MAKING VALID AND RELIABLE DECISIONS IN DETERMINING ADEQUATE YEARLY PROGRESS

A Paper In The Series:

Implementing The State Accountability System Requirements Under The No Child Left Behind Act Of 2001

December 2002

ASR-CAS Joint Study Group on Adequate Yearly Progress

SCOTT MARION, Co-Chair
CAROLE WHITE, Co-Chair
DALE CARLSON
WILLIAM J. ERPENBACH
STANLEY RABINOWITZ
JAN SHEINKER

Comprehensive Assessment Systems for ESEA Title I
Accountability Systems and Reporting
State Collaborative on Assessment and Student Standards

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Council Of Chief State School Officers

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization of the public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public. Through its structure of standing and special committees, the Council responds to a broad range of concerns about education and provides leadership and technical assistance on major educational issues.

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) Project 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-four states and one extra-state jurisdiction participated in one or more of the eleven projects offered during the project year 2001-2002.

John Olson, Director of Assessments
Jan Sheinker, Coordinator
Comprehensive Assessment Systems for ESEA Title I SCASS
and
Rolf Blank, Director of Education Indicators Programs and Coordinator
Accountability Systems and Reporting SCASS

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431

(202) 408-5505
Fax (202) 408-8072
www.ccsso.org

MAKING VALID AND RELIABLE DECISIONS IN DETERMINING ADEQUATE YEARLY PROGRESS

A Paper In The Series: Implementing The State Accountability System Requirements Under The No Child Left Behind Act Of 2001

December 2002

ASR-CAS Joint Study Group on Adequate Yearly Progress

SCOTT MARION, Co-Chair
CAROLE WHITE, Co-Chair
DALE CARLSON
WILLIAM J. ERPENBACH
STANLEY RABINOWITZ
JAN SHEINKER

Comprehensive Assessment Systems for ESEA Title I
Accountability Systems and Reporting
State Collaborative on Assessment and Student Standards

COUNCIL OF CHIEF STATE SCHOOL OFFICERS—WASHINGTON, DC

ISBN 1-884037-80-1

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: Comprehensive Assessment Systems for ESEA Title I (CAS), and the Accountability Systems and Reporting (ASR). The members of the Study Group benefited tremendously from discussions among SCASS colleagues throughout 2002.

Jan Barth, West Virginia
Dale Carlson, Consultant
William J. Erpenbach, Consultant
Paul M. LaMarca, Nevada
Scott Marion, Wyoming
Bernadette Morris, Louisiana
Stanley Rabinowitz, West Ed

Alan Sheinker, CTB
Christine Steele, Wyoming
Charlene Tucker, Connecticut
Cathy Wagner, Minnesota
Charles Wayne, Pennsylvania
Carole White, Delaware
Grace Ross, ED, Ex-Officio

Rolf Blank, CCSSO, ASR Coordinator
Jan Sheinker, CAS Coordinator

The work of the Joint Study Group was informed by two CCSSO meetings concerning AYP issues. In May 2002, a video-conference was held with chiefs and state leaders across the country to discuss the issues with AYP requirements under NCLB. Presentations from four states on their planning efforts were extremely informative, and the key presenters were:

Louis Fabrizio, North Carolina
Robin Taylor, Delaware
H. Gary Cook, Wisconsin
J.P. Beaudoin, Louisiana.

In the final stages of writing, the Study Group also gained insight and ideas from discussions conducted by the AYP "brain trust meeting" of state and national experts convened by CCSSO in October 2002.

Many of the ideas presented in the paper can be traced to discussions among members of the National Assessment and Accountability Work Group—a small study group sponsored by West Ed, whose members include: (Chair) Stanley Rabinowitz, West Ed; J.Reginald Allen, Minnesota; Sri Ananda, West Ed; Dale Carlson, Consultant, Brian Gong, Center for Assessment; Edward Haertel, Stanford University; Joan Herman, CRESST/UCLA; Richard Hill, Center for Assessment; Paul M. LaMarca, Nevada; Scott Marion, Wyoming; Scott Norton, Louisiana; and William Schafer, University of Maryland.

CCSSO received financial support for the development of this paper from the SCASS member states and the U.S. Department of Education (ED).

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 ED.

Table of Contents

APPENDIX A

APPENDIX B

APPENDIX C

APPENDIX D

Chapter 1

Introduction: Defining the Issues

This paper, together with its accompanying *Executive Summary*, 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. The *Executive Summary* provides an important overview of key issues, decision points, decision consequences, and policy implications related to making valid and reliable decisions in the calculation of adequate yearly progress (AYP). This paper addresses those topics in greater depth including a full exploration of the related technical aspects of validity and reliability in AYP determinations by States. It also explores unique issues that arise in designing accountability systems under the NCLB Act and critical variables related to decisions that States must make in finalizing these systems. The *Executive Summary* and this paper are intended to be viewed as complimentary, companion pieces.

Finally, it is anticipated that, in addition to State Educational Agency personnel, Peer Reviewers of State Accountability Systems will also find the paper instructive as they begin reviewing plans for statewide accountability systems in early 2003.

Key Issues for States

The following key issues have arisen as States began work toward developing plans for determining AYP of schools and districts—a State plan that meets the requirements under the *No Child Left Behind* (NCLB) Act and meets existing State policies, priorities, and needs. These issues led to the analysis, research, and writing for this paper, and the intent of the Council is to assist State leaders with addressing the issues:

- **Multiple, separate indicators**. The concept of AYP defined by the NCLB Act is based on students attaining a target level of achievement across a number of separate indicators and disaggregated student subgroups for each school and district; while many statewide accountability systems developed under the 1994 ESEA Reauthorization are based on student achievement and improvement on a combined indicator score or rating for the school and district. (See Ch. 1, pp. 8-10)
- **Definition of proficient**. School and district determination of AYP will rely on each State's establishment of a "proficient" level of student performance on State assessments that are aligned with State content standards. While expectations for the proficient level will vary by State, AYP is based on the percent of students meeting proficient and the expected

percentage increases over time. Many States previously established measures of growth or improvement toward proficient as part of their accountability system. (Ch. 1, pp. 12-15; Ch. 2, pp. 28-39; Ch. 3, pp. 55-58)

- **Selecting assessments and other indicators**. The State accountability system as defined by the NCLB Act must include assessment of reading/language arts and mathematics at six grades, as well as science (2005-06) and other indicators such as attendance; however, no assessment or other indicator can help compensate for poor student performance on the core subjects. (Ch. 1, pp. 5-11, 14-15)
- **Starting points and goals**. Under the NCLB Act, each State sets starting points for measuring school and district progress based on existing State assessment results and sets intermediate goals (2-3 years) toward all students proficient by 2014. Some States have found that setting separate starting points for subgroups or schools could provide more flexibility in AYP but more complexity. However, setting separate starting points for student subgroups or schools is not permitted under the regulations [see § 200.16(c)(2), Final Regulations, p. 71716, and related comments/discussion, p. 71742]. (Ch. 3, pp. 69-70)
- **Minimum "n" per student group**. For AYP determination, States must set a minimum number of students for a reliable indicator (e.g., number of Hispanic students that met proficient level on 4th grade State math assessment). While States find that a higher "n" yields fewer schools failing AYP, expert analysis has shown more valid and reliable AYP decisions can be made with a confidence interval, statistics-based, approach. (Ch. 3, pp. 63-65)
- **Include all students and schools.** State accountability systems are required to account for all K-12 public education students and move all toward the proficient level. A required AYP indicator is that 95 percent of students enrolled at assessed grades do take the assessments. Small schools are less likely to be held accountable if AYP is based on a high minimum "n." (Ch. 1, p.14; Ch. 3, pp. 58-60)
- **Multiple years of data**. States have found advantages in deciding to employ uniform averaging or rolling averages across multiple years of assessment results for making annual AYP determinations. Having the capacity to monitor AYP success/failure by student sub-group by school from one year to the next can reduce the number of schools identified for improvement. (Ch. 3, pp. 70-72)

Intent and Purpose of this Paper

The purpose of this paper is to address a specific provision found in Title I of the 2001 Reauthorized Elementary and Secondary Education Act (better known as the *No Child Left Behind Act of 2001*) concerning making valid and reliable decisions in the determination of the Adequate Yearly Progress of schools, districts, and States toward the goal of **all** students meeting State standards for proficiency in reading or language arts and mathematics by 2013-14 [Section 1111(b)(2)(C)(v)(II)(dd)]. Title I, Improving the Academic Achievement of the Disadvantaged, is the largest Federal program providing assistance to public elementary and secondary schools. Title I provides funds and a program framework to support the improvement of education for students attending schools with high concentrations of students from low-income families (see adjacent text box). The NCLB Act authorizes nearly $1B in annual funding to support the various programs covered under the law.

Intent and Purpose of Title I of The No Child Left Behind (NCLB) Act

‘‘The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments. This purpose can be accomplished by— . . .
‘‘(4) holding schools, local educational agencies, and States accountable for improving the academic achievement of all students, and identifying and turning around low-performing schools that have failed to provide a high-quality education to their students, while providing alternatives to students in such schools to enable the students to receive a high-quality education;” [P. L. 107-110 ‘‘No Child Left Behind Act of 2001,’’ Title I-Improving the Academic Achievement of the Disadvantaged, Section 1001, Statement of Purpose.]

According to a summary of the 2001 Reauthorization prepared by the Education Commission of the States (2002),

> This new law, a potent blend of new requirements, incentives and resources, poses enormous challenges for States. It sets deadlines for them to expand the scope and frequency of student testing, revamp their accountability systems and guarantee that every classroom is staffed by a teacher qualified to teach in his or her subject area. It requires States to make demonstrable progress from year to year in raising the percentage of students proficient in reading and math[ematics], and in narrowing the test-score gap between advantaged and disadvantaged students. And it pushes them to rely more heavily on research-based approaches to improving school quality and student performance. (p. 3)

Important Information for SEA Staffs.

This paper provides particularly important information for State Educational Agency staffs as they address the Title I requirement related to the development and implementation of a single, statewide accountability system. States are required to submit to the U. S. Department of Education (ED) a status report detailing their progress on AYP by January 31, 2003. In a recent letter to Chief State School Officers (Neuman, 2002), ED provided additional guidance to States related to AYP plan submittals—initial guidance had been provided by ED in June 2002. To help States complete their accountability plans, ED prepared an “accountability Workbook.” States were advised that, “Once completed, the Workbook will serve as a State’s Consolidated Application for the January 31, 2003 deadline and the basis for the May 1, 2003 deadline.” Although earlier described by ED representatives as a “status report” to detail a State’s progress with the AYP regulations, completion of the Workbook will likely require substantial time, detail, and documentation.

The information provided in this paper is also intended to assist States by

- Providing information that will guide them in identifying their needs for further understanding/background regarding specific aspects of the Adequate Yearly Progress/ State accountability systems requirements under the NCLB Act.
- Providing information for designers of State accountability systems to alert them to potential negative or unintended consequences that may arise from accountability systems designs and implementation strategies and to suggest steps that may minimize the impact of such consequences. For example, a State may choose to set a relatively high minimum group size number, having the effect of identifying fewer schools for improvement but which, in turn, may result in masking lower performance by small subgroups of students with the greatest needs, especially at the school building level. Thus, additional instructional support and resources needed by these students might not be made available.
- Providing information about and options to consider in approaching the flexible elements of the NCLB Act's AYP requirements that may help State leaders make better informed decisions about selecting or designing a single, statewide accountability system and evaluating the reliability and validity of the information to be yielded by that system.
- Approaching, analyzing options for, and making decisions about how State officials will define the AYP of schools and local educational agencies (LEAs) under the NCLB Act. State AYP definitions must be "statistically valid and reliable" [NCLB Act, Part A, sec. 1111(b)(2)(C)(ii)—see Appendix A]. Among the major questions States must consider are:
    - What should one consider with regard to the validity and reliability of accountability decisions?
    - How does one assess the validity and reliability of accountability decisions?
    - How does one decide how to ensure the validity and reliability of accountability decisions?
    - What are the potential sources of error or the consequences resulting from inappropriate decisions?
- Providing sufficient technical information to assist States in conducting simulations/analyses of their own assessment data to examine the validity and reliability of various methods for making AYP decisions including:
    - Determining starting points.
    - Setting the minimum "n".
    - Using standard error approaches (including confidence intervals).
    - Aggregating data.
    - Establishing annual and intermediate measurable objectives.
- Helping State leaders design systems that are sufficiently flexible to satisfy forthcoming regulations.

Accountability Requirements Under the NCLB Act—Shifting Emphases and New Challenges

State level elementary and secondary education policymakers and others across the nation are currently engaged in considerable discussion and work examining the likely consequences of implementing changes to the Elementary and Secondary Education Act of 1965 (ESEA) enacted under the NCLB Act of 2001. Across the nation, every conference or meeting of major educational interest groups now includes significant parts of the agendum devoted to presentations and papers concerning the new law. Articles about a myriad of issues related to implementing the NCLB Act's requirements for new assessments, expanded accountability requirements, schools identified for improvement, and the implementation of sanctions such as public school choice, supplemental educational services, and corrective action appear with regular frequency in newspapers and other media throughout the country.

The passage of the NCLB Act marked a significant shift in Federal educational policy from an emphasis on standards and assessments to an emphasis on accountability—school, district, and State accountability for students' academic achievement such that **all** students reach, at a minimum, proficiency on State's academic achievement standards and State academic assessments by 2013-14. Unlike previous ESEA Reauthorizations, the NCLB Act provisions became effective upon enactment of the law in January 2002. Thus, States were required to put into place most of the changes by the beginning of the current school year.

Some of the key differences related to accountability requirements between the 1994 and 2001 ESEA Reauthorizations are summarized in Table 1. This summary is intended to be illustrative and not exhaustive in analyses or nature.

TABLE 1. SUMMARY OF KEY ACCOUNTABILITY REQUIREMENT DIFFERENCES BETWEEN THE 1994 AND 2001 ESEA REAUTHORIZATIONS

	1994 ESEA Reauthorization (Improving America's Schools Act)	2001 ESEA Reauthorization (No Child Left Behind Act)
Transition Period	Almost one full school year with additional time to bring on line aligned standards, assessments, and accountability systems.	None—law was effective on enactment although standards and assessments beyond those previously required will be phased in gradually.
Assessments	Reading or language arts and mathematics at least once annually in the three grade spans—3-5, 6-9, and 10-12.	Same except that the assessments must be administered at least once annually in each grade, 3 through 8 by 2005-06 (and once within grades 10-12) with science administered at least once in each of the three grade spans by 2007-08.
Statewide Accountability Systems	Statewide system, using assessments administered to all students (not just those in Title I) to measure progress of schools and districts participating in Title I.	Single, statewide system required to measure progress of all schools and districts, not just those participating in Title I.

	1994 ESEA Reauthorization (Improving America's Schools Act)	2001 ESEA Reauthorization (No Child Left Behind Act)
AYP Measures	States were required to establish AYP standards that could be limited to schools and districts receiving Title I funds. Identification of schools and districts based on the performance of ALL students with no pre-determined annual growth rate or period of time for all students to master a State's academic content standards. Multiple measures of student performance could be applied in AYP determinations.	Each of at least 9 subgroups of students must reach proficient or advanced achievement levels in reading or language arts and mathematics by 2013-14 (Uniform progress is required beginning in 2002-03.). AYP determinations are based solely on student achievement results on State assessments. At least 95% of the students in each subgroup must participate in the assessments and all must meet the State's performance target in another academic indicator as prescribed in the law.
Rewards and Sanctions	*Rewards:* States were to identify especially successful schools and distinguished educators and were authorized to use Title I funds to provide additional support. *Sanctions*: Many possibilities identified but most could not be taken until standards and assessments were fully implemented and none were required.	*Rewards:* States must identify rewards and may use Title I funds in support of the rewards. *Sanctions:* A set of progressive sanctions required to be applied to low-performing schools and districts receiving Title I funds. Most sanctions are automatic although districts and States have some discretion regarding the extent of the number and scope of sanctions related to corrective actions that are applied under the law.
English Language Acquisition	Acquisition of English language proficiency not required.	States must set annual measurable objectives for increasing English language proficiency by limited English proficient students, and districts must annually assess same.
National Assessment of Educational Progress	State, district, and school participation not required.	State participation required as well as district and school participation required if district receives Title I funds and any of its schools identified for participation.

The NCLB Act did not include transitional language providing further direction to States in moving from the 1994 to 2001 ESEA Reauthorizations. Final regulations related to standards and assessments were published earlier this year, and final regulations related to accountability systems have just been promulgated. Non-regulatory guidance has not been issued in either of these areas.

Some provisions of the NCLB Act that may have considerable impact on decisions States must make regarding the validity and reliability of their AYP definitions seem, on initial reading, ambiguous (e.g., whether AYP will be based on the **same** performance requirement for two consecutive years or based on **any** performance requirement for two consecutive years such as reading scores for 4th grade LEP students versus mathematics scores for 4th grade economically disadvantaged students). One reading of the law appears to support the former while the draft regulations on accountability issued on August 6, 2002, seem to support the latter.[1] The final rules

[1] Section 1111(b)(1) requires each LEA to "identify for school improvement any elementary or secondary school served under this part that fails, for 2 consecutive years, to make adequate yearly progress as defined in the State's plan under section 1111(b)(2)." Section 1111(b)(2) requires, among other things, that States establish "separate measurable annual objectives for continuous and substantial improvement" for each of the subgroups identified in subsection 1111(b)(2)(C)(v).

did not address this issue either. Perhaps it will be addressed in subsequent guidance at which time some other options may present themselves for consideration by States.

An example of a **possible** other area of flexibility some States were hoping for was the option to develop separate starting points for student subgroups. This option was suggested in Secretary Paige's July 24, 2002, letter to SEAs (see Section 1111(b)(2)(E) and criterion 5). However, the final regulations on accountability did not provide for this option (see Comment/Discussion under Section 200.16, p. 71742.).

Another question concerning flexibility in the determination of starting points is whether States could average data across two consecutive years (e.g., 2000-01 and 2001-02) instead of using only 2001-02 data as set forth in Section 1111(b)(2)(E). Small States with many small student subgroups may find it advantageous to determine their starting points using a larger student performance database. The final regulations on accountability permit averaging of up to three years data provided that 2001-02 data are included [§200.20(d)].

The provisions in this latest ESEA Reauthorization receiving by far the greatest attention are those pertaining to requirements for

- Additional standards-based assessments of student performance (aligned by standards);
- New requirements for State accountability systems; and
- Expanded requirements for the determination of adequate yearly progress (AYP).

Prescriptive Requirements and Sanctions.

The NCLB Act sets forth highly prescriptive requirements for how States must design and implement a single, statewide system of educational accountability for all schools and districts regardless of the extent of their participation in Title I. As noted earlier, accountability, accompanied by sanctions and rewards, constitutes one of the most significant differences between the 1994 and 2001 Reauthorizations. Whereas States could previously use multiple sources of information to make accountability determinations and set the timelines for all students to achieve at the proficient or advanced levels, the law now prescribes both the nature of accountability measures and the timelines (see following discussion regarding **compensatory** and **conjunctive** models or approaches to accountability decisions).

Under the 1994 ESEA Reauthorization, States were required to develop a set of sanctions and rewards for low- and high-performing schools and districts. However, while the law then included examples of sanctions, districts and States had considerable discretion in applying any of these potential sanctions. The primary requirements for schools and districts identified for improvement then were the development of improvement plans and additional professional development activities for teachers and other school personnel.

The current reauthorization sets forth additional, **required** sanctions beginning the first year that a school or district is identified for improvement. For schools and districts identified for improvement prior to this school year (under the 1994 Reauthorization), the extent of sanctions may be immediately more extensive depending on the number of prior years that they have been in that status (see Section 1116(f) of the law and § 200.32 and § 200.50 of the December 5, 2002, final regulations on accountability). Additional sanctions come into play the longer that a school or district remains so identified. With respect to States, beginning in 2005, a listing of States not

On the other hand, § 200.32(a)(1) of the December 5, 2002, final regulations states, "An LEA must identify for school improvement any elementary or secondary school served under subpart A of this part that fails, for two consecutive years, to make AYP as defined under § 200.13 through 200.20 (meaning any two AYP measures; not the same two measures)."

making AYP will be included in the Secretary of Education's annual report "to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health, Education, Labor, and Pensions of the Senate…" (Section 6164).

The 2001 ESEA Reauthorization continues the national emphasis on standards, annual assessments of student learning, accountability for results, and school improvement. The Reauthorization adds new requirements for the annual assessment of the acquisition of English language proficiency by limited English proficient students and mandatory requirements for school districts (receiving Title I funds) and their schools, if selected, to participate in the National Assessment of Educational Progress (NAEP).

Accountability—A Change in Direction and Emphasis.

