

Executive Summary

A FRAMEWORK FOR EXAMINING VALIDITY IN STATE ACCOUNTABILITY SYSTEMS

A PAPER IN THE SERIES:

**IMPLEMENTING THE STATE ACCOUNTABILITY SYSTEM
REQUIREMENTS UNDER THE
NO CHILD LEFT BEHIND ACT OF 2001**



February 2004



The Council of Chief State School Officers (CCSSO) is a bipartisan, 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.

DIVISION OF STATE SERVICES AND TECHNICAL ASSISTANCE

The Division of State Services and Technical Assistance supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing and using information for decision-making; development of assessment resources; creation of high-quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success. The division combines existing activities in the former Resource Center on Educational Equity, State Education Assessment Center, and State Leadership Center.

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

The State Collaborative on Assessment and Student Standards (SCASS) Program was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects.

SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation. A total of forty-three states and three extra-state jurisdictions participated in one or more of the eleven projects offered during the project year 2003-2004.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Ted Stilwill (Iowa), President

David P. Driscoll (Massachusetts), President-Elect

Michael E. Ward (North Carolina), Vice President

G. Thomas Houlihan, Executive Director

Julia Lara, Deputy Executive Director,
Division of State Services and Technical Assistance

Rolf Blank, Director of Education Indicators Programs and Coordinator
Accountability Systems and Reporting (ASR) SCASS

Jan Sheinker, Coordinator
Comprehensive Assessments Systems for ESEA Title I (CAS) SCASS

COUNCIL OF CHIEF STATE SCHOOL OFFICERS
ONE MASSACHUSETTS AVENUE, NW, SUITE 700
WASHINGTON, DC 20001-1431

(202) 336-7000
FAX (202) 408-8072
www.ccsso.org

Call (202) 336-7016 for additional information on CCSSO publications

EXECUTIVE SUMMARY

A FRAMEWORK FOR EXAMINING VALIDITY IN STATE ACCOUNTABILITY SYSTEMS

A PAPER IN THE SERIES:
IMPLEMENTING THE STATE ACCOUNTABILITY SYSTEM REQUIREMENTS
UNDER THE NO CHILD LEFT BEHIND ACT OF 2001

Ellen Forte Fast and Steve Hebbler

with

ASR-CAS Joint Study Group on Validity in Accountability Systems

February 2004

ACCOUNTABILITY SYSTEMS AND REPORTING
COMPREHENSIVE ASSESSMENT SYSTEMS FOR ESEA TITLE I

State Collaboratives on Assessment and Student Standards

COUNCIL OF CHIEF STATE SCHOOL OFFICERS—WASHINGTON, DC

Financial support for the development of this paper came from the member states of the Accountability Systems and Reporting and the Comprehensive Assessment Systems for ESEA Title I State Collaboratives on Assessment and Student Standards (SCASS) projects. The Council of Chief State School Officers claims Copyright © 2004, for this material for the benefit of those member states.

Acknowledgements

This paper resulted from the work of the Joint Study Group on Adequate Yearly Progress (AYP) comprised of state education specialists and consultants from two SCASS projects: Accountability Systems and Reporting (ASR) and Comprehensive Assessment Systems for ESEA Title I (CAS). The members of the Study Group benefited tremendously from discussions among SCASS colleagues throughout 2003:

Reginald Allen, Minnesota
Jan Barth, West Virginia (co-Chair)
Wes Bruce, Indiana
Ron Carriveau, Arizona
H. Gary Cook, Harcourt
Tom Deeter, Iowa
Dorothy DeMars, Alabama
Steve Hebbler, Mississippi (co-Chair)
Ellen Hedlund, Rhode Island
Pat Roschewski, Nebraska
Ron Houston, Delaware
Ted Jarrell, Delaware

