assessment that may be expected to cover content dealing with numbers and operations, measurement, geometry, algebra, and probability and statistics, end-of-course assessments can go into greater depth in the circumscribed content of the course. As was previously noted, end-of-course assessments are generally expected to increase the focus and rigor of high school courses. A natural starting place for an evaluation of the consequences of end-of-course assessments is with a careful examination of the curriculum and instruction offered in the courses in question.

Questionnaire surveys of teachers and of students and the collection of instructional materials such as classroom assignments and tests could be used to provide evidence regarding the scope and rigor of a course. Ideally, a comparison would be made of the rigor of the curriculum and instruction in the courses before and after the introduction of an end-of-course assessment. In most cases, however, it is unlikely that surveys will be conducted prior to the introduction of an end-of-course assessment. Thus, in practice it is

likely to be necessary to ask participants to provide retrospective reports to have a basis of comparing curriculum coverage and instruction before and after the end-of-course assessment becomes operational. Focus groups provide an alternative to or a supplement to the use of questionnaires. Getting information about changes in curriculum and instruction that are linked to the end-of-course assessments may be accomplished more readily in some cases through the use of focus groups than the use of structured questionnaires.

The particular effects of end-of-course assessments are likely to depend on the uses that are made of the results. For example, the impact may be different for assessments that are part of graduation requirements than ones that only contribute a small percentage to course grades. It is plausible, however, that end-of-course assessments might discourage some students from taking the courses where the assessments are given when the courses are not required for graduation. On the other hand, if passing an end-of-course assessment is required for graduation, it may affect graduation and dropout rates. Thus,

information about changes in course taking patterns, in dropout rates, and in graduation rates should be

monitored as part of the evaluation of the consequences of end-of-course assessments.

Validity of School Quality Inferences from Student Assessments

The use of student assessment results to identify schools that need improvement and are therefore subject to various types of corrective actions or sanctions while other schools are identified as making adequate yearly progress rests on an implicit assumption that the observed school-to-school differences in student achievement are due to differences in school quality. For example, an inference is made in the NCLB accountability system that a school that makes adequate yearly progress (AYP) is better or more effective than a school that fails to make AYP (Linn, 2006). The validity of the school quality inference needs to be evaluated.

If school A makes AYP while school B fails to make AYP in a given year, it could be that school A is more effective than school B. But, it may be that students in school A were higher achieving in earlier years than students in school B. Moreover, it might also be that school B is serving a large number of students with disabilities or a large number of English language learners while school A has few, if any, students facing such challenges.

Evaluating the validity of inferences about the quality of schools from test-based accountability results requires the elimination of potential explanations of the observed student test results other than differences in school quality. Ruling out alternative explanations for differences in the achievement of students in different schools is poses the biggest challenge to validating the inferences about differences in school quality. As Raudenbush (2004a) has argued, the conclusion that school A is more effective than school B requires an inference that the schools and their instructional programs have caused the observed differences in achievement. The validity of such a causal inference, however, depends on the degree to which many competing explanations can be eliminated. Differences in achievement at the start of the school year, for example, would provide an explanation of the differences in test performance at the two schools in the spring that resulted in one school making AYP while the other school failed to make AYP.

There are many plausible explanations other than differences in school quality that might explain the differences in student performance on tests administered toward the end of the school year. For example, students at the school with the higher scores on the state assessment might have received more educational support at home than students at school

B. The student bodies attending different schools can differ in many ways that are related to performance on tests, including language background, socio-economic status, and prior achievement.

The difficulty in justifying inferences about school quality from assessment results at a single point in time has contributed to the interest in growth models. The longitudinal tracking of individual students across years provides a stronger basis for eliminating competing explanations for school differences in performance other than differential school quality than do accountability systems that rely on current status measures. Prior achievement during the span of grades for which students test results are included in the longitudinal tracking is no longer a viable explanation for differences among schools because that prior achievement is taken into account in analyses of the longitudinal student achievement data.

Although the use of growth models to judge school quality makes it possible to eliminate many alternative explanations for between school differences and thereby make the causal conclusion that differential school effectiveness explains the performance differences, strong causal claims still are not warranted (see, for example, Raudenbush, 2004b; Rubin, Stuart, & Zanutto (2004). Rubin, Stuart and Zanutto (2004) argued that growth model results “should not be seen as estimating causal effects of teachers or schools, but rather as descriptive measures” (p. 113).