However, the *No Child Left Behind Act of 2001* has most surely and markedly changed the direction of educational accountability for the nation's public school districts, public schools, and States. Under the 1994 ESEA Reauthorization, all parties—schools, districts, and States—were held accountable for annual, measurable increases (i.e., progress) toward the goal of having all students achieve to high academic standards (which could be based on a **compensatory model** to determine whether the required amount of progress was being made).

A compensatory model or approach allows higher scores on some measures to offset (i.e., compensate for) lower scores on other measures. For example, higher performance in language arts could be used to offset lower performance in mathematics. The most common example of the compensatory approach is the simple average. There are many more complex compensatory methods, such as many of the standards-setting processes used in large-scale assessment programs, but for now, thinking about the compensatory approach as a simple average will suffice. Consistent with the 1994 Reauthorization then, States began to set their own timelines and AYP requirements for the attainment of this goal. Although performance data were required to be reported publicly by disaggregated groups, AYP determinations were made on the basis of the performance of **all** students and not on the basis of the performance of subgroups of students.

In the new accountability system under the NCLB Act, adequate yearly progress requirements are precisely set in the law and each student subgroup is required to meet or exceed its annual measurable objective each year—a measure of status; requiring that each of the various subgroups of students be at or above the performance targets for a given school year (which means that only a **conjunctive model** can be employed). The NCLB Act holds all parties—schools, districts, and States—accountable for helping **all** students, including specific subgroups of students, to achieve to each State's proficient or advanced levels in reading or language arts and mathematics within 12 years beginning with the 2002-03 school year. (There are three proficiency levels specified in the law—advanced, proficient, and basic—although they are not specifically defined in the law or by regulations.)

In a conjunctive model or approach, scores on all measures used must be above the criterion point (cut score) for the student to have met the overall standard. If three measures were used to determine whether or not a student has met a standard, the student would have to be above the cut score on each measure to be considered proficient. This is a fairly stringent approach and typically leads to the lowest pass rate.

The NCLB Act, contrary to the 1994 Reauthorization, also sets a minimum assessment participation rate applicable to **all** students and to each subgroup of students specified in the law. The Act further requires States to include at least one other academic indicator (in addition to the State assessment) of student performance in the determination of AYP. In this regard,

- States may select any academic indicator for the elementary school level—the regulations refer to an indicator for elementary schools and another for middle schools (e.g., attendance).
- Under Title IX of the NCLB Act, States define the grades included at the elementary and secondary school levels. Nevertheless, the practical effect of the law as it pertains to assessments (in the grade spans, 3-5, 6-9, and 10-12) is to define elementary as including grades K-8 and secondary as including grades 9-12. The Federal regulations on accountability introduces use of the term "middle school" at § 200.19(a)(2) although that term is not used in Title I of the Act.
- Graduation rate is required for the secondary school level indicator [see section 1111(b)(2)(C)(vi)]. Graduation rate is defined in section 1111(b)(2)(C)(vi) of the NCLB Act as, "the percentage of students who graduate from secondary school with a regular diploma in the standard number of years." The ED's final accountability regulations add that the definition does not include students earning GEDs [§ 200.19(a)(1)(A)] and further provide for, "(B) Another definition, developed by the State and approved by the Secretary in the State plan, that more accurately measures the rate of students who graduate from high school…" (Fed. Reg., December 5, 2002).
- States may also include **other** academic indicators at either or both the elementary (includes middle) and secondary school levels. However—with the exception of meeting the "safe harbor" provisions (see also Appendix B regarding use of these provisions in accountability determinations) under section 1111(b)(2)(I)(i)—additional academic indicators, beyond those required in the law, may only be used to identify additional schools for school improvement or in need of corrective action or restructuring, and they cannot be applied to change the status of schools previously identified for improvement. Student performance on the academic indicators required in the NCLB Act cannot be used to offset (or compensate for) student performance on the reading or language arts and mathematics assessments or the requirement related to participation rate for the assessments.
- The NCLB Act does not appear to require States to set improvement or growth targets with respect to the academic indicators. The final regulations related to AYP requirements issued by ED on August 6, 2002, affirm this providing that, "The State may, but is not required to, increase the goals of its other academic indicators over the course of the timeline . . . [through 2013-14]" [§ 200.19(d)(1)]. States may choose to have the indicators remain constant over time or to increase them over time.
- As with student achievement and assessment participation rates, additional academic indicators added at a State's discretion to its statewide accountability system also apply to school districts and to the State itself in AYP determinations. And, as noted elsewhere in this chapter, using **additional** academic indicators has only the result of identifying more schools and districts for improvement; they cannot be used to "compensate" for other missed performance targets that are required in the determination of AYP.
- USDE originally proposed that student performance on the other academic indicators required at the elementary and secondary school levels would have to be disaggregated and used in AYP determinations (Draft regulations, 2002, August 6). However, the final regulations published on December 5, 2002, provide at §200.19(d)(2)(ii) that States, "Need not disaggregate those indicators for determining AYP…." Student performance on the other academic indicators, then, will be used for AYP purposes only in terms of whether ALL students met the annual progress measurement **and** in instances where the

performance of subgroups of students are reviewed consistent with a "safe harbor" determination of AYP (see Appendix B for further discussion).

- Student performance on the other academic indicators must, however, be disaggregated for purposes of school, district, and State public reporting under Section 1111(h). In determining not to include disaggregated reporting for AYP purposes, USDE commented, "The Secretary is confident that publicly reporting disaggregated data on the other academic indicators will ensure that schools, LEAs, and the State are held accountable for subgroup performance" (Federal Title I Regulations, 2002, December 5, p. 71742).

AYP—At Least 37 Determinations. Under the NCLB Act, the adequate yearly progress of schools, districts, and States would be based on up to 37 determinations of student performance related to the annual State assessments (in each of the grade spans 3-5, 6-9, and 10-12) in at least reading or language arts and mathematics. The 37 determinations (per grade span) are based on the following indicators related to the performance of all students as well as the performance of the eight subgroups of students as specified under the law—economically disadvantaged students; students from major racial and ethnic groups[2] (in this example); students with disabilities; and students with limited English proficiency [Section 1111(b)(2)(C)(v)]. To illustrate:

- The assessment of student performance in at least the two required subject areas

  (2 x 9 = 18);
- At least a 95% student participation rate in each of the required assessments

  (2 x 9 = 18); and
- The performance of **all** students on at least one additional academic indicator such as attendance rate in the elementary school grades and graduation rate which is required at the high school level (1 x 1 = 1).

Each student's results would appear in at least two of nine subgroups (all students and a major racial/ethnic group). **There is a potential for up to 21 determinations of an individual student's performance in this example.**

Table 2 portrays how one arrives at the 37 performance determinations **for each** of the three grade spans (3-5, 6-9, and 10-12) for which assessments and accountability determinations are presently required. Beginning in 2005-06, the assessments in reading or language arts and mathematics must be expanded to include each grade, 3-8, and in 2007-08, assessments in science must be included in each of the three grade spans (but the science assessments will not be factored into AYP determinations).

The NCLB Act also permits States to establish a uniform procedure for averaging data, which includes using data across grades in a school [Section 1111(b)(2)(J)(iii)]. In States exercising this option, it is possible, for example, in a K-8 school building, for AYP determinations to be made, using cross grade averaging, on the basis of 37—not 74—measures of student performance. In this instance, accountability is based on the combining and averaging of student performance data from assessments administered at two grade spans (under present NCLB Act provisions) and not on the basis of the assessment results at each of the two grade spans.

[2] Federal agencies are required to implement no later than January 1, 2003 the following Race/Ethnicity categories as published in the Federal Register on October 20, 1997: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian, or Other Pacific Islander, and White.

TABLE 2. AN ILLUSTRATION OF THE 37 STUDENT PERFORMANCE DETERMINATIONS

	Reading/LA		Mathematics		Other Academic Indicator
	Participation Rate	% Meeting Standard*	Participation Rate	% Meeting Standard*	
All Students					
Economically Disadvantaged					
Racial/Ethnic Group 1					
Racial/Ethnic Group 2					
Racial/Ethnic Group 3					
Racial/Ethnic Group 4					
Racial/Ethnic Group 5					
Students with Disabilities					
LEP Students					

*Enrolled Full Academic Year

In examining data that might be reported in the above summary, answers to the following questions would determine whether a school, district, or State made the applicable AYP requirements:

1. Did at least 95% of each student subgroup (takers) enrolled in the school/district/State for a full academic year take the assessments (with or without accommodations or alternate assessments)?
2. Did each student subgroup at each grade [or across grades if a State permits the use of data across grades in a school consistent with the provisions of Section 1111(b)(2)(J)(iii)] at the school/district/State levels meet the AYP growth requirement in reading or language arts and mathematics?
3. Did the all students group meet the graduation rate target (secondary school level) or the other academic indicator target (elementary school level)?

Analysis of Accountability Issues for States

Many unique issues arise for States in refining statewide accountability systems, developed under the 1994 ESEA Reauthorization due to the prescriptive, conjunctive nature of the AYP requirements under the NCLB Act. The essential components of the AYP requirements include:

- An aligned system of academic content standards, academic student achievement standards, and assessments of student performance;
- Annual assessments of student progress in attaining the student academic achievement standards;
- School, district, and State accountability decisions based on the performance of specific subgroups of students designed to ensure that **all** students are proficient in reading or language arts and mathematics by 2013-14; and
- A system of rewards and required, progressive sanctions to encourage and support high- and low-performing schools.

Following enactment of the NCLB Act, an early concern for States was the extent to which existing AYP models would have to be modified. In his July 24, 2002, *Dear Colleague* letter, Secretary Paige signaled a willingness to support these models provided that they integrated AYP as defined under the Act. That position was reaffirmed in the Comments/Discussion accompanying promulgation of the final regulations on accountability:

> The Secretary realizes that the accountability systems currently in place in many States may not fully meet the statutory and regulatory requirements. To meet the requirements in the ESEA and these final regulations, a State may continue to use its current State accountability system consistent with the Secretary's July 24, 2002, *Dear Colleague* letter, **if that system integrates AYP as defined in the statute and regulations** [emphasis added] (Federal Register, 2002, December 5, p. 71740).

According to ED, in issuing the final regulations, the Department's Peer Review process will be used to determine the extent to which a State's "current" accountability system meets the NCLB Act requirements (p. 71711).

The most critical of the NCLB Act accountability issues, as far as their impact on the re-design of State accountability systems are identified and discussed below. These issues are woven in and out of the discussions in Chapters 2 and 3 regarding decisions States must make in order to ensure valid and reliable AYP determinations consistent with the NCLB Act requirements.

- **Whether some States will modify their academic content and student academic achievement standards in light of statewide accountability system requirement changes between the 1994 and 2001 ESEA reauthorizations.** As noted by Linn, Baker, and Betebenner (2002), "Although many states have already established performance standards for their tests, the standards were not set with an awareness that they would be used to determine AYP objectives with the stipulation that all students reach the proficient level or higher by 2014. In a number of cases, the proficient level has been set so high that it may be completely unrealistic to expect all students to reach that level by 2014" (p. 4). The authors further state, "The content standards used by states to develop tests vary in specificity and in rigor. Content standards and associated tests are much more ambitious in some states than in others. The performance standards states have set that determine the cut scores used to define proficient on the test also vary widely from one state to another. The combination of these differences among states regarding their content standards, the rigor of their tests, and the levels of performance required for a student to be considered proficient means that states are not starting on a level playing field" (p. 4).

Kane, Staiger, and Geppert (2002), also forecast the likelihood that States would revisit their standards in light of the high stakes nature of new accountability requirements under the NCLB Act, observing that, "One flaw in the formula [AYP] is that it provides a strong incentive for states to lower the score students must exceed on their state test in order to achieve 'proficiency.' The problem is that redefining proficiency simply because of the new federal requirements may create a credibility problem for the standards movement in a number of states."

Nevertheless, in spite of such a caution, some States have already taken steps to change their standards or what it means to be proficient. In his recent article, "States Revise the Meaning of Proficient," Hoff (2002) describes the steps taken by three States—Colorado, Connecticut, and Louisiana—to "ease" their "standards for what it means to be 'proficient' in reading and math because of pressures to comply with a new federal law requiring States to make sure that all students are proficient on State tests in these subjects within 12 years" (p. 1). In each of these cases, the States reported that standards adopted under the 1994 Reauthorization went well beyond the requirements of the 2001 Reauthorization. As a result, these States believe that their recent changes to not constitute a "watering-down" of current standards.

The NCLB Act continues a provision originally included in the 1994 ESEA Reauthorization addressing such a possibility by providing in Section 1111(b)(1)(F), "EXISTING STANDARDS.—Nothing in this part shall prohibit a State from revising, consistent with this section, any standard adopted under this part before or after the date of enactment of the *No Child Left Behind Act of 2001*." [However, States are reminded under section 1111(f)(2), "ADDITIONAL INFORMATION.—If significant changes are made to a State's plan, **such as the adoption of new State academic content standards and State student achievement standards, new academic assessments, or a new definition of adequate yearly progress, such information shall be submitted to the Secretary** (emphasis added)."]

- **Minimum "n" determinations.** The NCLB Act requires States to determine the number of students in a group necessary to yield statistically reliable information as well as the number of students required to be in a group to ensure that the results will not reveal personally identifiable information about an individual student. (States must report this determination to the U. S. Department of Education by January 31, 2003.) State decisions here are likely to have a significant, but not long lasting, impact on the number of schools and districts initially identified for improvement. Higher minimum "n's" may initially cause some schools or districts with low-performing student subgroups to not be identified for improvement but, over time, that "masking" may disappear when performance data are examined using "uniform averaging" provisions under section 1111(b)(2)(J).

  States should not approach their minimum "n" determinations solely from this perspective. The result will inevitably be the search for the highest possible "n" some States may feel they can support or justify rather than a more concerted effort to address the intent of the NCLB Act—helping and ensuring that all students become proficient learners. As addressed in Chapter 3 of this paper, such an approach can lead to several adverse, unintended consequences. Smaller "n's" may lead to the inclusion of more students in accountability determinations but, perhaps, less reliable AYP decisions. Conversely, larger "n's" may lead to more reliable accountability decisions but fail to identify schools (or districts) for AYP where significant needs exist (a consequential validity issue).

- **How to include small schools and small subgroups in a valid and reliable school accountability system that meets the letter of the NCLB Act.** Many States are concerned about how to make valid and reliable decisions about student performance at the school building level when the number of students in various subgroups is small. In these cases, some schools may be judged as having met an AYP determination simply because of the small number of students enrolled for a full academic year taking the assessments at a given grade level even if the students consistently fail to meet AYP requirements.

  Consider the simple example of a K-6 school with a total of 29 students in 4th grade—100% of those enrolled in the school for at least one full academic year—taking the State assessments when the State requires a minimum "n" of 30. In this example, the number of students taking the assessments is insufficient to make a valid and reliable decision about the status of the school so the school is, for all practical purposes, judged to have met AYP in the absence of evidence to the contrary. In districts with small schools like this one, the result could be that a district may not meet its accountability targets while all its schools are considered to have made AYP because no other determination is possible under the circumstances.[3] A State with several small districts may find itself in a similar situation as performance data are aggregated across public schools and districts statewide.

  Another example might involve a small school (or district) where the number of ALL students exceeds the State's minimum "n" but the number of students in some or all subgroups is less than the State's minimum "n." In this example, an AYP determination could still be made for the school (or district) provided that the results of any subgroup are aggregated into the ALL students group and they are "included at the next higher level assuming the subgroup reaches the appropriate size" (Federal Register, 2002, December 5, p. 71743). In a similar situation, States may want to develop policies that would result in tracking and making AYP determinations regarding the performance of small subgroup populations across two or three school years if that would result in a group size sufficient to make valid and reliable accountability decisions (and then keeping rolling up data to enable annual determinations).

- **What assessments to include in the statewide accountability system.** The NCLB Act requires, under Section 1111(b)(2), that States design and implement a single statewide accountability system. The practical effect of changes made to AYP requirements in the NCLB Act compared to those set forth in the 1994 ESEA Reauthorization may well prove to result in States making more accountability decisions (e.g., at least 8 student subgroup performances versus performance decisions made solely on the basis of all students) based on fewer assessments (no multiple measures). This will serve to further challenge ensuring the reliability and validity of the resulting decisions. Basing accountability decision on fewer, perhaps even less challenging, assessments may also serve to lower overall expectations for improving teaching and student learning.

[3] It is important to recognize that a State's obligation to annually determine school and district accountability under the law does not end in such instances. If no determination can be made, at least for all students, for the school because of minimum "n," the State must develop additional policies and procedures to ensure that each school (and district in the case of very small LEAs) can still be reviewed annually for effectiveness consistent with the intent of the law. One option might be "rolling up" two consecutive years of student performance data to establish a sufficient minimum "n" to make a valid and reliable accountability decision. It is hoped that theEDwill develop related guidance and provide additional examples to assist States in this matter.

States that currently include assessments in areas other than reading or language arts and mathematics to determine school and district performance are likely to consider changing those requirements and include only the assessments prescribed by the NCLB Act. States, in the absence of other State laws, are not required to include other assessments in AYP determinations under the NCLB Act. (Results of the science assessments required to begin in 2007-08 do not have to be included in the State's AYP definition and determinations.)

Dropping existing assessments from State accountability decisions could permit some to argue that such decisions serve to narrow the curriculum and diminish the importance and contribution of other academic areas to students' overall education. On the other hand, as with the inclusion of additional academic indicators in AYP determinations, the use of additional assessments can only have negative consequences—the results can be used to identify more schools and districts for improvement, corrective action, or restructuring; they cannot be used to compensate for low student performance on other measures required under the NCLB Act.

- **Aligning State and Federal definitions related to student academic achievement standards.** Both the 1994 and 2001 ESEA reauthorizations require States to set at least three levels of student academic achievement standards (referred to as "performance standards" in the 1994 reauthorization). It is likely that some States, as a part of reviewing their standards, will re-evaluate the definitions they previously developed to describe student academic achievement levels. States may, consistent with the NCLB Act requirement for a single statewide accountability system, also find it advantageous in communicating information about students' academic performance to various publics to use consistent "labels" and definitions. This might help ensure that student, school, district, and State reports are communicated using terms that various publics are accustomed to seeing in other media regarding the NCLB Act.
- **Differentiated responses (sanctions) based on the extent to which schools or districts do not make AYP requirement.** This issue concerns whether States may—and if so, the extent to which they may—provide for differentiated responses (sanctions) given the extent to which a school or district identified for improvement continues to remain in that status. Although not addressed in this paper, we raise this issue here in connection with AYP determinations because it addresses one of the primary concerns surrounding the prescriptive nature of the NCLB Act requirements—how to fairly and equitably treat schools and districts when the range of "need for improvement" is as considerable as it will be under the law.

  What differentiation of sanctions might be appropriate, given the real difference between schools that fail to meet AYP due to essentially random events as opposed to schools that demonstrate consistent patterns of failure with the total group or a specific student subgroup over time? In his July 24, 2002, letter, Education Secretary Paige stated, "States are free to build on the statutory requirements and to develop differentiated responses based on the degree to which a school has not made AYP. The law does not prescribe how States must officially designate schools that do not meet AYP requirements." Nevertheless, under the NCLB Act, the consequences (and resulting sanctions) of missing one or multiple numbers of AYP determinations is the same—the Act does not differentiate based on the extent to which AYP targets are missed. Schools identified for improvement based on missing one of 37 performance decisions must provide for public school choice and supplemental educational services in the same manner as schools that miss all 37 (as in the example in this chapter).

Further, the law, in Section 1116, sets forth a set of prescriptive, substantial steps that apply once schools or districts have been identified for corrective action or restructuring. LEAs or SEAs **must** implement at least one of the prescribed actions. The August 2002 draft and the December 2002 final regulations on accountability did not address this issue and the ED has not yet signaled whether States will eventually be permitted some latitude in applying sanctions, especially those related to corrective action or restructuring, under the NCLB Act.

- **Other Issues.** The 2001 ESEA reauthorization has also resulted in renewed debate regarding (1) the use of Standard Error of Measurement in connection with interpreting cut scores related to student academic achievement standards and (2) the decisions States might need to make regarding the use of uniform averaging in AYP calculations.

Initial State Analyses

Several States have already analyzed the effect of varying the levels of group size on reliability and on the percentage of schools identified as not meeting AYP. The findings and initial conclusions reached regarding these State analyses are presented in Chapter 3. Examining the probable impact of setting various minimum "n" sizes was also the subject of a Joint SCASS ASR-CAS meeting in May 2002 in Salt Lake City, Utah. Several States presented their initial findings in this regard using prior years of student assessment results to estimate likely AYP outcomes had the NCLB Act been in effect 3 or 4 years earlier. Among the findings were (1) confirmation that raising the minimum "n" to levels high enough to have a noticeable effect on reliability would require samples so large that it would be impractical for many States to set such high thresholds and, assuming that States establish minimum "n's" that are practical, then (2) most States will find a high percentage of schools identified for improvement when the conjunctive model or approach required under the NCLB Act is the only one used in determining AYP regardless of the minimum "n" used.