Robin Jarvis, Louisiana
Susan Ketchum, Wisconsin
Sandra McQuain, West Virginia
Les Morse, Alaska
Jason Nicholas, Wyoming
Kenna Seal, West Virginia
Alan Sheinker, CTB
Gary Skoglund, South Dakota
Christine Steele, Wyoming
Michael Taylor, Utah
Robin Taylor, Delaware
Charles Wayne, Pennsylvania
Jeffrey Zaring, Indiana

Rolf Blank, ASR SCASS Coordinator

Jan Sheinker, CAS SCASS Coordinator

Several others served as critical resources during the development of this paper. These include:

Bill Erpenbach, WJE Enterprises

Dale Carlson, StandBACC Consulting

J. P. Beaudoin, Research in Action

Paul LaMarca, Nevada Department of Education

Scott Marion, Center for Assessment

This publication and any comments, observations, recommendations, or conclusions contained herein reflect the work of the authors. They do not necessarily reflect the views of the Council of Chief State School Officers, its members, or the U.S. Department of Education.

A Framework for Examining Validity in State Accountability Systems

Executive Summary

This *Executive Summary* provides an important overview of key issues and strategies related to the examination of validity in state accountability systems. The full paper addresses these topics in greater depth and is intended primarily for Chief State School Officers and their immediate staff members, especially State assessment directors, Title I directors, and others involved in statewide educational accountability policy development and implementation. It also explores unique issues that may arise as states implement and evaluate the accountability systems they have developed or revised in response to the *No Child Left Behind Act of 2001* (NCLB). This *Executive Summary* and the full paper are intended to be viewed as complimentary, companion pieces.

BACKGROUND

States and local education agencies (LEAs) across the country are now implementing accountability systems that must meet a number of highly specific requirements under NCLB. States have an obligation to evaluate these systems to determine whether they are associated with achievement of the intended goals and not with unintended, negative outcomes. This obligation is especially critical because these systems encompass high stakes consequences for schools and LEAs that do not meet certain criteria. In addition, the tight timeline under which these requirements must take effect has meant little time for planning and preparation prior to implementation. This combination of high stakes and limited planning and preparation time means that errors may be both more likely and more costly.

One of the first questions that policymakers and staff may pose with regard to accountability systems is whether the “right schools” are being identified for improvement. The answer depends on what kinds of schools were meant to be identified, which, in turn, depends on the goals the accountability system is intended to achieve. In fact, policymakers and staff should answer several questions to support their list of identified schools. For example, how trustworthy are the data on which the decision was based? How were these data combined in the decision-making model? What happens once schools are identified? What consequences and reforms are implemented and are they appropriate and effective?

To address these questions will require the accumulation and evaluation of a large body of evidence. When carried out systematically and

rigorously, this process can be considered *validation*. The purpose of the full paper is to provide a framework for this process.

FOUNDATIONS

1. **Every state must evaluate the validity of its accountability system.** Validation evidence is necessary to support the accountability claims made about individuals and agencies and the accompanying imposition of stakes. Professional standards for practice (APA/AERA/NCME, 1999; Baker, Linn, Herman, & Koretz, 2002) also highlight validation as being intrinsic to the decision-making process and the imposition of high stakes. The NCLB legislation itself makes 59 references to the need for validity with regard to assessment and/or accountability.
2. **States need accessible and flexible guidance on how to conduct this evaluation.** Validity with regard to accountability systems has received little formal attention to date; at present, there exists no framework to guide states in carrying out this work.
3. **The validation process as described in the full paper is basically grounded in the conceptions of validity and methods for evaluating validity as represented in the field of educational assessment.**

What Are Accountability Systems?

Accountability systems can be defined in the following way:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

This definition encompasses four critical concepts:

1. Performance indicators
2. Decision rules for attaching consequences to performance indicators
3. Consequences¹
4. Goals

For purposes of clarity, the concepts and the relationships among them are illustrated in a simplified manner in Figure 1.