Direct evidence regarding instruction in the schools needs to be obtained to evaluate the implicit assumption that schools with different student achievement outcomes differ in terms of the quality of their instructional offerings. This is needed regardless of whether the school accountability relies on measures of current status or on measures of student growth. Surveys of teachers and principals could be used along with direct observations and the collection of instructional artifacts to obtain evidence regarding the nature of and quality of instruction in schools that are rated high or low based on student performance. For example, classroom assessments might be collected from a sample of schools that do and do not make AYP and the relationship of the classroom assessment tasks could be compared to the state assessment in terms of content coverage and the level of cognitive demand of the assessment tasks. Teacher reports of instructional coverage of the

content of the standards could also be evaluated for | the two categories of schools.

Conclusion

Validation of the myriad uses and interpretations of state assessment results is a non-trivial undertaking. Although states have made substantial progress in accumulating evidence that is relevant to the evaluation of the validity of the uses and interpretations of their assessments, questions about the consequences of the uses of assessment results have been given little, if any, attention despite the fact the *Peer Review Guidance* requires that evidence be provided regarding the consequences of the uses and interpretations of state assessments.

There are a number of reasons that consequences have largely been ignored in validity investigations conducted or commissioned by states. The inclusion of consequences as a legitimate focus of validity investigations is controversial. Even if it is agreed that evidence of consequences should be gathered, either within a validity framework or as part of a more general evaluation of the uses and interpretations of assessment results, there are many

challenges that must be confronted. The implicit causal connection that underlies a claim that an assessment has had a particular consequence presents a major challenge. It is unrealistic to expect that a causal connection can be unambiguously established. The best that can be done is to present evidence that buttresses the plausibility that an assessment has a particular consequence.

There are a variety of techniques that can provide evidence that is relevant to judging the plausibility that assessment results have a specified set of effects. Questionnaires, interviews, observations, focus groups, the collection of data of record (e.g., course-taking patterns, graduation and dropout rates), and the collection of instructional artifacts (e.g., student assignments and classroom tests) can be used to collect the needed evidence. Once collected that evidence needs to be used in the development of a coherent validity argument.

REFERENCES

- Achieve, Inc. (No Date). American diploma project: Algebra II end-of-course exam. Fact sheet. <http://www.achieve.org/node/842>.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- 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.
- American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.
- Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). *Standards for educational accountability systems* (CRESST Policy Brief 5). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Brennan, R. L. (2006). Perspectives on the evolution and future of educational Measurement. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). (pp. 1-16). Westport, CT: American Council on Education/Praeger.
- Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Center on Education Policy.
- Chudowsky, N., & Behuniak, P. (1997, March). *Establishing consequential validity for large-scale performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp443-507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement*, 5, 99-108.
- Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer, & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621-694). Washington, DC: American Council on Education.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure. *Educational Researcher*, 36(8), 449-455.
- Goals 2000: Educate America Act of 1994, Public Law 103-227, Sec. 1 et seq. 108 Stat. 125 (1994).
- Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, 36(8), 456-462.
- Green, B. F. (1990). A comprehensive assessment of measurement. *Contemporary Psychology*, 35, 850-851.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, Wiley.
- Hamilton, L. (2003). Assessment as a policy tool. In R. L. Floden (Ed.), *Review of Research in Education*, 27, 25-68.