Variables Impacting Accountability System Design Decisions.

Based on related discussions at various conferences, meetings, and so forth as well as a thorough analysis and current understanding of the NCLB Act provisions, there are a number of variables in addition to minimum "n" that will ultimately impact the decisions States must make with respect to developing and implementing their accountability systems for determining AYP. Most, but not all, of these variables are addressed in the following chapters of this paper and include:

- Modifying State assessment systems including developing new assessments, dropping assessments in areas not covered under the NCLB Act, and re-evaluating alignment with academic content and student achievement standards.
- Defining student academic achievement labels (what it means to be Basic, etc.).
- Student academic achievement standards—where the cut scores are set and how they are applied.
- AYP decisions based on all students in a school or on only students receiving Title I services in such schools (which would be all students in schoolwide projects and students with the greatest need for academic assistance in Targeted Assistance schools).
- The impact of 'uniform averaging procedures" in AYP determinations.

- Determination of starting points including whether States decide to average student performance data across two or three years school years providing that 2001-02 data are included in the average.
- Setting the 2001-02 baselines to be first applied in 2002-03 (one each for reading or language arts and mathematics—two starting points OR as many as six starting points based on one for each of the two subject areas in the 3-5, 6-9, and 10-12 grade spans?) The final regulations on AYP provide that States can do either [§ 200.16(c)].
- How the State sets its Intermediate Goals for AYP—the goals must increase in equal increments over the timeline 2001-02 to 2013-14 and each incremental increase must occur in not more than three years. States must set at least four Intermediate Goals (2004-05, 2007-08, 2010-11, and 2013-14). States can also set more than four Intermediate Goals as long as each increases in the same amount within the timeline. An example might be having six Intermediate Goals set in this pattern: 2004-05, 2007-09, 2010-11, 2011-12, 2012-13, and 2013-14 (provided that the incremental proficiency increase is the same at each point, e.g., 8%). In this example, the annual measurable objectives and the Intermediate Goals will be the same from 2011-12 through 2013-14.
- School and District size (especially small vs. large districts).
- Establishing the cell size/minimum "n" (a single "n" for all student subgroups and AYP decisions).
- Establishing the cell size/minimum "n" (two or more "n's" such as one for schools and one for LEAs, again a variable not addressed in the final regulations).
- Students counted in multiple performance determinations (up to 21 of 37 measures in the example illustrated in this chapter). This is primarily an "awareness" issue—the extent to which the performance/subject mastery of some students can impact two or more AYP measures.
- How long the performance of LEP students and Students with Disabilities (SWDs) no longer receiving services in those areas may continue to be included in disaggregating data and making related AYP determinations. Such inclusion does not seem to be prohibited under either the Act or the regulations.
- Whether the identification of schools and districts for improvement will be based on "missing" the same AYP requirement two consecutive years or "missing" any AYP requirements for two consecutive years. As noted earlier in this chapter, this variable does not seem to have been directly addressed in the final regulations.
- Validity and reliability of the system/decisions.
- The outcomes of the required input/agreement of various publics in developing the accountability system.
- A State's structure for implementing Title I provisions regarding the opportunity to review and present evidence—see sections 1116(b)(2) and 1116(c)(4) permitting a school and a district identified for improvement, respectively, to review the data and to present evidence if they believe that the identification is in error (see also Appendix B).
- Capacity of the SEA and LEAs to provide assistance when large numbers of districts and schools identified for improvement each year.

- The extent to which LEAs and SEAs will have latitude in applying corrective action and restructuring requirements under Section 116 given the wide range of need (from minimal to severe) evident in schools and districts that continue to be identified for improvement.

Organization of the Paper

The following two chapters comprise the "heart" of this paper and address the paper's central issue—**the NCLB Act requirement that States ensure that their single, statewide accountability systems will result in valid and reliable decisions in determining the adequate yearly progress of schools, districts, and States themselves.**

Chapter 2 serves as an essential, critical foundation to the discussions that must follow in each State as options and alternatives are considered in the development, refinement, or revision of statewide accountability systems in order to ensure that the systems will indeed result in valid and reliable decisions about student, school, district, and State performance. It establishes the groundwork by presenting a general description and analysis of the concepts of validity and reliability as they apply to State educational accountability systems. The design of accountability systems is approached from a basic theoretical point rather than focusing specifically on the NCLB Act requirements arguing that there is great need for intellectual clarity about the scientific origins and definitions of these concepts before applying them to the requirements of the law.

Chapter 3 focuses on a decision process and framework for States to consider as they develop their processes and procedures for defining and calculating Adequate Yearly Progress consistent with the NCLB Act requirements. Results of State simulations and data analyses to project the likely impact of the new AYP requirements over the next few years are also presented. The authors focus their analyses and suggestions on helping State accountability system designers minimize potential negative consequences while maximizing the potential outcomes of this law. Finally, critical factors impacting the design of AYP parameters including the determination of starting points, minimum-n, using confidence intervals, standard error approaches, the aggregation of data, and establishing annual and intermediate measurable objectives are discussed.

Chapter 4 provides an the overview of the essential findings and points identified in the prior chapters and summarizes conclusions reached by the Study Group regarding a number of the critical variables States must consider in finalizing these systems.

Four Appendices are included at the end of the paper:

1. **Appendix A** lists excerpts from the NCLB Act related to educational accountability systems requirements.
2. **Appendix B** outlines information on using the "safe harbor" and "opportunity to review" provisions under Title I of the law to substantiate or reverse accountability decisions.
3. **Appendix C** is excerpted from another CCSSO publication and offers ten critical questions and their underlying major considerations for educational policymakers to guide their deliberations and planning as they design and implement statewide educational accountability systems.
4. **Appendix D** has a glossary of terms unique to educational accountability system components and requirements.

References

Education Commission of the States. (2002). *No state left behind: The challenges and opportunities of ESEA 2001.* Denver, CO: Author. [On-line]. Available: www.ecs.org.

Erpenbach, W. J., Carlson, D., LaMarca, P. M., & Winter, P. W. (2002). *Incorporating multiple measures of student performance into state accountability systems; A compendium of resources.* Washington, DC: Council of Chief State School Officers.

Federal Title I [Draft] Regulations for the U. S. Department of Education, 67 Fed. Reg., 50986-51027 (2002, August 6), (to be codified at 34 C.F.R. pt. 200).

Federal Title I [Final] Regulations for the U. S. Department of Education, 67 Fed. Reg., 71710-71771 (2002, December 2), (to be codified at 34 C.F.R. pt. 200).

Gong, B. (2002). *Designing school accountability systems: Towards a framework and process.* Washington, DC: Council of Chief State School Officers. [On-line]. Available: www.ccsso.org

Hill, R. (1997, June). *Calculating and reducing errors associated with the evaluation of Adequate Yearly Progress.* Paper presented at a meeting of the CCSSO Annual Large-Scale Assessment Conference, Colorado Springs, CO.

Hoff, D.J. (2002, October 9). States revise the meaning of "proficient." *Education Week*, 1, 24-25.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). *Randomly accountable.* [On-line]. Available: http://www.educationnext.org/20021/56.html

Neuman, S.B. (2002, December 5). *Letter to chief state school officers*. Washington, DC: U.S. Department of Education.

Paige, R. (2002, July 24). *Dear colleague letter to education officials regarding implementation of the No Child Left Behind Act of 2001.* Washington, DC: U. S. Department of Education. [On line]. Available: http://www.ed.gov/News/Letters/020724.html

U. S. Department of Education (2002, December). *Consolidated state application accountability workbook*. Washington, DC: Author.

U. S. Department of Education (2002, June 10). *Review guide for Title I, Parts A, B, and D, and Title VI, section 6111, of the 2002 Consolidated State Application.* Washington, DC: Author.

Chapter

2

Improving the Validity and Reliability of State Accountability Systems

"Validation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle, combining the attractions of chess and mud-wrestling" (Cronbach, 1988).

The *No Child Left Behind* Act of 2001 requires States to develop valid and reliable State accountability systems. Although the references to validity and reliability are ubiquitous in the law[4], our focus is on three specific areas: assessment systems[5], accountability systems, and definitions of Adequate Yearly Progress. The law specifies the following:

- Both the assessments and the other indicators that a State chooses to use are to be "valid and reliable, and…consistent with relevant, nationally recognized professional and technical standards."
- Each State's definition of Adequate Yearly Progress must be "statistically valid and reliable."
- Although each State is required to disaggregate the results for specified groups of students, it specifically relaxes this requirement in cases where "the number of students in a category is insufficient to yield statistically reliable information…" [Section 1111(b)(I)(ii)].

The goal of this chapter is to present a general description and analysis of the concepts of validity and reliability as they apply to State educational accountability systems. This will entail a description of State accountability systems (of which assessment systems are a component), and an analysis of their relationship to States' overall educational/instructional systems. It will deal with accountability systems in a general way, rather than focusing specifically on the requirements of

[4] In fact, the law uses the words in this phrase 59 times. In 21 of those references, the law requires something to be "reliable" and to have "validity" 18 times. Another 20 times an assessment or process is required to be either "valid and reliable" or "reliable and valid."

[5] The treatment of validity and reliability of assessment systems is only touched upon tangentially since they are viewed as relatively separate systems. Furthermore, the U. S. Department of Education has set forth separate requirements for assessment systems under pending regulations.

the NCLB Act. This is not because the new law will not dramatically alter the shape and magnitude of assessment and accountability programs, but because of the need for intellectual clarity about the accepted scientific origins and definitions of these concepts before applying them to the vision of accountability as set forth in the NCLB Act.

This chapter is intended to provide useful ideas or criteria for evaluating the validity and reliability of State accountability systems. More specifically, it is designed to serve as a foundation for the discussion of the options (topics) and alternatives that State personnel will be considering as they design their accountability systems. Some of these options (topics) are discussed in Chapter 3 in connection with an AYP Decision Framework:

- ***Nature of data sources,***
- ***Number of starting points,***
- ***Sample size issues,***
- ***Aggregation issues, and***
- ***Setting intermediate goals.***

Readers are encouraged to keep these topics in mind as they read this chapter.

This chapter has four main parts:

Part 1. First, the concepts of validity and reliability will be defined, beginning with their usage in non-educational fields, and moving to their extensive use in the world of testing and assessment.

Part 2. The concept of a State educational accountability system is then discussed in some detail.

Part 3. These two strands will be brought together to see how the concepts of validity and reliability apply to accountability systems. This includes a definition and a suggested set of procedures for judging the relative validity and reliability of a given system.

Part 4. Finally, some illustrations of the process are offered, especially by looking at the ways in which validity and reliability can be thwarted.

It is important to note what this chapter will <u>not</u> do:

- *It will not describe the best approaches and steps in developing an accountability system. Some useful documents are included in the list of references, however.*
- *It will not review, in any depth, the procedures for checking the validity and reliability of assessment systems (as opposed to accountability systems).*

Ten key points that will be made in this chapter are listed below. The reader is encouraged to watch for them and to evaluate their validity:

1. An accountability system must be judged in accord with its purposes.
2. Since an accountability system is a system, it is crucial that all parts of the system work together in a coherent fashion, and that they work toward the same goals and purposes.
3. An accountability system is not the same as an assessment system, but it relies on the results of the assessment system.
4. The method of analyzing student achievement results (the accountability model) is a major issue since it relates directly to the goals of the system, the ways in which they might be accomplished, and the ways of judging the system's validity/effectiveness.

5. Establishing the validity of a system is not that different from establishing the validity of a measure, a conclusion, or a hypothesis in medicine or science.
6. In judging a system, it is just as important to look for unintended—both negative and positive—consequences as for intended consequences.
7. Both validity and reliability are important, but validity is more important.
8. Establishing the validity of an accountability system is a complex process, requiring the collection and analysis of a variety of information.
9. Summarizing the results is a process of determining the preponderance of evidence for a conclusion or inference.
10. Different persons, using different criteria, will reach different conclusions about the validity of a given system.

Part 1. What Is Validity? (*and What about Reliability?)*

Validity and reliability have come to symbolize the ideal criteria for judging for all educational efforts and programs, especially those related to assessment and accountability. They have become a veritable mantra for our times. But what do they mean?

The two terms, validity and reliability, are often used interchangeably. In this general sense, they are both synonyms for truth, accuracy or trustworthiness. Furthermore, even in the professional literature there is considerable confusion; characteristics that one writer attributes to validity another ascribes to reliability, and vice versa (Winter, 2000). The characteristic that is more frequently linked to reliability, however, is that of replicability, dependability, or consistency—especially in the realm of psychological and educational measurement. It is often defined as the likelihood or probability of a given result or finding occurring under repeated observations or administrations. The classic example is that of the probability of a student earning the same score if a test were to be repeated. Other examples will be discussed in the context of accountability systems. This chapter will, therefore, observe that tradition; reliability will be discussed as a component of validity—the lack of which is one of many obstacles to validity.

The Need to Study Reliability of Accountability Systems

This is not to say that reliability is not of importance, in fact, the reliability of the results for accountability systems—typically taking the form of ratings or classifications—has not received the attention it deserves. Hill (2001; 2002) has shown why this is extremely unfortunate; the reliability of most accountability systems has never been estimated, and where it has, the results have been very disappointing. It is important to study the likelihood of making different types of mistakes (Rogosa, 2002). In fact, one could argue that a State's description of its accountability system must include a discussion of the State's position on the relative merits of falsely identifying good schools (false positives) versus failing to identify needy schools (false negatives). Both of these will occur, even for the best systems with the best data, but it is possible to tilt the system toward minimizing one of these, usually at the expense of the other. It is important that the balance be the product of rational discussion, rather than coming as a surprise at some later point, or worse—never knowing what the balance is.

The traditional—and logical—view is that reliability is a necessary but not sufficient condition for validity. If one of the measuring instruments provides random, or partially random measurements, for example, it would hardly be possible to trust any overall changes in scores or performance, whether of a student or of a school or a system. This is one reason why reliability is treated as one aspect of validity in this chapter—because it is.

Too Much Reliability.

On the other hand, reliability can undermine validity. Too much emphasis on reliability can be undesirable. How could this be? Two examples are sufficient to show how that can happen. The first example is drawn from assessment. If the process of instrument development relies too heavily on maximizing the overall reliability of a student-level score, the temptation is to narrow the range of the content and the skills on the assessment, which leads to a higher reliability coefficient.[6] Unfortunately, this narrowing lowers the validity of the assessment.

The second example is noted in Chapter 3 in the context of selecting a minimum number of student scores in a given subgroup as a basis for deciding how to treat small schools in the definition of AYP. It is shown that if one were to blindly use a minimum cutoff, some small schools—which actually do have low performance and should be identified for improvement—would be excused from the accountability rules. This would be a direct violation of the validity goal; a valid system identifies all the schools that should be identified and does not identify those that shouldn't be.

The reader was alerted earlier to the topics discussed in Chapter 3 (nature of data sources, number of starting points, sample size issues, aggregation issues, and setting intermediate goals). It will be seen that some of these relate directly to validity and some relate to the reliability of the data or the decisions, which, in turn, relate to validity. Either way, the end-game is validity.

A Definition of Validity

A quick scan of the web produced thousands of hits for the word "validity." Five of the more colorful yet serious references to validity include:

- Patterns studied by the American Polygraph Association (Hurlock, 2000);
- The pain ratings in the National Pain Data Bank (Clark & Gironda, 2000);
- The special theory of relativity (Schewe, Riordon, & Stein, 2002);
- The dictum that one ought to drink eight glasses of water each day (Drink, 2002); and
- The Lithuanian Whiplash Study (which claimed to find that whiplash incidents didn't seem to occur in countries without trial lawyers) ("Dynamic," 2002).

Two respected dictionaries provide remarkably similar definitions. Both Webster's Third International (1977) and the New Shorter Oxford Dictionary (1993) offer at least three main meanings. The first is that of being "strong" or "having legal strength or force;" the second is that of "being well grounded, sound and defensible;" and the third is that of "effectiveness, efficacy, and able to accomplish what is designed or intended."

Seldom have dictionary definitions been so helpful. It can be shown that all three of these facets are involved in a definition that might be applicable to accountability systems. This is important because there is no agreement on the meaning of validity of accountability systems. In fact, there is virtually no body of literature that can be summoned. In lieu of an accepted definition, the strategy will be to show how the word is used in some respected areas of human endeavor,

[6] This occurs because the indices are not really measures of stability or consistency over time, but are measures of the homogeneity of items within an assessment. It is both logical and true that items that use the same format, for example, will yield high "reliability" indices, and tests that contain a variety of formats and/or measure a wider range of subject matter or skills will yield lower indices. As always, a balance must be sought.

including science, medicine, and law. The word will then be presented as it has evolved—through very wide use—in the testing and assessment arena. From these two platforms, a working definition for accountability systems will be proposed and applied for the reader's consideration.

Validity in the Real World

A common definition in the world of research is often stated as "the degree of correspondence between a measurement and the phenomenon under study." *The Glossary of Clinical Epidemiology and Evidence-Based Medicine* (2002) defines validity as the best possible approximation of the truth. Noted researchers (Cook & Campbell, 1979, p. 12) define it as the "best available approximation to the truth or falsity of a given inference, proposition or conclusion."

In some areas of human endeavor, the process of establishing the validity of an assertion is as easy or as difficult as establishing the validity of a document. The business and legal world relies on valid documents and has developed standard procedures for verifying the authenticity of documents, ranging from drivers' and marriage licenses to wills, property titles (and corporation annual reports). While not foolproof, these procedures are relatively straightforward, and the outcome is binary—they are judged as either valid or invalid. This process differs markedly from the procedures used in other areas of law, or in medicine or science (or in educational accountability). In these fields, the process is more complex and the outcome is seldom as clear cut.

In spite of decades—if not centuries—of study and debate, the legal community is still struggling with the problem of establishing the validity of evidence. In this day of heavy reliance on expert testimony, the burden on judges to evaluate the validity of the testimony has raised the need for new rules and guidance for that process (Faigman, 2001). Only relatively recently did the American Association for the Advancement of Science and the National Academy of Sciences file an amicus curiae (friend of the court) brief to the Supreme Court, to set guidelines for determining what scientific evidence is admissible in a court of law. The brief states that the "courts should admit scientific evidence only if it conforms to scientific standards and is derived from methods that are generally accepted by the scientific community as valid and reliable" (Francis, 1993).

The search for validity (e.g., the process of validating evidence in the field of medicine) is becoming increasingly crucial. The amount of evidence is accumulating at an alarming rate. It is estimated that medical knowledge is increasing exponentially, doubling in its content every 10 years. To keep up with the 10 leading journals in internal medicine alone, it is necessary for the clinician to read as many as 200 articles and 70 editorials per month. Practicing physicians are expected to be aware of the essential information contained in the more than six million articles published in some 20,000 biomedical publications annually. Medical science is aware of more than 30,000 diseases that can present a near infinite number of combinations (Savitha, 1999).

In selecting an appropriate treatment (out of 15,000 available therapeutic agents), the physician must cut through the confusion about the data on benefits and risks of various treatment options. This usually stems from the fact that the quality of the supporting evidence varies considerably, often leading to completely opposite and contradictory recommendations. This information overload has led to the field of Evidence-based Medicine, defined as the process of systematically identifying and summarizing information from the scientific literature and applying the results to clinical practice ("Definition," 2002).

The solution is not simple. It is a matter of integrating an entire body of relevant medical research and then assessing the strength of that collection of research. Conclusions of any synthesis of indirect research evidence are inferential and based on a combination of facts, arguments, and analogies. One writer puts it this way, “Integrating evidence is invariably a subjective process, dependent on the skills and values of the individuals who are trying to synthesize multiple pieces of diverse medical evidence. The emphasis must be on consistent and coherent results across multiple types and sources of evidence (“Proof and Policy,” 2001).

The goal is to attain reasonable assurance of the validity of a claim. Carnap (1966) reminds us that total certainty is out of our reach; the mere existence of one counter-example falsifies a law or a conclusion. There is no choice but to form the best possible conclusion by rigorously applying the strongest summarization rules of logic to a body of evidence. The various pieces of theoretical, logical, and empirical evidence need to be woven into a coherent argument. This is not easy. Cronbach said it was “doing your damnedest with your mind, no holds barred” (1988—in paraphrasing Eddington who was referring to science).

The price of not attending to the validity of the evidence for medical claims can be so severe as to be newsworthy. The furor over problems caused by breast implants may be the most well-known example. Although the number of claims awarded by the court took the main supplier to bankruptcy, the European Committee on Quality Assurance and Medical Devices in Plastic Surgery reported that the evidence was “conclusive” that implants did not cause autoimmune or connective tissue diseases, and that “there is no scientific evidence” of other silicone maladies (Bandow, 1999).