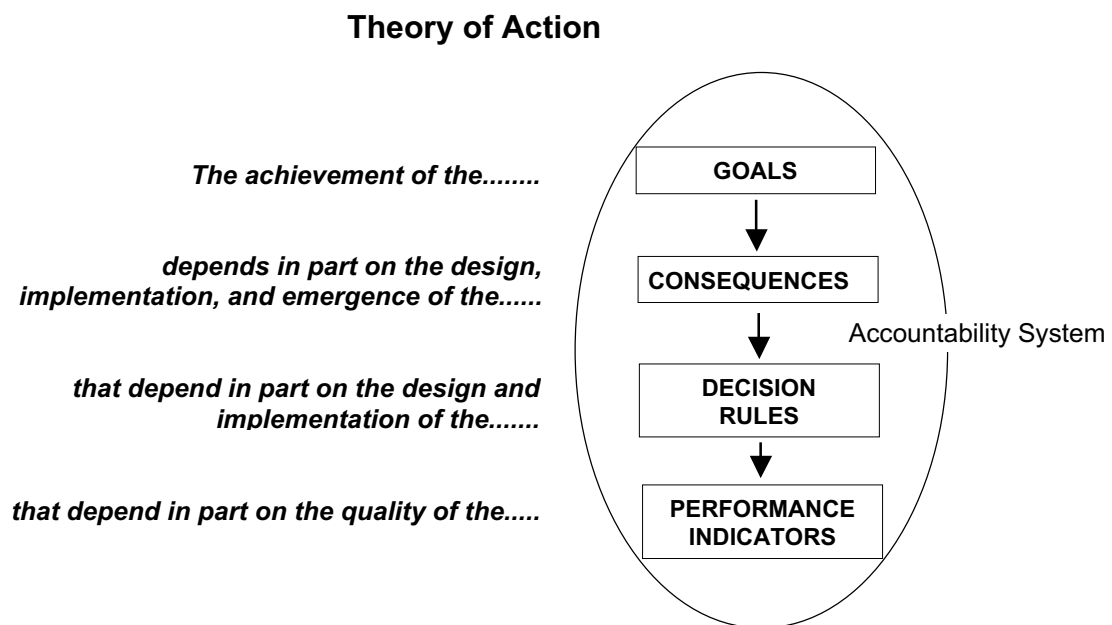
There are two types of consequences within accountability systems—imposed and emergent. Imposed consequences are those that are imposed by a state or LEA according to results of the accountability decisions and are consequences of performance as captured by the AYP model. Emergent consequences are the conditions in a school or LEA

¹ It is important to note that many people think of “consequences” as meaning something bad. That is not what is meant by this term here. Rather, consequences are simply one set of conditions that follow from another set of conditions.

that follow imposed consequences chronologically. Changes in resource allocations or professional development practices are some possible emergent consequences. Negative, unintended consequences (e.g., use of inappropriate test preparation techniques, loss of experienced faculty) would also be considered emergent consequences.

A Theory of Action specifies how emergent consequences are related to imposed consequences as well as how the other components of the accountability system are related to one another. This is illustrated in Figure 1. When represented in this way, it becomes clear that achievement of the end goals depends, in part, on the appropriate functioning of each of the other components. Further, unless each of these components is understood, it would be impossible to know why the goals were or were not achieved. Evaluation of the system, then, must involve studies of each of the parts and how they relate to one another.

Figure 1. Relationship between Components of Accountability Systems



Why Validate Accountability Systems?

When a state implements an accountability system, it is essentially making a claim that the system will help to achieve specific educational goals (e.g., improved achievement in reading/language arts and mathematics). Unless the state evaluates its system, it can have no evidence to support this claim. The state would not be able to defend its accountability decisions or the imposition of sanctions and would almost certainly lose credibility with its stakeholders and challenges in court. It is impossible to tell why a system does or does not work—or to defend the system against perfectly reasonable and inevitable challenges from affected stakeholders—unless it has been evaluated systematically.