- Improving America's Schools Act of 1994, Public Law 103-382, Sec. 1 et seq. 108 Stat 35424 (1994).
- Joint Committee on Standards for Educational Evaluation. (1994). *The Program evaluation standards: 2nd Edition. How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage Publications.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Author.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). (pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Kane, M. B., Khattri, N., Reeve, A. L., & Adamson, R. J. (1997). *Assessment of student performance*. Washington, DC: Studies of Education Reform, Office of Educational Research and Improvement, U.S. Department of Education.
- Khattri, N. Kane, M. B., & Reeve, A. L. (1995). How performance assessments affect teaching and learning. Research Report. *Educational Leadership*.
- Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing*. Yearbook of the National Society for the Study of Education (pp. 99-118), Vol. 104, Part I.
- Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. MR-792-PCT/FF. Santa Monica, CA, RAND.
- Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: Perceived effects of the Maryland school performance assessment program*. CSE Technical Report 409. Los Angeles, CA: UCLA, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Lane, S. & Parke, C. (1996, April). *Consequences of a mathematics performance assessment and the relationship between the consequences and student learning*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Lane, S., Parke, C. S., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational measurement: Issues and practice*, 17(2), 24-27.
- Lane, S., Parke, C. S., & Stone, C. A. (1999, March). *MSPAP impact study: Vol. I: Mathematics*. U.S. Department of Education, Assessment Development and Evaluation Grants Program (CFDA 84.271) for the Maryland Assessment System Project.
- Lane, S., Parke, C. S., Stone, C. A., Cerrillo, T. L., & Hansen, M. A. (2000a, September). *MSPAP impact study: Language arts (reading and writing)*. U.S. Department of Education, Assessment Development and Evaluation Grants Program (CFDA 84.271) for the Maryland Assessment System Project.
- Lane, S., Parke, C. S., Stone, C. A., & Cerrillo, T. L. (2000b, November). *MSPAP impact study: Social studies*. U.S. Department of Education, Assessment Development and Evaluation Grants Program (CFDA 84.271) for the Maryland Assessment System Project.
- Lane, S., Parke, C. S., Stone, C. A., Cerrillo, T. L., & Hansen, M. A. (2000c, December). *MSPAP impact study: Science*. U.S. Department of Education, Assessment Development and Evaluation Grants Program (CFDA 84.271) for the Maryland Assessment System Project.
- Lane, S., & Stone, C. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational measurement: Issues and practice*, 21(1), 23-30.
- Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23, no. 9, 4-14.
- Linn, R. L. (1997). Evaluating the validity of assessments, *Educational Measurement: Issues and Practice*, 16(2), 14-16.
- Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education*, 19, 5-15.
- Linn, R. L. (in press). *Validity and reliability of student assessments*. Washington, DC: The Urban Institute.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.

- Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12*. Executive Summary. National Science Foundation study. Chestnut Hill, Boston College Center for the Study of Testing, Evaluation, and Educational Policy.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Educational Policy Analysis Archives*, 6(13). 1-30.
- Messick, S. (1975). The standard problem: Meaning and value in measurement and evaluation. *American Psychologist*, 30, 955-966.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, 3rd ed. (pp. 13-103). New York: Macmillan.
- Miller, M. D. (1998). *Teacher uses and perceptions of the impact of statewide performance-based assessments*. Washington, DC: Council of Chief State School Officers, State Education Assessment Center.
- Mislevy, R. J. Validity by design. *Educational Researcher*, 36(8), 463-469.
- Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, 36(8), 470-476.
- No Child Left Behind Act of 2001*, Pub. Law No. 107.110.
- Popham, W. J. (1997). Consequential validity: Right concern – wrong concept, *Educational Measurement: Issues and Practice*, 16(2), 9-13.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in practice. *Educational Researcher*, 31, 3-14.
- Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators*. (CPRE Research Report Series No. RR-048). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.
- Rafferty, E. A. (1993). *Urban teachers rate Maryland's new performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA, April.
- Raudenbush, S. W. (2004a). *Schooling, statistics, and poverty: Can we measure school improvement?* The ninth annual William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.
- Raudenbush, S. W. (2004b). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 120-129.
- Reckase, M. D. (1998). Consequential validity from the test developers' perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.
- Smith, M. L., Noble, A., Heineche, W., Seek, M., Parish, C., Cabay, M., Junker, S., Haag, S., Tayler, K., Safran, Y., Penley, Y., & Bradshaw, A. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona student assessment program (ASAP): Equity and teacher capacity building*. CSE Technical Report 425. Los Angeles, CA: UCLA National Center for Research of Evaluation, Standards and Student Testing.
- Smith, M. L. & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11.

- Smithson, J. & Blank, R. (2007). *Indicators of quality of teacher professional development and instructional change using data from surveys of the enacted curriculum: Findings from NSF MSP-RETA Project*. Washington, DC: Council of Chief State School Officers. (February).
- U.S. Department of Education. (2004). *No Child Left Behind. Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, April 28.
- U.S. Department of Education. (2007). *Not Child Left Behind. Modified academic achievement standards: Non-regulatory guidance draft*. April.
- Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: a volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.