Validity in the World of Testing and Assessment

Validity and reliability probably have a longer and more extensive history of use in testing and assessment than in most any other field. The meanings of these terms, however, especially that of validity, have undergone major changes. The classic definition of validity as given in introductory textbooks, and dutifully memorized by students is, “A valid test is one that measures what it purports to measure.” Ah, there’s the rub—how does one know if a test measures what it claims to? The efforts to answer that question have followed several paths over the years. Three of these main waves or phases will be examined briefly.

The First Wave—Criterion and Content Validity.

When mental testing was a new endeavor, the main purpose of a test was to predict a person’s chance of success in a job, or in college or some training program. There was a criterion; a person either succeeded in college or on the job or didn’t—or succeeded to some quantifiable degree—and this criterion could be used to *validate* the success of the test in predicting a candidate’s success. Hence, the term “predictive validity,” one of several forms of criterion-related validity, came to be used.

The second main strand of this first phase originated with the use of tests in education. If tests were to be considered as measures of the quality of learning or teaching, it was natural that people would define validity in light of the degree to which a test adequately covers or samples the content that is specified in the curriculum documents. This area of “content validity” is now typically included in the discourse on alignment. Both of these two early forms of validity were set forth clearly in the first set of technical standards for testing (APA, 1954).

The Second Wave—Construct Validity. Over the years it became obvious that the validity of a test is a function of the use of the instrument and the results. A test could be valid for some uses and invalid for others. In fact, validity doesn't actually pertain to the test or the data, it pertains to the *inferences* that are drawn and the decisions that are made, which are a function of the purpose or use of the assessment and the results. It refers to the process of defining the "construct"—the trait or concept in question.

Testing historians point back to an article by Cronbach and Meehl (1955) as the origin for the idea of construct validity. They set forth the concept of a "nomological network"—the organized set of data and hypotheses, bound together with a set of interpretive rules that could be used to define the nature of a construct. Or, as Messick (1988) stated, "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" (p. 89).

The need for the idea of construct validity is obviously a necessary tool in the process of identifying various personality traits, for example. But it is no less important in the areas of schooling that seem to be too clear and straightforward to need such a complicated framework. Even a well-known topic such as "reading comprehension" is actually a construct (i.e., it obviously cannot be seen or weighed or measured in any direct way). It must be inferred on the basis of the kinds of questions students can answer after they read a passage (the most popular form of reading assessment).

This understanding of validity is now the standard view, as embodied in the most recent set of standards for psychological and educational testing:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. . . .The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated (AERA, 1999, p 57).

The Third Wave—Consequential Aspects of Validity. Although test users have long been advised to consider issues of adverse impact and test bias, and to think about the probably consequences of their assessments, it was Messick (1988; 1989) who laid out the theoretical arguments for placing these issues at the heart of validity. He emphasized the values connected with the measurement and the potential social consequences of the measurement for examinees. The use of an unfair test could, for example, have negative effects for different ethnic groups, and it could also produce unfair gender differences.

A poignant historical example was the use of verbal ability tests to select jurors, until the process was banned on the grounds that it resulted in very unrepresentative panels. Others have been concerned about the overuse of machine-scored assessments and the possible impact on students' ability to construct arguments and solve the kinds of problems they will face in the world of work. It is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal.

Messick (1989) saw validation is a process consisting of four aspects. According to his model, the main questions for the validation process are:

- How does the measurement correspond with the theoretical construct?
- Is the measurement relevant for the specified purpose?
- What values can be connected with the construct and the measurement?
- What are the consequences of the measurement?

One can see that validation has changed from being the final “quality check” in research to becoming a central, never-ending process. During the research process, it has become necessary to use multiple techniques to continuously value, question, and check the inferences and interpretations that are being made.

Although there are still some dissenting voices against this consequentialist view of validity, to some degree it is a matter of semantics. All parties agree that the impact of assessment needs to be systematically examined. It is a question of (a) whether it should be considered a part of validity or whether it is a separate process, and (b) who should be responsible for consequences.

This third wave might better be identified as the era of a unified view of validity. The focus on consequential aspects of the inferences and the decisions is not separate from the nature of the construct, and the process certainly relies on the various types of criterion-related evidence that is available.

Part 2. Defining a State Educational Accountability System

Accountability has many meanings and applications in public education. This paper deals primarily with the relatively new concept of a State educational accountability system, as seen in the ESEA reauthorization of 1994 and ratcheted up considerably by the NCLB Act. This chapter views State accountability systems from both their temporal and their structural aspects. First, the three-phase nature of these systems will be explored, and then they will be analyzed according to six components or elements.

An accountability system exists to improve the functioning of the educational system. It is part of the public educational enterprise, but (as argued here) it is not a part of the educational system, per se. One could make the case that as a feedback and control function, the accountability system is an integral part of the whole educational system. This document takes the other position, arguing that it is more productive to think of the accountability system as a quasi-separate system. Figure 1 illustrates how it exists along side or “hovers over” the educational system. The arrows show how it depends on data that are part of the on-going operation of the educational system, and how it feeds data and actions back into the educational loop.

The Three Phases of an Accountability System

It is argued here that State accountability systems are best viewed as three-phase systems, as illustrated in Figure 1. An accountability system is an information system and an indicator system, but it is more than that. It is also an intervention, in both a general sense and a specific sense. In a general sense, it exists (especially as viewed by the authorizing policymakers) as a source of motivation for school personnel, which is expected to lead to greater effort—and along with other mediating steps, is expected to lead to greater student achievement. In a specific sense, it exists as a set of rules for classifying schools and triggering various reform actions, such as school choice.

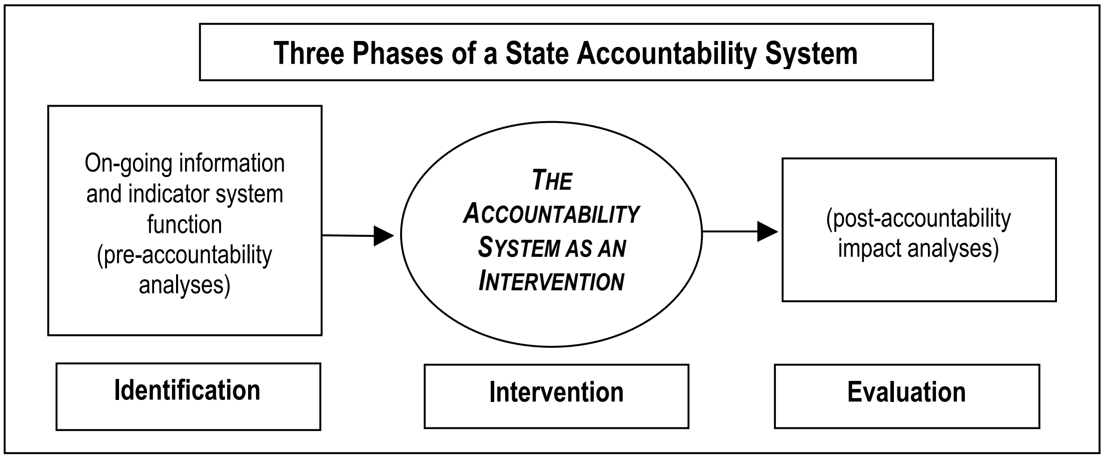
Both of these aspects are important, although it may be virtually impossible to separate their effects.

The three phases are the

1. Pre-intervention, identification phase;
2. Intervention phase; and
3. Post-intervention, evaluation phase.

The information that the system collects and reports pertains to decisions that are made in the first phase and implemented in the second phase. That and related information is then used to evaluate the impact of the interventions, especially the specific reform elements. In its pre-intervention role, it monitors the general state of student learning. After its introduction as an aspect of the school reform program, it serves as a post-intervention measure of the impact of the reforms, including the impact of its own existence.

FIGURE 1. THE "BEFORE, DURING, AND AFTER" PHASES OF AN ACCOUNTABILITY SYSTEM


Three Phases of a State Accountability System
On-going information and indicator system function (pre-accountability analyses)
THE ACCOUNTABILITY SYSTEM AS AN INTERVENTION
(post-accountability impact analyses)
Identification
Intervention
Evaluation

The information that an accountability system reports is also differentiated by level—statewide or school-specific. Different types of questions need to be addressed at each level as illustrated in Figure 2.

FIGURE 2. QUESTIONS TYPICALLY RAISED AND ADDRESSED AT EACH PHASE AND LEVEL OF A STATE ACCOUNTABILITY SYSTEM

	Pre-Intervention Phase	Interventions	Post-Intervention Phase
State-Level Applications	What is the general level of achievement in our state? How many schools are not meeting the criteria, not meeting AYP? How does our State perform relative to other States? What are the patterns of strengths and weaknesses for our students?	**Identifying schools for different kinds of assistance, sanctions, and rewards.**	Searching for both intended and unintended consequences: **Positive impact:** • Are the schools improving? • Are they improving as a result of the accountability system—is the system having the right impact? **Negative impact:** • Are more students dropping out? • Are good teachers leaving the impacted schools? • Is the public perception of the quality of public schools dropping?[7]
District-Level Applications	What is the general level of achievement in our district, as a distribution of schools? What are the patterns of strengths and weaknesses for our students? How many schools are not meeting the criteria, not meeting AYP (on a pre-accountability analysis)?		Did the district make its goal? If not, is it eligible for the next stage of the accountability ladder of sanctions? Which schools made AYP, and to what degree did they drive the district results? How should the district allocate its Federal and State resources to improve the schools' programs?
School-Level Applications	Which schools are in need of improvement, that is, are not meeting some (pre-accountability system) criteria? What are the patterns of strengths and weaknesses for our students?		Did the school make its goal? What rewards or sanctions are appropriate? (Each question in the State-Level Applications box above should be asked about each school and district.)

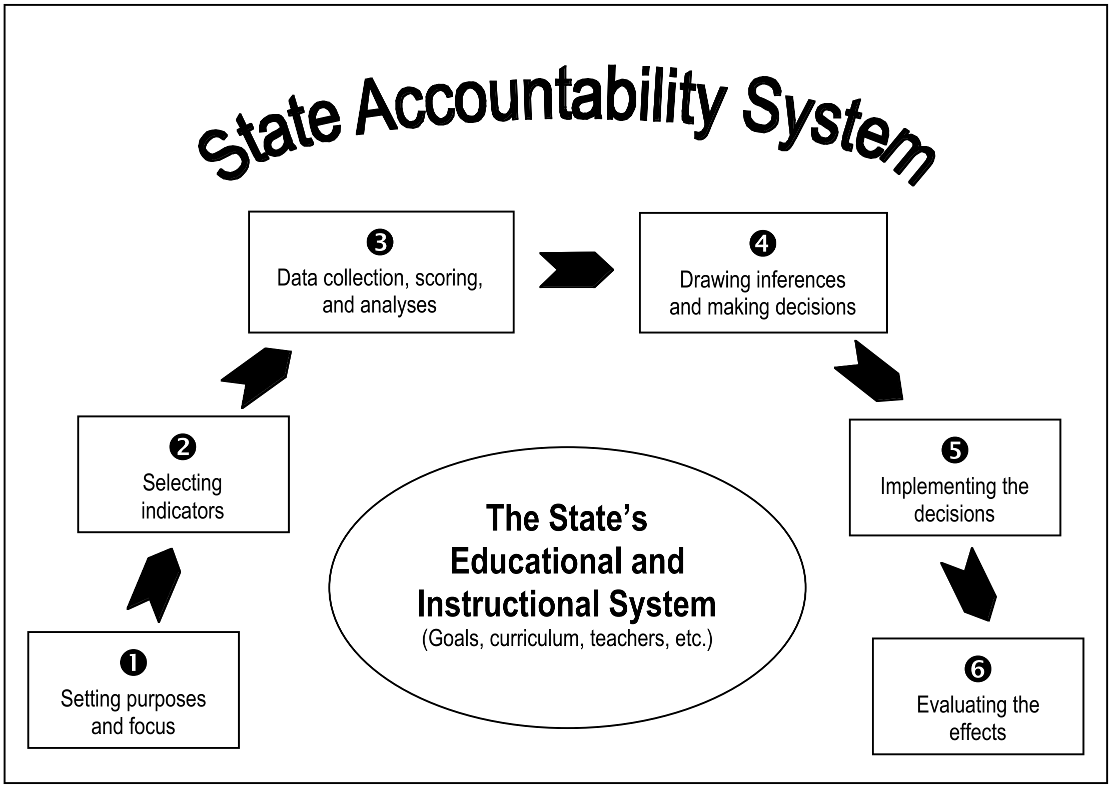
Accountability Elements or Components

As systems, accountability systems have purposes and have parts or elements to help the system reach those goals. While systems have approximately the same number of elements or components, different writers parse them in different ways. The most common focus is on the large chunks: standards, assessments and rewards/sanctions. In an effort to help those who must design an accountability system, Gong (2001) listed 10 questions or issues that must be answered,

[7] See the standards for accountability systems proposed by CRESST for a discussion of unintended consequences (CRESST 2001). Also see Stecher and Hamilton (2002) for an analysis of how the best intentions and designs can go wrong.

either implicitly or explicitly. Similarly, LaMarca (in Erpenbach et al., 2002) set forth ten critical questions and their underlying major considerations that States should address in the process of developing their systems.[8] See Appendix C for a listing of those questions and underlying major considerations. For the discussion in this paper, all the elements and components are grouped together into six areas, as shown in Figure 3.

FIGURE 3. THE COMPONENTS OF A STATE EDUCATIONAL ACCOUNTABILITY SYSTEM


State Accountability System
❶
Setting purposes and focus
❷
Selecting indicators
❸
Data collection, scoring, and analyses
❹
Drawing inferences and making decisions
❺
Implementing the decisions
❻
Evaluating the effects
The State's Educational and Instructional System
(Goals, curriculum, teachers, etc.)

1. Purposes and Goals. There are at least three key facets that need to be addressed in the statement of purposes and goals:

a. **The overall goals (of the accountability system, not the educational system):** What is expected to happen—what are the intended outcomes? To what degree is it improving student achievement, per se, versus other goals?

b. **The focus or target:** What kind of schools should be identified for improvement? What types of students (achievement-wise) are the focus, and what does success look like for these populations?

c. **The logic or theory:** What is the theory of action underlying the reform strategy? How are the levers of accountability to accomplish the goals?

[8] Both of these documents also deal at some length with validity issues related to different designs.

Overall Goals

Accountability systems actually have many purposes, some explicit and others implicit. The most obvious and explicit goal is that of improving student achievement. Other purposes, such as, increasing the support for public education, are also important—indeed, they may be the real reason for the existence of some accountability systems. Accountability systems can also serve to increase motivation to improve, to help identify areas most in need of improvement, or assist in State resource allocation decisions.

Some have suggested that the goals be structured according to the positive and the negative. On the positive side, the goals would be listed but set in the context of the primary ones and the secondary ones. The negative side would be a foreshadowing of the things that could go wrong; it would list the things that must be avoided and show evidence of a strategy or mechanism for detecting early signs and taking appropriate action.

This chapter takes the position that **standards** originate and function in the realm of the educational system (which as described later, is separate from the assessment system and the accountability system), and, therefore, are not dealt with in this document. Yet, standards are at the heart of both the assessment system and the accountability system. Alignment of the assessments to the content standards is the central issue for assessment systems. For accountability systems, the question is about the rigor or appropriateness of the performance standards as the basis for determining the percent of students meeting the proficient level, and therefore the level of learning in a school, and, hence, its possible need for improvement. This, in turn, will depend on the context in which the standards were set and the purposes that they were intended to meet.

The goal descriptions should include a statement of the **guiding principles**. This would include a description of the ways in which the system is to be developed and implemented, including the groups who would be involved in the design and evaluation of the system. It also would include a discussion of the things that the system would be specifically designed to avoid or at least minimize. The two best examples here are (a) the ways in which the system would be designed to detect—or at least not provide an incentive for—the inevitable and logical efforts to "game" the system or to obtain artificially positive results; and (b) to ensure that, in order to obtain positive results, teachers and administrators are not rewarded for engaging in behavior that would defeat the overall goals, including the narrowing of the standards, or the exclusion of certain types of students from assessment or from participating in various programs.

The need to keep the system honest is closely related to the **credibility of the system**. One could think of credibility both as a component of validity—and as an outgrowth of validity. In some ways it is the equivalent of "face validity" in assessment. The point is that without credibility, the impact of the system cannot be sustained. Although many things can lead to a lack of credibility, prime examples include the items mentioned above related to "gaming the system" or showing indefensible results. Illogical results also undermine credibility; stories are told of schools that receive rewards and sanctions simultaneously, or very high scoring schools that are identified for improvement. While there may be very rational explanations for such phenomena, the credibility of the system is stretched—at best. If a system's procedure for identifying low-performing schools lacks reliability so that schools bounce in and out of school improvement status, no amount of rationalizing is likely to help. Similarly, a system that identifies virtually all schools for improvement is likely to get an incredulous response—and, unfortunately, have only a trivial impact on the schools that are most in need of improvement.

The Focus of the Reforms

The second aspect of the statement of purposes is a declaration of the types of students and schools that are the object of attention or reform. This is the essence of the defining of the construct—including the definition of the schools that merit improvement status. This might seem like an unnecessary step—the focus is on schools where the students aren't learning and the test scores reflect it, what's to talk about? It turns out that this is only true if (1) all results were based on several years of data; (2) if there were no student mobility, and (3) if all schools drew from the same types of communities. Since these three conditions are hardly ever met, it does make a difference how one looks at the scores for a school; different—and equally rational—ways of looking at the results yield quite different results for many schools. Hanushek and Raymond (2002) confirm the need to think carefully about this topic; as they see it, this is just one of the ways that accountability systems fail to show any evidence of a rational design or plan.

One conception uses a simple two-by-two table to look at four main models or approaches (Carlson, 2001). Two other documents (Gong et al., 2001; Hill, 2001) expand on that conception by drawing out the underlying assumptions about the nature of schooling and school reform, showing how each would identify schools of different types and presenting what is known at this point about their technical characteristics. Figure 4 presents an abbreviated description of the three most popular models.

FIGURE 4. THREE COMMON WAYS OF REPRESENTING PROGRESS OF SCHOOL ACHIEVEMENT RESULTS

School-level Model	Example of the Type of Data Used	Advantages	Disadvantages
A. Status	68% of the 4th graders meet the Proficiency level in reading in 2001.	Easy to understand; school scores are quite reliable.	Scores indicate the level of current achievement only.
B. Change, Successive Groups, or Improvement	Difference between percent meeting the Proficient level for two years for students at the same grade level (e.g., change of 4% between 2001 and 2002 for 3rd graders in Lincoln School).	Easy to compute.	Differences between cohorts are large and random, leading to wide variability and lack of reliability (Kane & Staiger, 2001, 2002; Linn & Haug, 2002).
C. Growth or Longitudinal[9]	Average progress from one grade to the next (or from fall to spring) for a set of students (e.g., average growth from 3rd to 4th grade in Lincoln School is 33 scale score points).	Allows inferences about school effectiveness.	Requires assessments at adjacent grade levels; scores are also quite unstable from year to year; and some believe that this method requires vertically-scaled assessments.

All accountability systems have to define the construct by answering the question, "What kinds of schools should be identified for improvement?" It is an obvious exaggeration, but one might say that good schools are very similar, but bad schools are bad in different ways. This is why it is important for the accountability system to clearly communicate the definition of a school identified for improvement. There are two main philosophical positions, both of which are rational and defensible, but lead to the identification of different schools:

[9] The distinction is usually made between true-longitudinal (where the same students are tracked over a year, or more), and quasi-longitudinal (where a whole group of students is tracked over time, but not individuals are present in all calculations, e.g., where the third grade for a school in one year is compared with the fourth grade for the next year).

- **Ineffective schools** should be identified for improvement—because they have weak programs.
- **Low-scoring schools** should be identified for improvement—because they have a large number or proportion of low-scoring students.[10]

Schools with effective programs (those that do well under Model C analyses—growth or longitudinal) are able to bring students to high levels of achievement over the years—as judged by the increasing achievement of individual students (based on individual growth curves). If there is high student mobility, however, this effectiveness may be masked and difficult to detect—leading to the false inference, if Model A or B is used—that they are ineffective. Those who argue that ineffective schools ought to be the focus of reform efforts would prefer to identify the weak schools based on the lack of growth for the students who have had an opportunity to benefit from the schools program (the students who have been there at least a year, for example). It has even been suggested that under the NCLB Act, a State should use the status model as required, then use the longitudinal results for the schools as a second filter. This way, the lowest scoring schools, which are also the least effective —would receive the most or the earliest attention, and the others—the schools with low-scores but apparently effective programs—would be in a second tier.