Whether an accountability system “works” to achieve certain goals relies on three fundamental points:

1. the performance indicators are meaningful and relevant
and
2. the decision rules by which performance indicators are combined and attached to consequences function as intended
and
3. the imposition of specific consequences—
 - a. can ultimately lead to the intended goals by instigating or supporting intended reform activities;
 - b. will not lead to unintended, serious, and negative changes; and
 - c. is preferable to doing either nothing or something else.

Regardless of how obvious these points may seem, they are actually only tentative statements that must be tested empirically; they are hypotheses. Testing these hypotheses and the theory of action that subsumes them comprises the process of validation.

What is Validity?

“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*

- ◆ Validity with regard to accountability systems refers to the degree to which evidence and theory support the indicators, decisions, and consequences, individually and combined as established via the theory of action, as used for the purpose of achieving specific goals.
- ◆ “An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that: [1] the components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and [2] the system is accomplishing what was intended (and did not accomplish what was not intended)” (Marion et al., 2002, p. 105).
- ◆ Accountability is not Adequate Yearly Progress (AYP). The overarching validity question is not “Does this accountability system pick the right schools?”, but rather, “Does this accountability system do what it is intended to do?” Selecting the right schools is only part of the answer.

- ◆ Validity is not a property of an accountability system or of a decision made as part of that system (e.g., this school needs improvement), and validity cannot be captured conclusively. Rather, a judgment must be made regarding whether a body of evidence supports the system and each of its components as implemented for the intended purpose.
- ◆ The purpose of a validation process is not to prove claims true, worthy of making, or socially valuable; it is to “clarify for a relevant community what [a claim] means, and the limitations of each interpretation” (Cronbach, 1988, p. 3). A validation process cannot prove accountability systems worthy of implementing nor can the process prove any accountability decision true, worthy of making, or socially valuable.
- ◆ “Validation is never finished” (Cronbach, 1988, p. 5). An agency’s responsibility for validation starts the moment someone decides to design and implement an accountability system and continues as long as the consequences of that system are applied.
- ◆ Validity is a unified concept. Rather than different types of validity, there are different sources of evidence that should be considered as part of a validation process.
- ◆ The major professional statement on validity in relation to assessment is presented in the *Standards* (see reference below). This book and its subsequent revisions should be considered required reading for those engaged in educational assessment and/or accountability:

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

A VALIDATION FRAMEWORK

The basic elements of accountability systems (performance indicators, decision rules, consequences, and goals) are related to one another according to a theory of action. The validation framework presented here involves the evaluation of each of these elements as well as of the theory of action that underlies their association. This framework includes the following parts:

- ◆ Clarifying the **goals of the system and the theory of action** by which those goals are to be achieved
- ◆ Evaluating the **indicators** used to make accountability decisions
- ◆ Evaluating the **decision rules** (including AYP) used to determine how schools and LEAs are categorized for the purpose of distributing rewards, sanctions, and interventions

- ◆ Evaluating the **consequences** that are imposed for certain levels of performance as well as those that emerge subsequently

In practice, these pieces cannot be considered or addressed linearly; thus, they are not referred to here as “steps.” Ideally, system mapping would indeed be the first step in both the development and the evaluation of the system. However, the speed with which states have had to design and implement accountability decisions, and the prescriptive nature of NCLB requirements have pushed evaluation of the decision rules to the top of the evaluation priority list.

Given their limited resources, states may be best advised to prioritize their questions and to think of the “never-ending” validation process as a series of five-year plans. It is important for states to recognize that although validation requires the use of resources that could certainly be used elsewhere, *not* validating their systems could easily be far more expensive. The possible waste of time and resources by implementing ineffective programs, the loss of credibility, and the risk of lawsuits without a proper defense would cost much more than validation work.

States should use multiple data collection methods in their validation work. These can include, for example, extant data or data gathered via established collections, surveys, focus groups, site visits, and classroom observations

Clarifying Goals and the Theory of Action

To clarify system goals and the theory of action is to specify what the accountability system is meant to achieve, the means by which these goals are to be reached, and the outcomes and processes that will be the focus of the validation process. The major questions that should be addressed in this part of the validation process include:

- ◆ What are the goals this accountability system is meant to achieve?
- ◆ Who is to be held accountable for these goals in this system?
- ◆ What indicators are used to represent performance in relation to these goals?
- ◆ How and when are decisions made regarding performance toward the goals?
- ◆ What consequences are associated with different levels of performance?
- ◆ What changes are these consequences meant to affect?
- ◆ How are the intended changes thought to be related to the overall goals?

Standards-based accountability systems are supposed to work by applying specific criteria to indicators (e.g., test scores, graduation rates, attendance rates), categorizing each school and LEA as, for example, “excellent” or “needs improvement” based on whether the school or LEA met the criteria, and then assigning consequences corresponding to each

category. A state’s academic standards should provide the basis for accountability expectations. The consequences—and even the category labels themselves—are meant to affect change in areas including those listed above in order to improve specific educational outcomes. Once the consequences have been implemented, the outcomes are re-evaluated to determine whether the intended goals have been achieved.

That is essentially the theory of action for all accountability systems. To tailor this theory of action to an individual accountability system first requires answering a number of questions in relation to the accountability system being evaluated.

Evaluating the Indicators

Evaluation of indicators used for accountability purposes involves an investigation of their accuracy and meaningfulness. This investigation should address the following types of questions:

- ◆ What indicators are included in the accountability system and how is each used?
- ◆ How well do the definitions of these indicators capture what is intended?
- ◆ How reliable are the indicators that are used to make high stakes accountability decisions?

Indicators are values that represent constructs. Test scores, for example, are indicators of the construct(s) the test is meant to measure, such as reading comprehension or mathematics problem solving. Graduation rates may be considered indicators of how well a school or school system is preparing its students to meet graduation requirements.

The objectives in selecting indicators for use in an accountability system are to achieve as close a match as possible between the indicators and the system goals (Carlson, 2002; Gong, 2002; Marion et al., 2002) and to ensure that indicators can withstand the pressure of being used to make high stakes decisions. Validation of the indicator component of an accountability system involves evaluating the degree to which these objectives have been met.

States are not entirely constrained by the list of NCLB-defined indicators. They can include other academic indicators in the accountability systems either by using them as the “other academic indicator” for AYP or by using them to moderate or otherwise inform accountability decisions (Palmer & Coleman, 2003).

Test-Based Indicators

All states must use percent proficient indicators for reading and mathematics as part of their AYP decisions. The primary validity issues with regard to these assessment-based indicators are whether the indicators represent what they are meant to and are appropriate for use in high stakes decision-making. The evidence for consideration of these issues can be organized around two central questions, both of which should be addressed in a validation process:

- ◆ Does the indicator reflect the content and construct(s) it is meant to reflect?
- ◆ How reliable is the indicator?

Although states and testing contractors may be able to offer proof that they have conducted validation studies with regard to their tests for years, use of a test score for high stakes accountability purposes may be new, and this requires another look at assessment validity issues.

Evaluation of test-based indicators should address the following types of questions:

- ◆ Is the assessment aligned with the intended content domain and not other content domains?
- ◆ Does the assessment engage the intended cognitive and behavioral response processes and not other processes?
- ◆ Does the assessment minimize the biasing effects of irrelevant contextual or personal characteristics on test performance?
- ◆ Is the estimated reliability for the reported test scores adequate to support the use of these scores for making high stakes decisions?