Others argue that the purpose of an accountability system is to identify schools with large proportions of low-scoring students; the goal is to help low-scoring students and the best way to find and serve them is to identify low-performing schools. Two assumptions are involved here:

- First, it is assumed that these schools have ineffective programs that need to be overhauled. This might not be the case. Some research has shown that a significant portion of those schools actually have very effective programs, but that large in-migrations of low-scoring students make their programs appear to be ineffective. Some would say that resources spent on these schools would be better directed to truly ineffective schools.
- Second, it is also assumed that these schools house most of the low-scoring students in the State. All schools, even high scoring schools, have low-performing students. This may seem like an obvious and uninformative statement. However, the actual number or proportion of such students that are not in the "worst" schools might be considerably higher than expected. It depends on the homogeneity of schools within a state—and states differ considerably in this regard. Table 3 illustrates this phenomenon for two States. While the States obviously are very different, it still is clear that if the school reform efforts were to focus on only the lowest 20 percent of the schools, as many as 70% of the low-performing students would be still be untouched, at least in the short run (causing some to say that the NCLB Act actually means Leave No School Behind).

TABLE 3. THE DISPERSION OF LOW-SCORING STUDENTS ACROSS SCHOOLS FOR TWO STATES

State	Percent of All Low-Performing Students in the State That Are Found in the…		
	lowest 20% of the schools	next 30% of the schools	highest 50% of the schools
A	30	30	40
B	40	30	30

Note: The figures in this table have been rounded to the nearest ten percent.

[10] Model C would be the method used to identify ineffective schools; Model A would identify low-performing schools. Model B is not discussed here, although it is the traditional method that has been the basis for AYP definitions, in spite of its lack of technical merit.

An extrapolation of this line of reasoning is that the current concern about being sure that all schools are identified, especially the small schools, may be misplaced, or at least over-emphasized. It could be shown that the total number of low-performing students in all small schools is less than the number of low-performing students in larger schools that would not be identified for improvement (using the status model). Advocates of the status model would need to articulate their focus, whether it is on low-performing/ineffective schools, or just on low-performing students or schools—and if the latter, to be sure that the procedures accomplish that goal.

The risk is great of generating confusion in the reader's mind, where none existed, from such a short review of this complex topic. Nevertheless, it should be evident that the topic is important, and that any judgment of the validity of the system demands clarity of focus. It is the heart of the process of defining the construct.

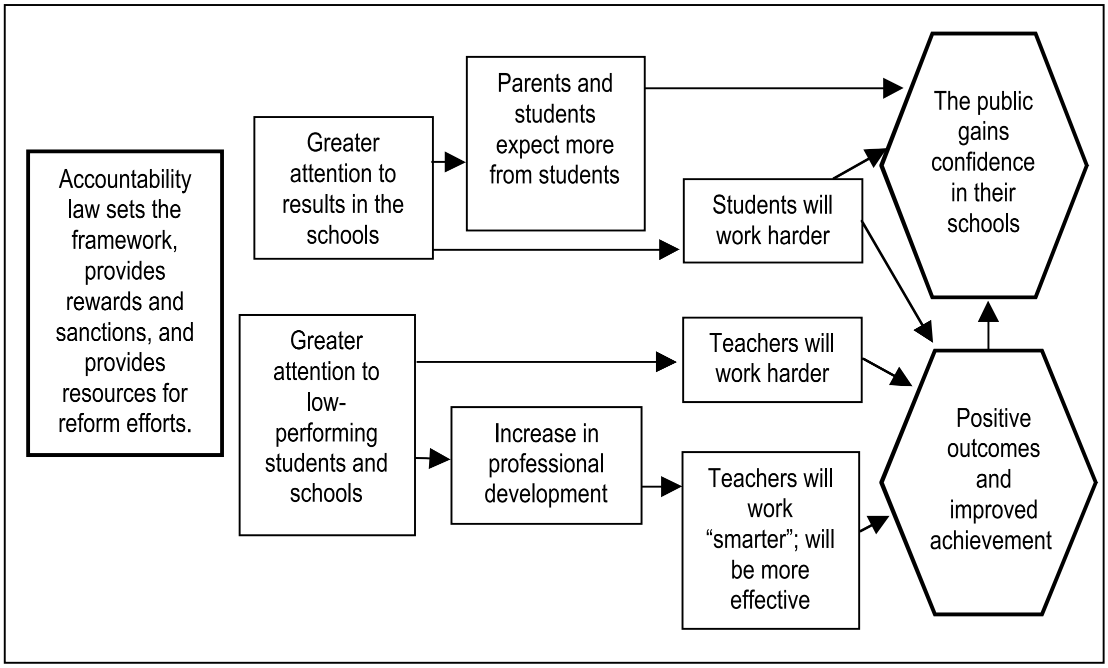
A Theory of Action

It is critical that the statement of purpose for the accountability system spell out its underlying theory of action. There are many questions that need to be considered:

- How are schools expected to improve?
- What factors will mediate this improvement?
- Will public pressure be sufficient, or is it meant to trigger other changes that will lead to improvement?
- How will this attention and pressure be modulated to stimulate teachers and administrators to make positive changes without stimulating them to take shortcuts or to "game the system"?
- What are the mechanisms that ensure proper attention to the goals of learning—especially those that are **not** part of the reward and sanctions system?
- How are the various types of incentives designed to work?
- How are teachers to learn how to "work smarter"?

Without this explicit theory of action—of action and reaction, which unwrap the system designers' assumptions about "how things work," it is virtually impossible to judge the validity of a system or to tell why it is or isn't effective (Weiss & Brickmayer, 2000). It is difficult to determine if the components are aligned and supportive of the system as a whole, or to detect where the system appears to break down. See Figure 5 for a graphic illustration of an over-simplified version of the elements and how they relate to each other. Visual displays allow all the assumptions and steps to be considered simultaneously, so the missing ones might be spotted. More importantly, the very process of agreeing on the theory will encourage essential discussion among the policymakers and others responsible for the design of the accountability system.

FIGURE 5. A SIMPLIFIED THEORY OF ACTION: HOW ACCOUNTABILITY EXPECTATIONS AND ACTIONS WILL LEAD TO HIGHER ACHIEVEMENT


Accountability law sets the framework, provides rewards and sanctions, and provides resources for reform efforts.
Greater attention to results in the schools
Parents and students expect more from students
Students will work harder
The public gains confidence in their schools
Greater attention to low-performing students and schools
Teachers will work harder
Increase in professional development
Teachers will work "smarter"; will be more effective
Positive outcomes and improved achievement

2. Indicator Selection. This refers to the all the processes required for data collection related to the academic content areas and grade levels; determining the nature and format of the assessment instruments; and the types of non-cognitive instruments and data. The indicator selection process is guided primarily by the purposes and goals of the system, but must also cope with the realities of obtaining data that can meet the exceedingly heavy demands of an accountability arena, such as the need to have rotating secure forms of the assessment instruments and heavy monitoring of other data collection processes.

While assessment is the key indicator for most accountability systems, it is seen here as a separate sub-system that provides the primary information for the accountability system. Figure 6 portrays this relationship between a State's assessment system and its accountability system. Other documents, including the *Peer Reviewer Guidance for Evaluating Evidence of Final Assessments* (U. S. Department of Education, 1999), describe the meaning of validity and reliability for assessment systems. This is not to say that those responsible for accountability systems are not obligated to ensure that all the aspects of validity of assessment systems have been reviewed and evaluated but it does assume that the processes of documenting the validity and reliability of the assessments and the assessment system is a relatively separate—and non-trivial—process. It would include reviewing the evidence for at least the following:

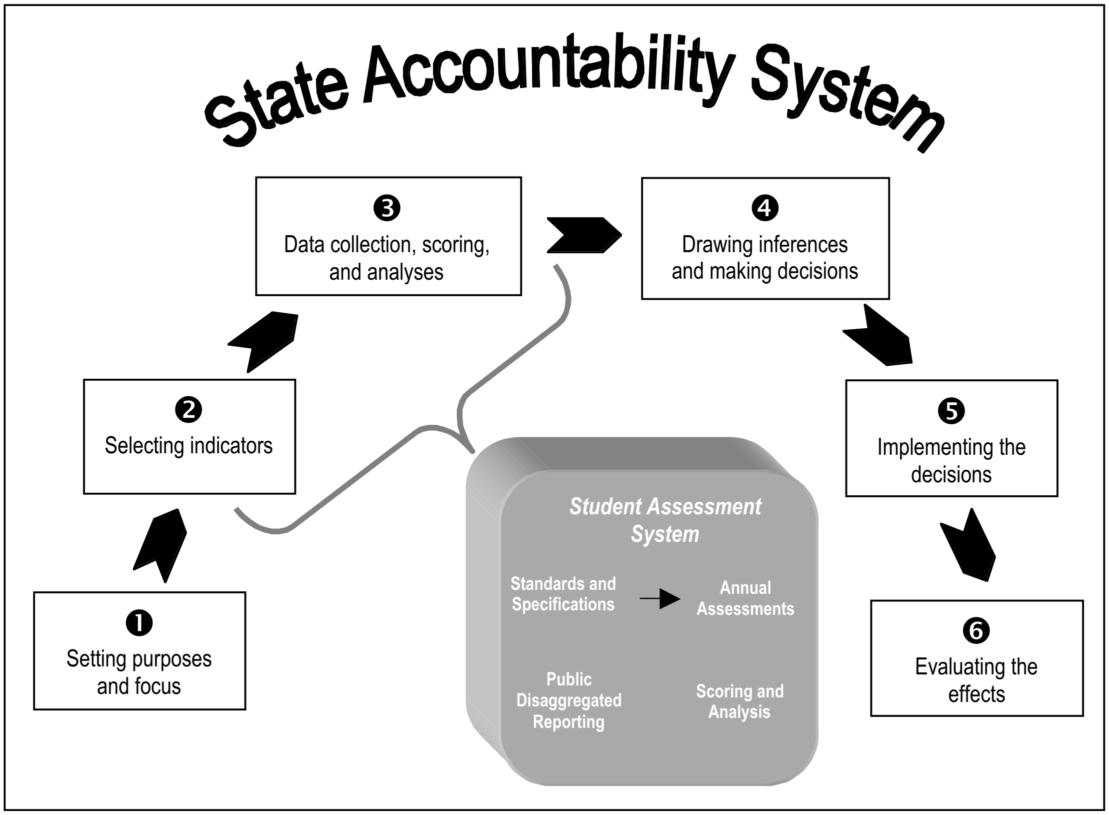
- The depth, breadth, and quality of the state content standards
- The quality of the procedures for setting performance standards, including the breadth of involvement of the community and major stakeholder groups[11]

[11] The content and performance standards are not discussed at any length in this document since they are seen as part of the educational system itself, rather than the accountability system. Suffice it to say that they do have a major bearing on the way that schools are judged in an accountability system and may have implications for the design of the accountability system itself.

- The alignment between the standards and the assessments
- The quality of the assessments and the coherence of the whole assessment system in providing different types of information for different purposes
- The degree to which the performance standards were set to serve as motivating goals, but not really reachable, versus actual criteria that could be reached by virtually all students who had access to a high-quality instructional program

Furthermore, it is incumbent on those responsible for accountability systems to be sure that the validity and reliability of assessment systems can be maintained and defended under different accountability designs.

FIGURE 6: THE PART-WHOLE RELATIONSHIP BETWEEN A STATE'S ASSESSMENT SYSTEM AND ITS ACCOUNTABILITY SYSTEM


State Accountability System
❶ Setting purposes and focus
❷ Selecting indicators
❸ Data collection, scoring, and analyses
❹ Drawing inferences and making decisions
❺ Implementing the decisions
❻ Evaluating the effects
Student Assessment System
Standards and Specifications
Annual Assessments
Scoring and Analysis
Public Disaggregated Reporting

3. Data Collection, Scoring. Most of the tasks in this component are related to the assessment system itself, but this section also includes the computations and derivations for the other indicators. All of the issues related to data quality, accuracy, and bias come to the fore in this stage. Although assessment systems are treated here as separate subsystems, it is clear that the reliability and validity of an accountability system must rest upon that of the underlying assessment system.[12] It is important to be sure that the assessment system includes all students, that the assessments are scored consistently; and that procedures provide for comparability across schools and across years. For non-achievement indicators, it has been found

[12] In contrast, Richard Hill has written about the ways in which the reliability of an accountability system is relatively independent from that of the assessments (see Hill, 2000).

that definitions of variables and procedures for collecting and editing that had seemed adequate for routine management and administrative purposes are usually found to be inadequate to ensure the integrity of the results for accountability purposes.

4. Developing the Model: Designing the Rules for Making Inferences and Decisions about School Success.

This may be the most important component; it surely is one of the most powerful determinants of the ratings and classifications of the schools. The model and its procedures follow philosophically from the statement of purposes and goals of the system, and are responsible for implementing the vision for the system, including the particular focus of the system in driving a particular kind of change for particular types of schools and students. It dictates how the indicators are to be weighted and used to make decisions about different types of schools, how different types of measurement and sampling error are to be dealt with, and how the integrity of the data and the process is to be ensured in the face of the natural forces that mitigate them. It includes such specifics as the relative emphasis on status or change; the relative focus on different types of students; the processes of combining the results to form a judgment, including compensatory and conjunctive approaches; the metric to be used, how small schools and subgroups are handled, and the combining of data across years and grades (and any other topics discussed in Chapter 3).

5. Implementing the Decisions.

In this phase, the decisions are being executed. Schools are being classified and specified reform strategies are being implemented. The key issues here revolve around fairness and fidelity of the process. It includes the strategies for tailoring the reforms to the nature of the schools and their student populations, and perhaps, at least for some states, to the degree to which the schools failed to meet adequate yearly progress (e.g., the number of subgroups that failed to meet AYP).

6. Evaluating the Effects.

The accountability system has its own evaluation component. This is different from the evaluative function that the system fulfills for the educational system; this component looks specifically at the impact of the accountability system itself. Is it having the right kind of effects? Is it being implemented properly? Which components are problematic? Where does it seem to be working best? An approach to conducting this evaluation is presented in the next section.

Part 3. A definition of validity for Accountability Systems

There is almost no literature on the validity of accountability systems.[13] Therefore, in this section, the definitions of validity as used in the worlds of science, business, medicine, and psychological and educational testing are applied to accountability systems. The following definition would seem to be a proper application:

An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that

- ***The components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and***

[13] The thought-provoking set of proposed standards for accountability systems developed by CRESST is one of the very few resources (2001).

- ***The system is accomplishing what was intended (and did not accomplish what was not intended).***

This definition accounts for all three of the critical requirements for the validity of any system:

1. The smooth interaction among the parts of the system, as they are aligned with the system's purposes
2. The outcome or effectiveness of the system in meeting those purposes
3. The breadth and strength of the evidence, and therefore the inference, about the functioning of the system (evidence about both 1 and 2)

The definition also implies that a system can be invalid either because it (a) lacks the evidence to support an inference about its effectiveness, or (b) that the evidence shows that the system is failing to accomplish its goals.[14]

The dual aspects of validity—construct and consequential—that are discussed in assessment can be seen in this definition for accountability systems as well.

- **Construct validity.** Does the accountability system focus on the right (agreed upon?) aspects of schooling? Is the concept of a "good school" or "school in crisis" thoughtfully set forth? Is that really the focus of the system, or are there mismatches in the design or implementation stages?
- **Consequential aspects of validity.** Is the accountability system working? Do the results support the intended impact of the system? Are there other unintended consequences?[15]

Even if satisfactory, this definition is still only a definition; it does not tell how that judgment is to be reached or what kinds of evidence are necessary. It also does not say how to diagnose the system to see why it isn't working. Some strategies to meet these needs are included in the next section.

Applying the Definition to Accountability Systems: An Illustrative Evidence-Based Template

This section looks at the process of evaluating or judging a system. It includes looking at the ways that a system might be invalid, specifically focusing on the main ways that it can go wrong. Popper has taught us that **an explanation (or a judgment about validity in this case) gains credibility chiefly from falsification attempts that fail.**

As seen in the "hard" sciences, judging the validity of a conclusion or a theory is a process of forming a "working" judgment based on a systematic look at a variety of evidence of different types. Judging the validity of an accountability system is, unfortunately, no less direct. This breadth of information, based on multiple measures, must pertain both to the construct side of the ledger (Is the system focusing on the "right" goals?) and the consequences side (Is the system having the desired impact?). It is interesting to note that the concepts of validity and reliability are most distinctly different in this regard. One can actually compute various types of statistical indices of reliability. One can study both the actual stability of scores and indicators over time or one can make estimates based on simulations (see Hill, 2001 and 2002). There is, unfortunately,

[14] It is important to be alert to the distinction between a lack of evidence and sufficient evidence of a lack of effectiveness.

[15] The reader may also agree that the two main aspects of the dictionary definitions are reflected above. The "strong" and "well-grounded" elements of the definition seem analogous to construct validity (i.e., having strong evidence that the system has the right focus). Similarly, the "effectiveness" and "efficacy" facets of the term seem to relate to the consequential aspects of validity (i.e., Is the program working?).

no satisfactory way to compute a coefficient of validity, especially of the validity of the overall system. Since there is no accepted approach, the following steps are offered for the reader's consideration[16]:

- Check for intended—and unintended—outcomes.
- Search for corroborating evidence.
- Evaluate each component.
- Study the implementation of the reforms.
- Study the levels of impact.

Each of these is briefly discussed below:

1. **Check intended and unintended outcomes.** The process would almost certainly begin with a close look at the intended outcomes, beginning with the assessment results. In most cases, this is a matter of seeing if there is a reduction of low-performing students and schools—as called for in the AYP definition.

   The search for unintended outcomes is trickier, partially because no provision was made for the collection of data—since these outcomes were unintended. The other outcomes may be positive or negative, but it seems that the negative ones gain most attention. One way to focus the search is to think of the impact of the accountability system on the different audiences or participants. Beginning with the students, one of the most obvious things to check for would be the opportunity cost, the loss of opportunity to learn other content areas, or even the broader concepts and skills within the subject areas of focus. The proposed standards for accountability systems (Baker et al., 2002) list some possible "side-effects" on the morale and retention of teachers and principals. The impact on the public has not been studied carefully. The goal of informing the public about the overall quality of their schools is frequently mentioned and laudable, but with the difficulty of reporting of complex assessment information in the popular press, one wonders if the public is developing a full and accurate understanding of the relative quality of the schools and their rate of improvement or decline.

2. **Search for corroborating evidence.** One would then conduct a disciplined search for additional information to corroborate the "official" overall findings. If the goal is to establish the validity of the system, the usual difficult questions are asked:

   - How do we know that the right schools were selected for improvement?
   - How do we know that the reforms were the right ones?
   - How do we know if the reforms were truly effective?

   For each of these questions, the only recourse is to look for corroborating evidence, or for evidence that contradicts what seems to be the logical conclusion—as well as evidence of unintended consequences.

   It is important to note that this process is independent of the level of success or validity of the system as indicated by the main findings. It is just as important to see if successful findings are truly trustworthy as it is to see if negative findings should be taken seriously.

[16] While the process of evaluating the effects of the accountability system (the evaluation component--#6, as discussed in a previous section), is not identical to the process of studying the different sources of information about the accountability system in order to judge the validity of the system and to evaluate the strength of the inferences underlying that judgment, these two processes will be discussed in common.

The natural tendency is to accept positive findings as valid. It could be argued that the temporary (spurious?) gains of most programs are often larger than the "real" gains (Koretz, 1996; Koretz et al., 2001). It is incumbent on the evaluator not only to try to separate the temporary from the long-term gains, but also to show the opportunity cost of those gains. If, for example, social studies or science is neglected to obtain greater reading achievement, the public needs solid information on the magnitude of the trade-offs.

Three kinds of other external evidence might be useful:

- **Other outcome measures.** It might be that some available assessment data are not included in the accountability system as primary outcome measures. Even if it may be less than desirable in terms of alignment with the standards, it might be useful adjunct information.
- **Process measures.** This could include ancillary achievement-related information, such as measures of the quantity or nature of writing assignments. These would be of interest in their own right, and they could be used in a triangulation process to judge the validity of the official scores. For example, if the quantity of writing assigned had not increased statewide, should one wonder about a major improvement in the scores? Baker and others have argued persuasively for monitoring the processes using these and other instructionally-focused "Early Indicators" and "First Wave Outcomes" as evidence of reform taking hold (Baker, 1999). Finally, any studies of the degree to which clusters of teachers focus on teaching the standards—as well as how their students rate this alignment between standards and instruction—can also be very useful in interpreting the achievement findings.
- **Attitude and opinion information.** Information on parent or client satisfaction is often useful. Some systems may include measures of parent satisfaction as part of the official outcomes. For others, this would be considered external information that would be useful in judging the overall impact of the system. One often hears a phrase, such as, "But we really know which schools are bad, don't we?" While the truth of that assertion may be debatable, or at least inconsistent, it may be useful to probe the common judgments of those familiar with the school and its staff—it may be that what this information lacks in verifiability, it makes up for in insights that are otherwise unobtainable.

3. **Evaluate each component.** After conducting a review of external evidence about the impact of the overall system, it is important to study each of the components. Each component has to be defined and implemented in harmony with the functions of the other components in light of the system's purposes. Again, this should be done even if the system seems to be having the right impact. It is possible for the system to be showing positive results, but for the wrong reasons. For example, a school may be showing improvement, but one would want to be sure that it isn't a function of increased drop-out rates.