Rate Indicators (other than percent proficient)

Rate indicators are proportions of a given population that meet a specific criterion or set of criteria; percent proficient, participation, and graduation rates are all rate indicators required for NCLB AYP. A number of states also use attendance rates as the other indicator for elementary and middle schools. In their simplest form, rates are calculated by assigning every member of a population a “1” if they met a criterion and a “0” if they did not, summing these ones and zeros, and then dividing by the total number of the target population.

To evaluate the quality of rates, it is necessary to examine the extent to which the definitions for the numerator and the denominator match the practices used to generate those figures. In addition, the reliability of the rate should be estimated. Thus, two primary questions for rate indicators result:

- ◆ How well do the practices for generating rates align with the definitions for the numerator and denominator?
- ◆ How reliable are the rate indicators?

Evaluating the Decision Rules

Evaluation of the decision rules involves examining the reliability and accuracy of accountability decisions, including AYP. This examination should address the following questions:

- ◆ Do the results of the AYP model support the goals of the accountability system?
- ◆ Were the “right schools” identified for rewards, sanctions, and interventions?
- ◆ Are the results of the AYP model stable over time?

Both the comparison of school or LEA performance to the accountability criteria and the process by which schools and LEAs can appeal their accountability designation are subsumed in an evaluation of decision rules.

Identification Process

While several studies have led the way toward an understanding of the reliability and validity of accountability systems, there is no general paradigm that can easily be applied to states’ AYP models to demonstrate that they are “reliable and valid.”

The AYP model prescribed in NCLB is a classification system. Each school and each LEA is classified annually into one of two groups—met AYP or did not meet AYP. If the model were perfectly reliable and accurate, every school and LEA would be correctly classified. Thus, the validity of the model is based on the degree to which the AYP determinations identify the correct schools—those in greatest need of improvement²—while not identifying schools that are doing an effective job educating all students.

However, no evaluation system produces perfectly accurate outcomes. Given a model that assigns schools and LEAs to two groups based on certain criteria, there are four possible outcomes for each decision, as illustrated in the four-fold truth table below.

Figure 2. Sample Four-Fold Truth Table for AYP

		AYP Determination (Results of the AYP Model)	
		Did Not Meet AYP	Met AYP
True School Effectiveness Status	Does Not Need Improvement	False Positive (Error)	True Negative
	Needs Improvement	True Positive	False Negative (Error)

A simple measure of the accuracy of the decisions made using the model would be the sum of the numbers in the “true” cells divided by the total number of schools or LEAs for which AYP determinations were made. In the case of a perfectly reliable model, the value would be 1.00, indicating that 100% of the schools or LEAs were correctly classified. A major challenge in this approach is the identification of a suitable

² It is recognized that failure to meet AYP targets for two consecutive years is required for identification for improvement under NCLB.

external criterion that allows one to determine the “true” AYP status for each school³.

Another complicating factor is that the AYP determination made using the AYP model prescribed by NCLB is based on a conjunctive (non-compensatory) set of standards. Failure to meet the standard on only one of many standards results in the AYP determination for the school or LEA being “not met.” The use of conjunctive standards “usually makes the accountability system much less reliable” (Gong, 2002, p. 9) and is seen as a common error in the design of effective accountability systems (Hill, 2000). Such a system might routinely classify schools/LEAs as needing improvement simply by chance due to the number of conjunctive standards that must be met from year to year (Kane & Staiger, 2002; Marion et al., 2002, p. 86).

This having been said, possible sources for an external criterion could be—

- ◆ information collected systematically via surveys or site visits from a range of schools, including those with relatively high and relatively low achievement (perhaps using a variable such as school-level poverty so that schools with higher and lower achievement than would be expected given their poverty levels could be identified).
- ◆ results from non-NCLB statewide school level accountability systems.

An alternate type of strategy to enhance the accuracy of AYP classifications is the differentiation between schools by the degree to which they missed their AYP targets. For example, states could—

- ◆ distinguish between schools and LEAs that miss their AYP targets by a large margin and those that miss by very little; and/or
- ◆ distinguish between schools and LEAs that miss all or most of their AYP targets and those that miss very few or only one.

While the AYP model in NCLB allows for no such distinctions (the same sanctions must be applied to both schools), some type of scale could be developed to represent the degree to which the school failed to meet the criteria in the AYP model. This scale could be used to determine the kind or level of resources and technical assistance that a state or LEA might provide to a school.

In addition to these studies, which address accuracy of identification, states must also evaluate the reliability of their decisions, which is defined here as the stability of AYP results over time (i.e., from year to year). A strategy for evaluating stability would be to rank all the schools in the state on the criterion measure described earlier. Taking only the schools falling at the top and bottom of the rankings (e.g., the top and bottom 10% or the top and bottom 25%), determine the extent to

³ Personal communication, Dale Carlson, September 16, 2003.

which schools in the high need and low need groups fall in the opposite group from one year to the next. The greater the degree of such change, the lower the reliability of the model.

When conducting evaluations of accountability decision rules such as AYP, states should bear in mind the following concerns:

- ◆ According to Gong (2002), “states should perform reliability analyses to ascertain that the level of error or uncertainty associated with accountability decisions is acceptable to the [state] and to key policy makers...states need this type of information for legal and professional defensibility of high-stakes programs” (p. 6).
- ◆ It is posited that for an AYP model, as for an assessment system, reliability is a necessary (Marion et al., 2002, p. 23), but not in itself sufficient, requirement for validity.
- ◆ There are many sources of error that affect the results of an accountability system. Many researchers agree that a chief contributor is sampling error (Cronbach et al., 1997; Hill, 2002; Hill & DePascale, 2002; Linn, 2001). To reduce misclassification errors that are affected by sampling error, states can employ confidence intervals with or without a minimum “n” for group size. States may also need to consider measurement error as a factor in misclassification.
- ◆ Although NCLB places strong emphasis under Title I (including provision of supplemental services) and other sections of the Act on the implementation of educational programs that are “research-based,” little if any prior study of the specific AYP model mandated for use by all states was conducted prior to the drafting of NCLB. Work done as the Act was being debated in Congress (Kane, Staiger, & Geppert, 2001; Riddle, 2001) and studies conducted since the Act became law in January 2002 (CEP, 2003; Hill, 2002; Linn, Baker, & Betebenner, 2002) overwhelmingly indicate that there are serious technical issues to be considered. This makes it even more important for states to have some way to judge the reasonableness of the results produced by their AYP models.
- ◆ Although NCLB provides for the consideration of growth through its “safe-harbor” provision, the improvement in student achievement is based only on the change in the percentage of students reaching the proficient level on the state’s academic assessments. This is problematic because much of the useful information and score variance in a state’s assessment program is lost by moving from units such as scale scores to a set of proficiency levels, and finally to the dichotomy (proficient or not proficient) required for use under NCLB. Several researchers have examined the problems associated with the use of coarse reporting statistics in accountability systems (Hanushek & Raymond, 2002; Hill, 1997; Hill, 2000; Linn, Baker & Betebenner, 2002; Thum, 2003).

Reviews and Appeals

An appeals process can encompass two categories of review:

- ◆ Reviews and corrections of raw data and indicators prior to the AYP analyses
- ◆ Reviews and appeals of preliminary AYP results

With regard to the reviews and appeals of preliminary AYP results, states make clear to schools and LEAs the following information:

- ◆ When preliminary AYP results will be available
- ◆ When and how appeals must be filed
- ◆ When state responses to appeals will be issued
- ◆ When AYP results will be considered final.