   While the goal is to arrive at an overall judgment of the validity of the system, both validity and reliability must be checked for each component—and usually for a number of sub-components. One should look for validity of the purposes statements and in the logical consistency among those statements, the validity of the indicators and their specific use in a given accountability system, the validity of the decision processes, etc. Similarly, one is obligated to look at the reliability or stability of the data underlying the indicators, the reliability of the combining and other analytic procedures, the reliability of the classification rules—across schools and years, and the more familiar reliability or consistency of the actual classifications.

Some of the key issues per component might include the following:

- **Purposes.** Are the purposes consistent with each other, and with the indicators and the decision rules? Are the guiding principles being followed? Are there plans for evaluating the credibility of any gains?
- **Indicators.** Do the indicators match the purposes? Are there gaps? Are they biased against certain types of schools?
- **Data quality.** What evidence is offered for the accuracy of the data?
- **Decision rules.** Are the rules spelled out clearly? Are they compatible with the purposes, and especially with the definition of schools worthy of identification for improvement? What proportion of the schools is identified? Is that a credible number? What proportion of the schools is probably inaccurately classified?
- **Reform implementation.** Is there clear evidence that the program improvement strategies and the sanctions are being implemented? Are the neediest schools receiving priority assistance?
- **Evaluation.** Is the system effective and are the main purposes being realized? What are the main obstacles to greater effectiveness? What improvements in the system are called for?

Not only should these components be examined separately, but their inter-component alignment needs to be studied as well. Two components may be defined and implemented quite thoroughly, but if they are in logical conflict, the system will lack validity. The fourth part of this chapter offers some illustrations of this problem.

4. **Study the implementation of the reforms.** The reviewer must look at the level and quality of the implementation of the reforms, from the actual classifications of the schools to the fidelity of the selected school reform efforts. It is important to determine (especially if the results are disappointing) if the theory of action is inadequate, if the data are incomplete—or if the programs aren't actually being implemented
5. **Study the impact and validity by unit or level of analysis.** As mentioned in the description of the sixth component of an accountability system, these analyses should be done on several levels—at least statewide and for particular types of schools, perhaps selected on the basis of size, geographic area and student population. Furthermore, although these five steps have focused on the validity of the post-accountability action, it is just as important that the questions in the pre-accountability action phase (see the first column in Figure 2**)** also be addressed.

The methods of judging the validity of a system will vary with the maturity of the system. In the early stages of implementation, mainline outcome information may not be available. Therefore, the focus must center on the clarity of the descriptions of the separate components, the completeness of the definitions, the coherence among the different components of the plans, and the feasibility of the plans. As systems are implemented, the focus can expand to the effects of the implementation efforts.

Degrees of Validity.

When is a system not valid? When any part of it is less than perfectly valid? If complete validity were required for every component, there probably could never be a valid system. The new testing standards state that "validation can be viewed as developing a scientifically sound argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA, 1999, p. 9). One might propose that a system is valid when the users and participants in the accountability policymaking arena agree that the data and the arguments linking them together support (a) the classifications and interpretations that the accountability system imposes on schools, and (b) the uses and the consequences of those inferences and the actions that were taken. This allows a variety of actions. In some cases the system may only need to be fine-tuned, some might be in need of major changes, and some may need to be abolished and re-constituted.

Part 4. Illustrating the Validation Process

The very definition of a system is that of a set of processes that work together in coherent fashion to accomplish some purpose. It assumes that each of the parts are present and functioning, both in an independent sense, and in relation to each other. In most systems, the output of one element or component serves as input for subsequent steps or components. In the old-fashioned, serially-wired Christmas tree light strings, if one bulb were defective, the whole string was dark. This analogy is not totally appropriate, however. In the case of accountability systems, an improperly defined component may not short out the whole system—or might not seem to; the system may appear to be working quite adequately. On closer examination, however, it usually can be seen that the final output or impact of the system is not what it could be—or not what it appears to be.

These illustrations take the path of sharpest contrast, illustrating validity from a negative perspective: In what ways can a system be invalid? How is it likely to go wrong? The sources of a system's lack of validity (or lack of reliability that leads to a lack of validity) are of **two general types**:

- The definition or implementation of any one stage or component is defective in some way, including unreliability.
- Two or more of the stages or components are in conflict or out of alignment, either conceptually or computationally.

These two types of problems, intra-component and inter-component, are illustrated below.

Intra-Component Problems

Three main classes of intra-component problems exist:

1. Errors of Various Kinds. The most frequently appearing errors are in the stages involving scoring, analysis, reporting, or in the application of decision rules, but the problems are by no means limited to these. Problems usually result from some type of human or technological error. The variety, creativity, and persistence of error are as impressive as the diversity found in the natural world. Variables get defined incorrectly, input and keyboarding errors occur, wrong data files get used, wrong scoring keys get used, wrong formulas get applied, and the list goes on and on.

Sometimes these errors affect only some schools or certain types of schools; sometimes all schools are affected. Sometimes errors occur only for a given year; sometimes the problem is buried in the complexity of a given routine and survives—unknown to the programmers, the schools, and the public for years. Frequently, these errors lead to lower reliability or stability of results.

Occasionally, however, the errors can artificially suppress the natural variability of results, and actually make the system appear more reliable than it is.

2. Conflicts between Design and Implementation.

It is crucial that the definitions and specifications for each component and subcomponent be inspected. Sometimes the personnel responsible for defining the data or procedures are not the same ones that implement the procedures, or more likely, some aspect of a definition or process was changed—a change that didn't get made in the other procedures. This could be as simple as a failure to include a given computational step for schools of a certain type, such as those which had authorization (perhaps because of their involvement in a pilot project of some type, for example) to report data differently for a given variable.

3. Improper Definition or Formulation of a Component.

The example that comes immediately to mind is that of the definition of a "drop-out." Different definitions mean very different things.

A less well-known problem of this type is illustrated below. With the seriousness of identifying schools for improvement, it is important that schools be accurately identified, and that process requires a proper conceptualization of a school as a place which educates not just the students who are enrolled a given year, but one that is responsible for—and must be judged on the basis of—how well it educates different groups of students over the years. The next paragraph explains this assertion.

One of the largest sources of unreliability in an accountability system pertains to **Student variability across years** or what is known as "student variance." This does not refer to the variability of students within a school, but to the variability of students from one class to the next, that is, from one year to the next. Ask any teacher about the "Good class-Bad class" phenomenon; classes vary greatly from year to year. This has great implications for the reliability of the results and decisions made by an accountability system. If this fact is not taken into account, the decisions about the quality of a school's program will seem—and may be—more random than real. A school will be judged as in need of improvement one year and judged as not in need of improvement (or even worthy of a reward) the very next year, with no change in the school's program. Linn and colleagues (2002) describe it this way:

> There seems to be little recognition that school-level results are often volatile from year to year because of differences in cohorts of students. . . .changes in scores for students tested at a given grade from one year to the next can be markedly unreliable. (p. 12)

There are some steps that can be taken to ameliorate the situation, although they are usually less than satisfying. These steps are addressed in other sections of this chapter.

One implication of this fact of life lies in the computation of proper standard errors. Some formulas (known as infinite) take this variability into account, and some (known as fixed or finite) do not. It is now clear that one should use the infinite formulas; they produce larger standard errors, thereby decreasing the chances of making false judgments. Cronbach and colleagues (1997) have clearly set forth the principle and thereby established our obligation in this process:

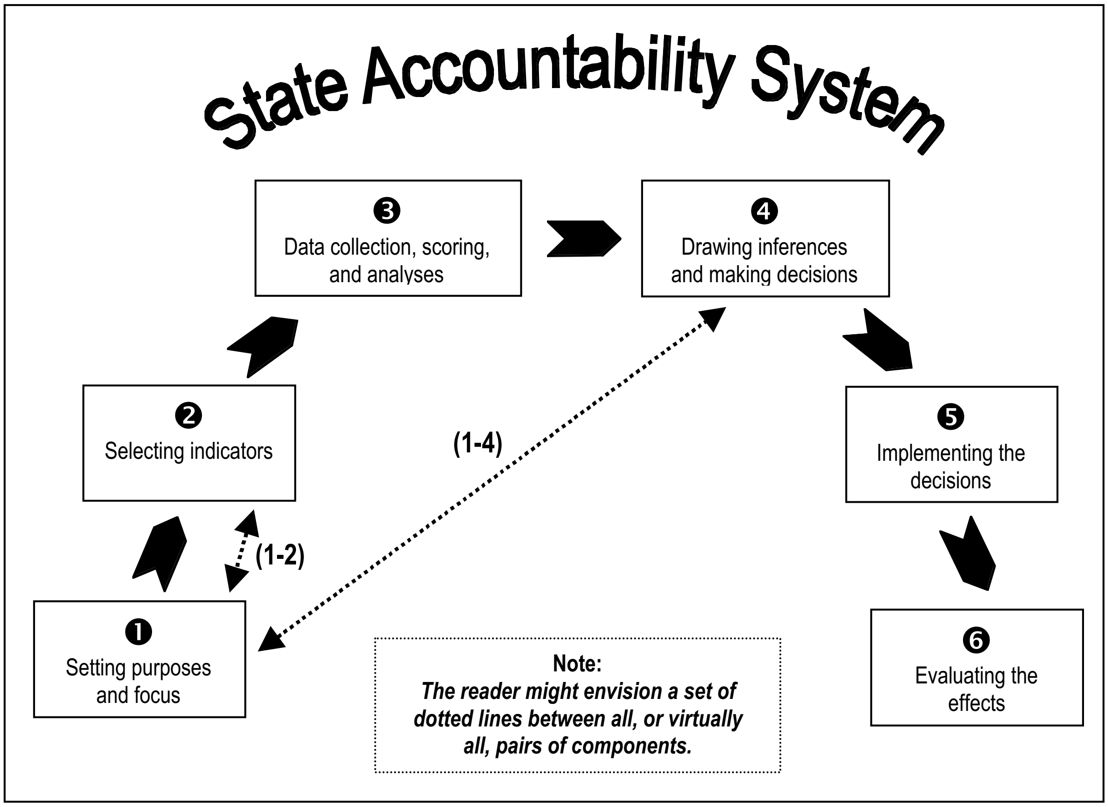
> Restricting inference to the historical statement would defeat the purpose of many school-accountability uses of assessment results, where inferences reach beyond students recently taught. Note, for example, that analysis of one year's data, developing the SE for the finite student body, cannot support such actions as rewarding a school for satisfactory performance, or imposing sanctions on a

> school that did poorly; the finite-population analysis provides no basis for assessing the uncertainty in the school mean that arises from random variation in student intake. Nor would the analysis support the inference that the program in school A is better than that in school B, even if they draw on equivalent populations. Nor, thirdly, does the finite-population SE provide a basis for arguing that a school mean higher in Year 2 than in Year 1 implies improved instruction, rather than a fortuitously superior intake of pupils. (p. 21)

Inter-Component Problems: Mal-Alignment among a System's Components

Two regions of inter-component conflict are illustrated here. Figure 7 illustrates where they occur. The first (1-2) is between the purposes (1) and the selection of indicators (2). The second (1-4) is between the purposes (1) and the decision rules for identifying schools for various kinds of intervention (4). Some of the more serious types of mismatches are described below for each of these pairs of components:

FIGURE 7. AN ILLUSTRATION OF THE REGIONS OF LIKELY MAL-ALIGNMENT BETWEEN ACCOUNTABILITY SYSTEM COMPONENTS



State Accountability System
❸ Data collection, scoring, and analyses
❹ Drawing inferences and making decisions
❷ Selecting indicators
(1-4)
❺ Implementing the decisions
(1-2)
❶ Setting purposes and focus
❻ Evaluating the effects
Note:
The reader might envision a set of dotted lines between all, or virtually all, pairs of components.

Example (1-2) Mal-alignment between purposes (1) and indicator selection (2). Nearly all States have standards delineated for all the academic content areas, yet their accountability systems only include indicators for certain areas that are judged to be crucial. This is certainly intentional, often justified, that some areas, like reading, are much more important than other areas. Other States restrict the focus to certain areas within a content area, often focusing on the more rudimentary skills. (Or, is this a case of mal-alignment between the goals of the educational program and the purposes of the accountability system?) States need to be aware that any restricted focus is almost certain to bring about a loss of instruction in other content areas. Research (Stecher, 2002) shows that teachers logically spend more time in areas that are under the accountability spotlight and that time is taken away from other areas, such as science. If this is spelled out in the purposes for the accountability system as an acceptable outcome, then the system is valid. If not, the technical definition of validity is violated; the system is not measuring or focusing on the purposes that it purports to.

Example (1-4) Mal-alignment between purposes (1) and the rules for interpreting the results and making school classification decisions (4). The opportunities for mismatches in this realm are legion. Deming (1995) described a system as a coherent whole that must be understood as one. Accountability systems also need to be dealt with as a whole, meaning that the parts must be in alignment. Coherence is king. What is the overall model or design? The model should be selected on the basis of the purposes of the system, and the views that it represents about the nature of schools and the best ways to help them improve. The model itself will usually dictate the rules for interpretation and decision-making. An earlier section reviewed the main accountability models (in terms of the type of focus on achievement—status, longitudinal, etc.). A lack of clarity about the actual functioning of the model and how it relates to the purpose of the accountability system can lead to alignment problems.

The following exhibit illustrates some of the problems that can emerge when the decision rules (component #4) are not compatible with the purposes of the system (component #1). It demonstrates the importance of specifying the purposes of the system as fully as possible—before selecting a design. The ways in which schools are classified is primarily a policy matter, not a technical one. As such, the classification process must be dealt with in the purpose section, as an outgrowth of the designers' beliefs and principles.

EXHIBIT 1. AN ANALYSIS OF SOME LIKELY SOURCES OF INVALIDITY DUE TO MAL-ALIGNMENT BETWEEN PURPOSES OF AN ACCOUNTABILITY SYSTEM AND ITS DESIGN AND DECISION RULES

Intent of the accountability model as stated or implied in its purpose section (Component 1)	Some possible decision rules (Component 4)	Nature and likely effects of a given mismatch between goals and decision rules
1. A stated goal of the accountability system is to provide accurate and meaningful information to the public about the quality of the schools.	A model identifies schools for improvement if any subgroup falls below a very rigorous cutoff.	Virtually all schools will be identified for improvement. The public will have great difficulty differentiating between a crisis of monumental proportions and the natural result of an instant application of extremely high standards to existing schools. At first, they will be shocked, and then realizing that the schools haven't suddenly stopped teaching completely, will lose faith in the state's accountability system.
2. A goal of the system is to identify schools with weak instructional programs.	A model calls for schools to be identified if they have a certain proportion of low performing students.	The likelihood of identifying schools of low effectiveness is close to random. Low-scoring schools may or may not be ineffective schools, depending mainly on the level and nature of the mobility of their student population. Schools with low-performing students can be very effective; however, it may not show if they have a large influx of low-scoring students.
3. A goal is to identify schools that truly are low-scoring—defined in this case as schools with consistently large populations of low-scoring students. This implies that schools that are not truly low-performing would not be identified.	A model identifies schools for improvement if any of their subgroups fall below the required percent of students scoring at the proficient level.	Many schools with very high scores for their general population but will be identified for improvement because of a small number of students with disabilities, for example. Should the whole school program be revised?
4. Schools are to be identified on the basis of their true (valid and reliable) weak areas.	The model identifies schools if one or more student subgroups (any of the subgroups) fail to meet the requirement.	Some schools will be identified because they were low in mathematics one year, then low in reading the next. How is the program to be strengthened? Some schools will be identified because their African-American subgroup was low one year, and the LEP group was low the next. Again, how is the program to be revised?
5. The goal is to have an efficient accountability system (i.e., one that serves the most schools possible, serves the most needy ones first, and gives the most help to the schools that need it most).	A model identifies schools for improvement if any subgroup falls below a cutoff.	A very large number of schools will be identified in most states. There is no clear method of offering triage to some. Many schools with very high scores for their general population but will be identified for improvement because of a small number of students with disabilities, for example.
6. The goal is to reward (i.e., to recognize schools that are making progress).	The model uses a single (high) achievement cutoff—and that cutoff is being raised each year.	Low-scoring schools that are making very strong progress will be identified for improvement if their level of achievement is just below the cutoff each year.

Intent of the accountability model as stated or implied in its purpose section (Component 1)	Some possible decision rules (Component 4)	Nature and likely effects of a given mismatch between goals and decision rules
7. The stated goal is to encourage teachers to work with the lowest-scoring students.	The model uses a single (high) achievement cutoff.	Since only the students near the cutoff have the potential to improve enough to raise the percent of students meeting the cutoff, teachers have a disincentive to work with the very lowest-performing students. Savvy teachers will work with the "bubble" students (i.e., those near the bar).
8. The goal is to encourage all students to do well in school and to graduate.	A model gives substantial weight to non-academic indicators.	Some schools could increase their achievement results by "encouraging" low-scoring students to leave. (If not weighted properly, the payoff from higher scores could offset the loss from a higher dropout rate.)

Summary

This chapter has outlined one perspective of validity and how it applies to State educational accountability systems. This is a new and evolving field, and it exists in each State and nationally in the most visible and controversial corner of the public policy arena—improving the quality of public schools. It is hoped this discussion will facilitate positive communication among various publics, the educational policymakers, and public school practitioners.

Although checklists lead to gross-oversimplifications, this chapter ends with the offering of a beginning checklist as a quick summary of the things to look for when examining the validity of an educational accountability system. It must be remembered that the answers to the questions on the checklist are really only useful if they help answer the following questions that define validity for an accountability system:

- Is the accountability system properly focused on its stated purposes; do all the components support that goal?
- Is the accountability system effective; are its goals being realized, and do they dominate over any unintended consequences?
- How good are the answers, that is, how strongly can the data support the inferences to be made about the program?

A “Checklist” Related to Examining the Validity of an Educational Accountability System

- □ What are the **goals and purposes** of the system?
  - What are the main expectations and outcomes? Are they explicit?
  - What type of students and schools are targeted? (What is the construct?)
  - What is the theory of action underlying the system? Is it consistent with the goals and the types of schools targeted?
- □ What are the **main indicators** that are used in the system? Are they completely aligned to the goals and purposes of the system?
- □ How are the **data** collected, scored, and analyzed? What is the quality of the data? How is reliability confirmed?
- □ What is the **decision model**; how are schools classified into the various quality classifications? How are the indicators “combined”? What is the reliability/consistency of the classifications? Is there an explicit value statement about false positives and false negatives, and how is that implemented?
- □ What are the main school **reform actions** dictated by the system? Are they being carried out faithfully?
- □ Is the **system evaluated** at least annually? How effective is it? What steps are taken to determine and improve the validity of the system? Are the key stakeholders involved in this process? What rules are used to combine all the findings and draw defensible inferences? How uniformly are these conclusions held by the various participants in a state—the teachers, the parents, the policymakers?

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author. [Copies available at AERA, 1230 17th St., NW, Washington, DC 20036].

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 201-238.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: Author.

Baker, E. L. (1999). *Future directions for program evaluation*. A presentation at the annual conference of the American Educational Research Association. [On-line]. Available: http://www.cse.ucla.edu/CRESST/aeraoh99/bakereval/sld006.htm

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* (Policy Brief #5). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Bandow, D. (1999, May 10). *The decade of "Junk Science."* [On-line]. Available: http://www.cato.org/dailys/05-10-99.html

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.

An adaptation of this paper, *Focusing state educational accountability systems: Four methods of judging school quality and progress,* available on-line at: www.aceeonline.org/presentations/Carlson_models.pdf

Carmines, E. G., & Zellar, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Carnap, R. (1966). *Philosophical foundations of physics: An introduction to the philosophy of science.* New York: Basic Books.

Clark, M. E. & Gironda, R. J. (2000). Concurrent validity of the National Pain Data Bank: Preliminary results. *American Journal of Pain Management.* [On-line]. Available: http://www.aapainmanage.org/default.asp

Clinical epidemiology & evidence-based medicine glossary: Clinical study design and methods terminology. Updated August 22, 1999. [On-line]. Available: http://www.vetmed.wsu.edu/courses-jmgay

Coleman, A. M. (2001). *Oxford dictionary of psychology.* Oxford: Oxford University Publishers.

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

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

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.

Definition of evidence-based medicine. (2001). *NCE based web tutorial.* [On-line]. Available: http://www.usc.edu/hsc/nml/lis/tutorials/ebm.html

Deming, E. (1995). Cited in J. Cortada & J. Woods (Eds.), *McGraw-Hill Encyclopedia of quality terms and concepts.* New York: McGraw-Hill.

Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). What does research say about assessment? [On-line]. Available: http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm

Drink at least 8 glasses of water a day—Really? Dartmouth professor finds no evidence for "8 x 8." (2002). [On-line]. Available: http://www.darmouth.edu/~news/

Proof and policy from medical research evidence. (2001). *Journal of Health Politics, Policy and Law, 26,* 2. [On-line]. Available: http://dukeupress.edu/jhppl/

Dynamic chiropractics. (2002).[On-line]. Available: http://www.chiroweb.com/contactus.html

The dynamics of reasoning in the sciences: Adaptive and interrogative perspectives. (2002). Belgium: Ghent University. [On-line]. Available: http://logica.rug.ac.be/centrum/

Faigman, D. L., Jaye, D. H., Saks, M. J., & Sanders, J. (2001). *Modern scientific evidence: The law and science of expert testimony.* [On-line]. Available: http://www.uchastings.edu

Francis, F.J. (1993). *Admissible scientific evidence in court* (Issue Paper 1). Amherst, MA: University of Massachusetts at Amherst, Department of Food Science.

Fuhrman, S., & Elmore, R. (in press). *Redesigning accountability systems.* New York: Teachers' College Press.

Gay, J. M. Home College of Veterinary Medicine. Pullman, WA: Washington State University, Field Disease Investigation Unit. [On-line]. Available: http://www.vetmed.wsu.edu/courses-jmgay/

The glossary of clinical epidemiology & evidence-based medicine. (2002). [On-line]. Available: http://www.vetmed.wsu.edu/

Gong, B. (2002). *Designing school accountability systems: Towards a framework and process.* Washington, DC: Council of Chief State School Officers.

Hanuskek, E. A., & Raymond, M. (2002, June). *Lessons and limitations of state accountability systems.* Paper presented at "Taking account of accountability: Assessing policy and politics," Harvard University, Cambridge, MA.

Hill, R. (1997, June). *Calculating and reducing errors associated with the evaluation of adequate yearly progress.* Paper presented at a meeting of the CCSSO Annual Large-Scale Assessment Conference, Colorado Springs, CO.

Hill, R. (2001). Issues related to the reliability of school accountability scores. In W. J. Erpenbach, et al., *Incorporating multiple measures of student performance into state accountability systems—A compendium of resources.* Washington, DC: Council of Chief State School Officers.

Hill, R. (2002, April). *Examining the reliability of accountability systems.* Paper presented at the 2002 Annual Conference of the American Educational Research Association, New Orleans, LA.

Hurlock, J. (2002). *Validity and accuracy of ratings.* [On-line]. Available: http://www.missouripolygraph.com/index.htm

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, T. J., & Staiger, D. O. (2001). *Volatility in school test scores: Implications for test-based accountability systems.* [On-line]. Available: http://www.brook.edu/gs/brown/brown_hp.htm

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). *Randomly accountable*. [On-line]. Available: http://www.educationnext.org/20021/56.html

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171-195). Washington, DC: National Research Council.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Lane, S., Parke, C., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17* (2), 24-28.

Linn, R. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Technical Report 539). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (June 2002). *Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001* (CSE Technical Report 567). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Linn, R.L., Baker, E. L., & Betebenner, D. W. (March-April 2002) Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher.*

Linn, R. L., & Haug, C. (2002). Stability of school building scores and gains. *Educational Evaluation and Policy Analysis, 24* (1), 27-36.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 89-103). Hillsdale, NJ: Erlbaum.

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

New Shorter Oxford English dictionary. (1993). Oxford: Clarendon Press.

Rogosa, D. (2002, September 9). What's the magnitude of false positives in GPA award programs? *Orange County Register* [Santa Ana, CA]

Savitha, S. (1999). *Why evidence-based oncology?* New York: Harcourt.

Schewe, P., Riordon, J., & Stein. (2002, January 2). A new limit on the overall validity of special relativity. *Physics News Update, 571*, 1. [On-line]. Available: http://www.aip.org/physnews/update

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (pp. 405-450). Washington, DC: American Educational Research Association.

Stecher, B. M., & Hamilton, L. S. (2002). Putting theory to the test: Systems of "educational accountability" should be held accountable. *Rand Review, 26*, 17–23.

U. S. Department of Education. (1999). *Peer reviewer guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act.* Washington, DC: Author.

Webster's third international dictionary (1977). New York: Merriam-Webster.

Weiss, C. H., & Brickmayer, J. D. (2000). Theory-based evaluation in practice: What do we learn? *Evaluation Review, 24*, 407-431.

Winter, G. (2000). A comparative discussion of the notion of "validity" in qualitative and quantitative research [58 paragraphs]. *The Qualitative Report 4,* 3/4. [On-line]. Available: http://www.nova.edu/ssss/QR/QR4-3/winter.html

Chapter

3

A Framework for Policy Decisions That Results in Statistically Reliable Disaggregated Information about Student Performance

The previous chapter described how the concepts of validity and reliability, familiar to those responsible for building large-scale assessment programs, could be applied to the creation and implementation of accountability systems required under the NCLB Act. By use of specific examples, we argued that both inter- and intra-component problems could threaten the validity (and reliability) of such systems. In this chapter, we discuss specific decisions policymakers need to consider that will affect the reliability and validity of the accountability system they choose to meet the NCLB Act requirements. Taken together, these chapters are designed to guide the development of reliable, valid, and *coherent* accountability systems, consistent with both the intent and spirit of NCLB, containing the highest possible technical qualities.

As with all comprehensive legislation, meeting the numerous requirements of the NCLB Act create various challenges for State policymakers. While many of the law's provisions appear relatively straightforward, implementing workable systems that meet both its letter and spirit has proven elusive for many States, particularly those with a history of reform based on clear, delineated principles.

No sections of the NCLB Act have created more uncertainty than those dealing with the calculation of adequate yearly progress (AYP). This is true for several reasons. First, many States were already in the process of developing and implementing their own version of AYP as a result of local reform movements and requirements under the 1994 ESEA Reauthorization. Determining similarities and differences between two complex systems (one very familiar and one somewhat hypothetical) has proven to be both time consuming and difficult, particularly given the delays in the development of regulations. Second, the "mechanics" of AYP truly are complex given the number of components and decision rules of which they are comprised—ranging from how to combine the results from various tests and grades into a school score, to how to properly disaggregate the numerous required subgroup scores, and to how to determine reasonable starting points and growth intervals at the school, local education agency (LEA) or school district, and State levels (and for each relevant subgroup of students).

Finally, AYP has proven difficult simply because it is so important. A properly developed system promises to separate successful schools from those in need of assistance. Shortcomings in the selected system will lead to misidentification of schools as in need of improvement or worthy of reward, diffusion of limited resources, confusion over what programs are working, and loss of public confidence in both our public schools and our attempts to hold them accountable.

Further complicating the process has been the evolving nature of the range of possible State AYP models. Initial draft regulations issued by the U. S. Department of Education (*Federal Register,* 2002, August 6) suggest AYP systems once thought to be unacceptable under the NCLB Act may be reviewed favorably. These draft regulations acknowledge "that there are rigorous models that States have already developed that may achieve the same fundamental principles of the statute, although through different approaches." States using such alternate models were invited to "comment on the statutory provisions that might affect their use, and how these requirements could be **incorporated into their current systems**" [emphasis added].

The ability of a State to continue to use an accountability model developed under the 1994 ESEA Reauthorization was re-affirmed by ED in promulgation of the final regulations on accountability (December 5, 2002) **provided that** the State can demonstrate how it has integrated the NCLB Act AYP provisions as required by statute and regulation (see Analysis of Accountability Issues for States in Chapter 1).

This chapter is intended to provide State leaders with a decision framework for developing systems for calculating AYP. In doing so, we will base our decision points on three primary considerations:

1. Language in the NCLB Act and subsequent negotiated rule making for regulations on standards and assessments (Federal Title I Regulations, 2002, July 5) draft regulations covering other parts of the law, and related correspondence from the DE;
2. Recognized professional practices for the development and implementation of high-stakes accountability systems; and
3. The apparent spirit of flexibility by the administration (Olson, 2002), as represented in the draft regulations on AYP (Federal Title I Regulations, 2002, August 6), and communications from the Office of the Secretary of Education (Paige, 2002).

As with various provisions of the NCLB Act, the range of possibilities for State AYP models may continue to evolve and is subject to final approval via the ED Peer Review process tentatively scheduled for early 2003. For example, the Peer Review process could place in operation some of the flexibility that the ED intended in Secretary Paige's "Dear Colleague" letter of July 24, 2002. At a minimum, States considering local variations of key NCLB Act provisions have the responsibility of providing evidence that such potential modifications do not violate the law's accountability expectations and will result in the success of schools being determined by the number of students meeting proficiency requirements across various content areas, especially reading or language arts and mathematics. ED officials may then choose to allow such modifications, based on the strength of the evidence provided.

Initial Factors

In building a State model for AYP, practitioners need a thorough understanding of the NCLB Act provisions and requirements, both as indicated in the Statute and as expanded on through the negotiated rule making on standards and assessments and draft regulations on other areas of the law. However, equally important is an awareness of several local factors that should be used to inform subsequent system building. Below we describe how these factors—*principles of reform, historical context,* and *practical considerations*—can help design an AYP system that is both consistent with the intent of the NCLB Act and right for a State.

Principles of Reform

Original review of the NCLB Act indicated that States would be required to implement a status growth model (i.e., one based on all schools meeting *identical* targets). Status models are based on two basic tenets:

1. minimum achievement levels can be determined below which students are at a clear disadvantage for subsequent success; and
2. the responsibility of the State is to ensure all students meet those minimum standards—performance beyond which may be desirable but not essential given limited time and resources.

Several States have implemented status accountability models (e.g., Texas), presumably based on the principles indicated above. Other States (e.g., Kentucky, California) have attempted to reward and sanction schools based on *growth* targets. Such models differ from their status counterparts in several important ways:

- progress towards mastery should be rewarded even if that level of performance may not yet meet minimum achievement levels (no floor exists); and
- all schools should be expected to improve, even those who may have exceeded minimum achievement levels (higher ceiling exists).

For the purpose of this paper, either the status or growth models are defensible given certain assumptions about education reform. States should examine their own local principles and (in the apparent spirit of flexibility) build an accountability model, including AYP parameters that match the principles driving the entire statewide reform movement.

Historical Context

Many States have been developing and implementing massive, comprehensive reform efforts for a decade or longer. Based on principles of reform, such movements usually require tremendous expenditures of political, social, and fiscal capital. The inclusive process undertaken typically involves representatives of the State's legislature, business and community groups, educators, parents, and State Educational Agency officials. The final product—the statewide accountability system—represents the consensus view of many of these constituents.

Provisions of the NCLB Act must be implemented with this historical context in mind. States should determine first which provisions of the law are identical or directly compatible with State objectives and provisions. Minor differences should be accommodated based on the NCLB Act requirements. For example, the law requires the goal that all students reach proficient or above in reading or language arts and mathematics within 12 years. If a current State accountability system projects 15 years for this same goal, the NCLB provision must take precedence. Other potential conflicts should be resolved under the umbrella of the NCLB Act requirements or intent. States must explain both why the local provision is essential to overall success **and** will lead to the same 12-year goal. For example, States that have developed accountability indices (rather than a percent proficient model, as suggested by the NCLB Act) must demonstrate that appropriate schools are identified for rewards and sanctions and no school escapes accountability *simply because of the* *accountability model proposed*. To the extent possible, intermediate goals should be determined to provide evidence that the final goal can still be reached via a State's mechanisms. This will serve to provide the necessary evidence that the State system meets the goals of and incorporates the spirit of the NCLB Act.

Practical Considerations

Key to the success of the NCLB Act is identifying failing schools and providing them sufficient assistance and resources to improve student achievement within the allotted time frames. States need to determine both the cost of improving schools at various achievement levels and the availability of resources (human and material) to assist failing schools. Various AYP decisions will result in identifying different subsets of schools. States should use current data and knowledge regarding successful and unsuccessful schools to help build an accountability system that identifies the "right" schools for rewards and sanctions and provides the correct resources to schools identified for improvement. Most States are already familiar with chronically under-performing schools; an important validity check on the proposed accountability system is whether such schools are identified as not meeting their AYP targets.

Careful attention to these initial considerations will result in an accountability system, including the calculation of AYP, meeting the intent of the NCLB Act and also being successful in the local context. States without a long or entrenched history of reform, or having no historical system to dismantle at a high political cost, may choose to move ahead by following each NCLB Act provision as explicitly set forth in the law. Even these States should examine their choices using the three primary considerations described above if they hope for implementation that minimizes resistance and maximizes long-term success.

AYP Decision Framework

The remainder of this chapter is designed to provide State leaders with a decision framework and a series of steps they can follow when developing systems for calculating AYP. We include discussions addressing such topics as

- *Nature of data sources,*
- *Number of starting points,*
- *Sample size issues,*
- *Aggregation issues, and*
- *Setting intermediate goals.*

Two factors drive the discussion for each issue: (1) adherence to the NCLB Act provisions and expectations; and (2) technical defensibility. It is important to note that at the time of this writing, the authors are working from assumptions of flexibility provided in Secretary of Education Paige's July 24, 2002 "Dear Colleague" letter. This flexibility may or may not be included in the final accountability regulations issued by ED. However, in this chapter, we will indicate what types of decisions are clearly supported by law or draft regulations and which decisions are based upon our understandings of the flexibility provided in the Secretary's letter.

Figure 8 presents a decision heuristic for State leaders to use as they design their AYP methodology. Some might be surprised to see the potential accountability approaches permitted under the NCLB Act divided into two major categories—improvement and increasing status (a common bar for schools in a given year). These are not distinct categories because any system needs to incorporate aspects of both approaches into a final design. The choice of improvement or increasing status approach is really a choice about order since starting from either approach, given the same decisions at other steps in the process, should lead to the same result. The initial decision to give priority to either an improvement or status approach can be based on many contextual factors such as political philosophy or beliefs about school improvement, but the nature of available assessment data will be a major factor in this decision.

Prior to discussing factors contributing to the choice of improvement or status models, it is worth discussing whether an improvement-based model would even be permitted under the law. The “safe harbor” provision [see section 1111(b)(2)(I)(ii)] clearly allows schools and districts [see § 200.20(b) of the final accountability regulations] to meet AYP if they reduce—for the subgroup that did not meet the status target—the percent non-proficient by 10%. Therefore, it appears to start with the improvement target? However, a school or district -using “safe harbor” to meet AYP for all 12 years is not likely to end up with 100% proficient by 2014. For example, a school starting with 0% proficient that barely makes “safe harbor” each year will result in only 71.8% by 2014. Although this outcome was clearly an unintended consequence of the law, States are not likely to be permitted to propose a model that leads to significantly fewer than 100% of **all** students reaching proficiency. Nevertheless, by adjusting the goal so that it reflects **all** students reaching proficiency in 2014, an improvement model can be used to meet both the letter and intent of the law.

Policy and Contextual Factors

Data Sources. There are several factors that will influence this initial choice in the decision framework. Figure 9 elaborates on the policy and contextual issues outlined in initially in Figure 8. For example, States that were fully compliant with the 1994 provisions of the Elementary and Secondary Education Act might only have one assessment at each grade level, which would make the use of a longitudinal system virtually impossible. Certainly, yearly assessment data allow for more options, especially data with unique student identifiers. The size of schools and student subgroups will also affect the decision. Small schools with data only three times through the K-12 span will have a difficult time using an improvement model. Mobility rates of students could also have a considerable impact on a State’s choice of approaches. Status approaches might better meet the intent of the law because they will tend to include more students than longitudinal methods that require matching of students across years. For example, if States had average mobility rates of 30% (the percentage of students who have been in their schools for less than one full year—from spring testing in year 1 to spring testing in year 2), the initial accountability decision would clearly be based on a non-representative sample of 70% of the students. States with this type of contextual issue would certainly need to use a status-type approach to supplement these initial findings and would probably need to define “a full academic year” as something shorter than a full calendar year.

Units of Analyses. The language in the statute has left many people confused about the appropriate unit of analyses for AYP decisions. Some have suggested that because there will eventually be testing in every grade, 3-8, States could actually determine whether each grade made AYP. While this is certainly possible, it does not make any sense when thinking about the inferences required under the law. The law requires States and LEAs to identify SCHOOLS for improvement, NOT grades for improvement; therefore, it only makes sense to consider schools as the appropriate unit of analysis.

However, defining a school is not always as straightforward as one might think. For example, if the State defines starting points and intermediate goals separately for elementary, middle, and high schools, what should a State do with a K-8 school? If the starting points and intermediate targets are set for K-5, 6-8, and 9-12 grade configurations (the law requires the grade spans to be 3-5, 6-9, and 10-12), the State or LEA might decide to provide two ratings for the school or average the scores across the multiple grades (and average the yearly targets) to provide a single rating for the school. In cases such as these, the State should probably consult with the LEA and school to decide *a priori* about the appropriate unit of analysis.

Full Academic Year.

The law requires States to define a full academic year in order to determine which students will be included in school and LEA accountability decisions. (Neither the 1994 or 2001 Reauthorizations nor the draft accountability regulations include requirements or parameters that States must address in defining "full academic year.") This has been a requirement since the 1994 Reauthorization, but given the stakes associated with the accountability system under the NCLB Act, States are more carefully examining how they have defined this term. There are clear tradeoffs at play here—choosing a longer year reduces the impact of student mobility on school and district scores and generally will lead to higher performance. On the other hand, a longer year can allow needy students to slip through the accountability system, which clearly violates the intent of the law.

As mentioned above, States focusing on a longitudinal approach might require that a student attend the same school (or district) for a full calendar year in order to be included in accountability determinations, whereas those focusing on a status approach might reasonably use an annual cutoff date, such as October 1st. In either case, a State should examine and document the percentage of students that will be excluded from the accountability system based on whatever dates it finally selects. Further, the State should examine the demographic characteristics of this excluded group to see if any of the "accountability subgroups" is disproportionately represented. If so, the State may need to adjust the definition of a full academic year to minimize this negative consequence.

FIGURE 8. DECISION-MAKING PROCESS FOR DESIGNING A STATE SYSTEM TO MEASURE ADEQUATE YEARLY PROGRESS UNDER THE NO CHILD LEFT BEHIND ACT: FACTORS AND CONSIDERATIONS