States vary widely in how they handle the appeals process. Of the states that have well-developed processes in place, notable characteristics are observed:

- ◆ Clear deadlines for filing appeals
- ◆ Readily available forms
- ◆ Limited allowable reasons for appeal
- ◆ Specific requirements for the types of data that will be considered in the state's review of the appeal

Evaluating the Consequences

Evaluating the consequences component of an accountability system involves consideration of how the imposed consequences are implemented and how they are related to both intended and unintended emergent changes in school and LEA functioning. It simply cannot be assumed that selecting the “right schools” and assigning them pre-specified consequences will lead to the intended reforms and ultimately result in the achievement of intended goals. Nor can it be assumed that it will not lead to any unintended, negative consequences.

Studies of consequences should address the following types of questions:

- ◆ How well are rewards, sanctions, and interventions implemented?
- ◆ How do school and LEA characteristics, as well as other facets of the context, moderate the implementation of the consequences?
- ◆ To what degree are the intended actions occurring in relation to the application of rewards, sanctions, and interventions?
- ◆ To what degree are negative, unintended consequences occurring in relation to the application of rewards, sanctions, and interventions?

- ◆ To what degree are the reform activities associated with achievement of the goals of the system?

The primary concerns with regard to accountability consequences are—

- ◆ how the consequences that are imposed on schools, which include rewards, sanctions, and interventions, are implemented;
- ◆ whether imposed consequences are associated with the emergence of the intended reforms, as indicated in the state's theory of action, and also the emergence of any negative, unintended consequences or activities; and
- ◆ whether the activities and conditions that emerge after the application of consequences are associated with the achievement of the accountability goals.

References

- American Psychological Association, American Educational Research Association, & the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* [Policy Brief 5]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Carlson, D. (2002). The focus of state educational accountability systems: Four methods of judging school quality and progress. In W. J. Erpenbach et al., *Incorporating multiple measures of student performance into state accountability systems—A compendium of resources* (pp. 285-297). Washington, DC: Council of Chief State School Officers.
- Center on Education Policy. (2003). *From the Capitol to the classroom: State and federal efforts to implement the No Child Left Behind Act*. Washington, DC: Author.
- Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Gong, B. (2002). *Designing school accountability systems: Towards a framework and process*. Washington, DC: Council of Chief State School Officers.
- Hanushek, E. A., & Raymond, M. E. (2002). Sorting out accountability systems. In W. Evers & H. Walberg (Eds.), *School accountability* (pp. 75-104). Palo Alto, CA: Stanford University, Hoover Press.
- Hill, R. (1997, June). *Calculating and reducing errors associated with the evaluation of adequate yearly progress*. Paper presented at the CCSSO Annual Large-Scale Assessment Conference, Colorado Springs, CO.
- Hill, R. (2000, March). *Common problems with accountability systems*. Paper presented at the Conference on Policy and Measurement Issues in Large-Scale Science and Mathematics Assessment, Washington, DC.
- Hill, R. (2002, April). *Examining the reliability of accountability systems*. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA.

- Hill, R., & DePascale, C. (2002). *Determining the reliability of school scores*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Kane, T. J., & Staiger, D. O. (2002) *Volatility in school test scores: Implications for test-based accountability systems* [Brookings Papers on Education Policy]. Washington, DC: The Brookings Institution.
- Kane, T. J., Staiger, D. O., & Geppert, J. (2001). *Assessing the definition of "adequate yearly progress" in the House and Senate education bills* [Unpublished manuscript].
- Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* [CSE Technical Report 539]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(6), 3-16.
- Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Palmer, S., & Coleman, A. (2003, September). *Developing a framework for policy and legal implications*. Presentation at the Council of Chief State School Officers' Workshop on Implementing AYP in States' Accountability Systems, Washington, DC.
- Riddle, W. (2001). *Adequate yearly progress under the ESEA: Provisions, issues, and options regarding House and Senate versions of H.R. 1*. [CRS Report RL31035]. Washington, DC: Congressional Research Service, The Library of Congress,
- Thum, Y. M. (2003). *No child left behind: Methodological challenges & recommendations for measuring adequate yearly progress* [CSE Technical Report 590]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.