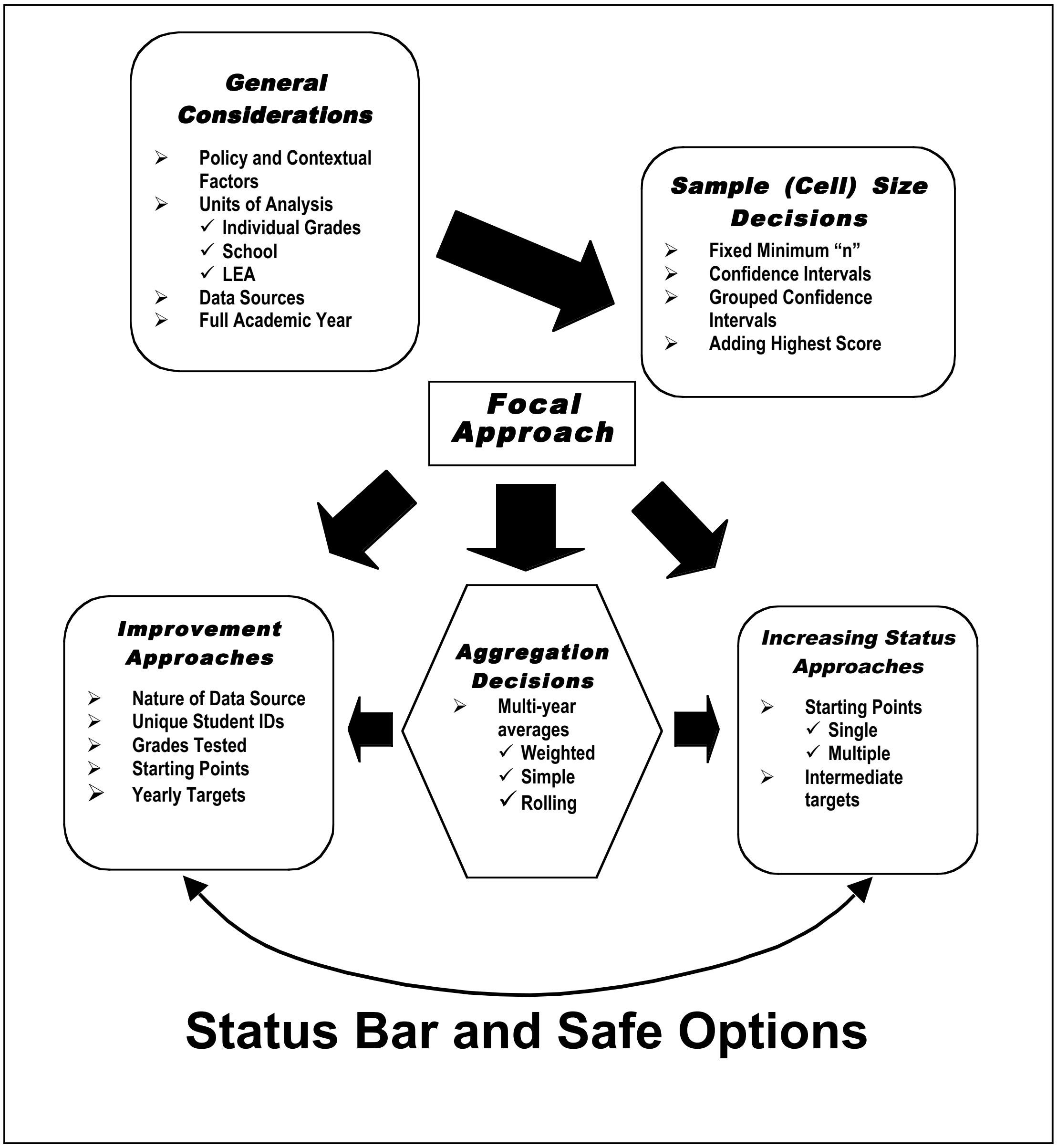
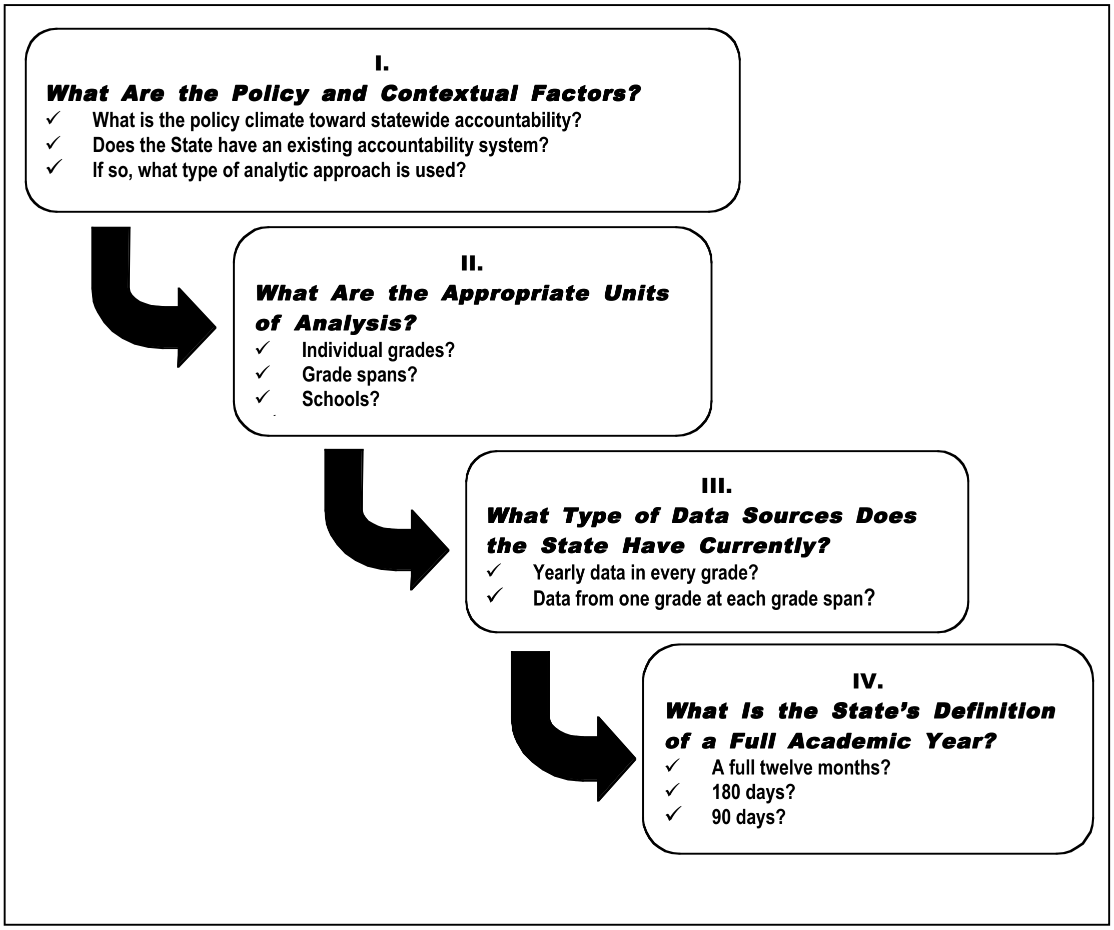


FIGURE 9. FACTORS TO CONSIDER AS STATE OFFICIALS BEGIN TO DEVELOP A METHOD FOR MEASURING AYP



I.
What Are the Policy and Contextual Factors?
- ✓ What is the policy climate toward statewide accountability?
- ✓ Does the State have an existing accountability system?
- ✓ If so, what type of analytic approach is used?

II.
What Are the Appropriate Units of Analysis?
- ✓ Individual grades?
- ✓ Grade spans?
- ✓ Schools?

III.
What Type of Data Sources Does the State Have Currently?
- ✓ Yearly data in every grade?
- ✓ Data from one grade at each grade span?

IV.
What Is the State's Definition of a Full Academic Year?
- ✓ A full twelve months?
- ✓ 180 days?
- ✓ 90 days?

Sample (Cell) Size Issues

How many students does it take to constitute a group? This has been one of the most discussed issues among State leaders since the NCLB Act was enacted. Many initially focused on the minimum sample size (minimum "n") necessary to yield statistically reliable results. Policy makers quickly realized that raising the minimum "n" to levels that could yield somewhat reliable results also allowed them to avoid identifying as many schools as they would with lower thresholds because—according to this early view—one or more student subgroups would have fewer students than the minimum "n" and therefore with fewer hurdles, more schools could meet AYP. Additionally, many individuals have confused the issue of minimum "n" for reporting (a Family Educational Rights and Privacy Act issue) with the minimum "n" for accountability purposes. The discussion here is only focused on sample size related to accountability purposes. State leaders will still need to decide on the minimum acceptable number of students constituting a group for reporting guidelines. Many States use five or ten students as this minimum **reporting** number. The Family Educational Rights and Privacy Act (FERPA) should guide this decision to make sure that no individual students can be easily associated with their particular test scores by virtue of their demographic or other characteristics.

Initial Data Simulations and State Analyses

After the enactment of the NCLB Act in January 2002, several States—perhaps following the lead of Kane, Staiger, and Geppert (2001)—began analyzing student achievement data from prior school years to simulate various impacts of the new AYP requirements on their States. Without exception, their findings parallel the findings of these researchers in their examination of North Carolina and Texas student assessment data for years prior to 2001-02—virtually all schools would be identified for improvement at given points in time. Increasing or decreasing the minimum number of students required to make performance determinations had some impact on both the number of schools identified for improvement and the timing of that identification. In the latter case, more schools tended to be identified while in the former case, some schools with small numbers of low-performing students in subgroups were not identified that should have been. A critical consideration now facing States in the development of their single, statewide accountability systems under the NCLB Act is the determination and justification of the minimum number of students that will be necessary to make statistically reliable judgments about disaggregated student achievement data in the determination of a school or district's measures of AYP. The need to accurately identify low-performing schools must guide States' efforts in this determination.

Minimum "n"

One option that some States have considered to reduce the number of schools identified for improvement is to establish a high minimum number of students (typically 30 or so) required before the results of a subgroup can cause the school to be so identified. While this approach does indeed reduce the number of schools identified (see cautionary footnote in Chapter 1), there is no evidence that it eliminates the *right* schools (e.g., those failing to make AYP on the basis of just a few of the possible 37 cells). Just because a school has less than 30 students in a subgroup doesn't mean that it is doing a satisfactory job with them. **Allowing schools to avoid serving their subgroups simply because those subgroups are relatively small is inconsistent with the intent and goals of the NCLB Act** (Hill, 2002).

The National Center for Education Statistics (NCES) uses a minimum "n" of 62 in order to REPORT subgroup results on the National Assessment of Educational Progress (NAEP). It is very important to understand the difference between minimum "n" used for reporting and minimum "n" used for AYP decisions. While the minimum "n" used for NAEP appears to be a reporting minimum, it is clearly more than that, although it is still not designed for making accountability decisions. The seemingly obscure number of 62 was selected by NCES because that was the sample size necessary to have an 80% chance of detecting an effect size of 0.5 (Allen, Jenkins, Kulick, & Zelenak, 1997). Importantly, this determination was based on calculating an effect size using scale scores, which have been shown to yield significantly more reliable estimates than proportion proficient (Hill, 2001). Therefore, using NAEP as a guide, one could easily justify using minimum sample sizes of at least 62 students. However, selecting a minimum "n" for the accountability provisions of the NCLB Act often requires balancing the technical, practical, and policy considerations.

Preliminary indications are that many States (e.g., Colorado, North Carolina) have recently proposed n=30 as the minimum "n" for accountability decisions. The rationale for this decision usually relates to the point in z or t statistical tables where "things start to level off." On the other hand, a State such as California has proposed using a minimum "n" of 100 or at least 15% of the school population (Crane, 2002). As we will discuss shortly, it can be difficult to rationalize any fixed minimum "n."

The Council of Chief State School Officers' (CCSSO) Accountability Systems and Reporting (ASR) and Comprehensive Assessment Systems for ESEA Title I (CAS) study groups, part of CCSSO's State Collaborative on Assessment and Student Standards, worked with several States earlier this year, using statewide assessment results from prior years, to conduct simulations to determine the extent to which schools in their States would have made AYP assuming the NCLB Act had been in effect during that time. Although most of the State simulations have used only assessment results and have not looked at test participation rate, other academic indicators, or the "safe-harbor" provisions under the NCLB Act, the results of these studies, together with additional data simulations subsequently provided to CCSSO by other States have been consistent. With few exceptions, the State simulation studies show that a high proportion of schools will likely not meet the new AYP requirements within two or three school years.

In May 2002, CCSSO distributed to State Teams, including Chief State School Officers, a report summarizing data simulations from eight States. Among the findings CCSSO reported,

- The percent of schools that would be identified for improvement after two years ranged from 49% to 88% (5 States used a minimum cell size of 10, one 15, one 30, and one did not report minimum cell size).
- In the five States using a cell size between 15 and 30, the percent of schools that would have been identified for improvement ranged from 31% to 88%.
- Clearly, as the minimum cell size is set higher, the percent of schools failing to make AYP declines. States with primarily rural and small schools would see these schools excluded from accountability systems with the full burden of accountability shifted to large schools.
- One implication of setting a higher minimum cell size for subgroups would be that more schools would be deemed to have made AYP but school districts might be determined to not have made AYP as student achievement data are aggregated across schools. A medium size school district may have fewer students in particular subgroups attending its elementary schools than the State requires in making accountability determinations but, district wide, more than enough students in the subgroup to make accountability determinations. Thus, the district is held accountable for the AYP of the subgroup(s) and not any of its schools.

When examining the results of these analyses by many States, it quickly became obvious that raising the minimum "n" to levels high enough to have a noticeable effect on reliability would require samples so large that it would be impractical for many States to set such high thresholds.

The analyses conducted by various States indicates that raising or lowering the minimum "n" is more an issue of consequential validity than of reliability. As seen in Table 4, raising the minimum "n" simply allows small schools an easier way through the system. The percent excluded in Table 4 refers to those schools for which there are not enough students to constitute a group (or school) and we could not say whether or not they met AYP. Therefore, these schools would be allowed to "pass" through the system in that year. Obviously, this effect is exacerbated in States like Wyoming with many small schools, but the principle applies to essentially all States.

TABLE 4. EFFECTS OF ALTERING MINIMUM “N” ON “PASS” RATES OF SCHOOLS IN WYOMING

MINIMUM “N”	% Meeting AYP	% Excluded	TOTAL “PASS
3	20.5	9.3	29.8
5	28.8	13.2	42.0
10	39.5	22.9	62.4
30	39.5	47.8	87.3

Given the potential stakes associated with the decision, some have suggested that a sample of n=200 be used to yield results precise enough to support such decisions (Gong, personal communication, August 13, 2002). However, even a sample size of 200 with 50% of the students classified as proficient yields a standard error of approximately 3.5% or a 95% confidence interval of ±7%. Linn (2002) has recently suggested that approximately n=25 might strike a reasonable balance between decision consistency and practicality, but Linn also cautioned that this could still result in many potentially unreliable decisions. We agree with the second clause in Linn’s statement because the standard error with n=25 and 50% proficient is 10%, meaning that for an observed proportion of 50%, one could be 95% certain that the true proportion would be between 30% and 70%. Even with 90% (or 10%) proficient, the standard error would be 6%, yielding a 95% confidence interval of ± 12%. This range of uncertainty easily masks most changes (e.g., 5-10%) one could hope to observe.

On the other hand, what if the minimum was n=30, the target performance was 40% proficient, and you had a school with a subgroup of 25 that consistently had fewer than 10% proficient for that group? Even though this school (or student subgroup) had fewer than the minimum “n”, we believe that State officials could be confident in identifying that school as not meeting AYP. In fact, it could be easily argued that not identifying that particular school could be a disservice to the students in that school. This type of situation has led many to believe that the minimum “n” has much more to do with the consequential aspects of validity (see Chapter 2) and less to do with reliability. Finally, it is important to point out that regardless of the minimum “n” a State establishes, almost all schools will end identified for improvement within five or six years; therefore, it makes sense to try to find the most appropriate way to attend to educationally disadvantaged subgroups in the near term. Increasing the *minimum-n* to an unrealistically high level will not serve this purpose.

Using Standard Error Approaches

Now that the flaws with selecting a single minimum sample size have been documented, State leaders still need a way of identifying schools for improvement and most would like to be able to do so with confidence that they are not misidentifying a significant percentage of schools. In the discussion of minimum “n” above, the principle of considering confidence intervals was introduced. We suggest that using a confidence interval approach can ameliorate many of the pitfalls associated with using a single minimum “n”, although that does not mean this approach is entirely free of potential problems.

The NCLB Act focuses on proportions of students meeting a particular target proficiency level and the observed proportion for a school (or school district) is compared to this target proportion. There is some debate nationally about the appropriate inference and therefore about which score the confidence intervals should be constructed. As mentioned in Chapter 2, it is the inferences made as a result of the accountability system that are valid or not; therefore, State leaders should attempt to explicitly articulate the inferences they intend from their accountability system.

If one wants to establish a measure of certainty regarding whether the observed proportion is characteristic of this school, then analysts can simply use the formula for the standard error of the proportion[17] to construct confidence intervals around the observed proportion. This would allow one to infer how well the observed score represented the "true" percent proficient for that school given a sample of all possible students who could attend that school. The resulting confidence intervals would then be compared to the target performance level. If the school or student subgroup had an observed score below the target, but the upper confidence interval was greater than this target, the school would be classified as having met AYP for that year.

If a State is trying to calculate confidence intervals for schools with small sample sizes, standard error approximated using the normal distribution may not be accurate enough for some applications. Glass and Hopkins (1984) offer the following guidelines for determining whether a sample is too small to use the approximate method. If the sample size (n) multiplied by the proportion proficient AND the sample size multiplied by one (1) minus the proportion proficient are both greater than five (5), an approximate standard error calculation can be used. If both of the Glass and Hopkins thresholds are not met, an "exact" test should be used. An "exact" method based on the binomial distribution as well as a Microsoft Excel tutorial is available at: http://www.itl.nist.gov/div898/handbook/prc/section2/prc242.htm.org.

Another inference that could be the focus of these analyses is one that asks, "Is this observed proportion different than the target proportion?" For example, if the target percent proficient and advanced for language arts is 50% and a given school with 100 tested students has 40% of these students scoring proficient and advanced, the question becomes, "Is 40% (for 100 students) significantly lower than 50? In this case, the appropriate statistic is the standard error of the difference and then using a simple z-test to see if the observed difference falls outside of whatever confidence intervals (e.g., 68%, 90%, or 95%) the State is using. This approach is illustrated as follows[18]:

- The hypothesis tested in this case is, "the observed proportion is different than the target (population) proportion."
- The formula for the z-test is: $z = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$

  Where π is the population proportion proficient (or in this case, the statewide target for proficiency) and p is the proportion proficient in the school or district.
- The value of z is then compared with the critical value of z to determine of the observed difference is statistically significant. For example, if we were testing this difference at the .05 probability, the value of z is compared with z_{crit} of 1.645[19] and if the observed z is greater than 1.645, we can conclude that the observed proportion is significantly different than the target proportion. If this observed proportion is less than the target proportion, it can be concluded that the school or district did not meet their AYP target.

There are some obvious communication/public relations issues that arise when using a confidence interval approach. For example, two schools could have observed proportions below the bar and one with 500 students could have a higher actual score than a school with 50 students, but because

[17] SEp = sqrt (pq/n).

[18] For purposes of this procedure, an approximate standard error calculation will be used. Obviously, if the sample is small, the exact method should be used.

[19] This value of z_{crit} is for a one-tailed test, which we argue is the appropriate test for these analyses.

the confidence intervals would be so much wider for the smaller school, it could have "made AYP" while the larger school might not. This would require providing some easy-to-understand information about why and how the State is using a confidence interval approach. Because calculating confidence intervals might be beyond the capabilities of many district officials, State officials should consider publishing an "approximate" confidence interval table similar to the one depicted in Table 5.

TABLE 5. REQUIRED DIFFERENCE BETWEEN THE OBSERVED PERCENTAGE OF STUDENTS SCORING PROFICIENT OR ABOVE FOR A GIVEN SCHOOL AND THE STATE PERFORMANCE TARGET TO ACHIEVE A MINIMALLY (68%) STATISTICALLY SIGNIFICANT* DIFFERENCE[20]

Percent of Students Proficient and Above	Number of Students Tested in School/District				
	10	25	50	100	200
10%	9%	6%	4%	3%	2%
20%	13%	8%	6%	4%	3%
30%	14%	9%	6%	5%	3%
40%	15%	10%	7%	5%	3%
50%	16%	10%	7%	5%	4%
60%	15%	10%	7%	5%	3%
70%	14%	9%	6%	5%	3%
80%	13%	8%	6%	4%	3%
90%	9%	6%	4%	3%	2%

*To find the required percentage to achieve a 95%, one-tailed statistically significant difference, multiple the percent in each cell by 1.645.

To use Table 5, one needs to find the intersection of the approximate number tested from the top row with the percentage of students scoring proficient or above from the left column. The cell at that intersection indicates the minimally required percentage difference between the school's performance and the State target for the school to be considered above or below the target performance. For example, if the State target is 50% proficient and a school with 50 students achieves a percent proficient of 40%, we find a value of 7%; therefore, we could be 68% certain that this school is below the target value. However, considering the stakes associated with this decision, most would argue for being at least 95% certain of this decision. Since we are really only concerned whether a school with a percent proficient lower than the target (not if it is lower than the observed percentage) is "truly" lower, a one-tailed statistical test is appropriate. Therefore, to find the 95% one-tailed confidence interval for the current example, we would multiply 7% by 1.645 (the critical z-value for a one-tailed, 95% confidence interval). The resulting value of 11.515 would be added to the observed 40% to arrive at 51.515%, which is higher than the target of 50% proficient. Therefore, we would not be certain enough to say that this particular school did not make AYP and would classify the school as "passing" (or having met AYP) for that year.

Another approach for handling potential communication issues is to use a single minimum "n" as a "first cut" and then to apply confidence intervals to school scores falling below the AYP target.

[20] From *Interpreting Wyoming Comprehensive Assessment System School and District 2002 Reports*. Cheyenne: Wyoming Department of Education.

Using this approach, a State could set a smaller minimum “n” than might be expected based on reliability concerns alone because State leaders could be confident that they will avoid misidentification by applying confidence intervals. For example, using this “compromise” approach, State leaders could establish a minimum “n” of 15 or 20 and then for any school/subgroup failing to meet the performance target, the State analysts can offer to take a “second look.” While we do not think that it takes any more work to calculate confidence intervals or z-tests for all schools compared with a subset, this approach might be easier to communicate to educators and the public[21].

The Highest Score Method

Some have suggested that using a confidence interval approach might be hard for State leaders to explain. Certainly, the confidence interval approach allows the State to hold all schools—no matter how large or small—accountable, while a single, fixed minimum-n allows many schools to appear to avoid accountability. Another method (Hill, 2002, personal communication) has been suggested that would use a single, fixed minimum-n, while holding all schools accountable. This would entail adding fictional proficient students to a school or subgroup with fewer than the fixed minimum-n to give the school the benefit of the doubt, while still holding them accountable. The following is a brief example to walk the reader through such a hypothetical approach:

- Assume the minimum-n is 30 students and the performance target is 50% proficient.
- Also assume the school has 25 students, 9 (36%) of whom are proficient.
- In order to have enough students to meet the minimum-n, we would “add” five (5) fictional proficient students so the school would have 30 students, 14 (46.7%) of whom would be considered proficient.
- Since 46.7% falls below 50%, the school would not meet its AYP target.

This system is not based on traditional statistical methodology (as is the case with a confidence interval approach), but it clearly gives schools the benefit of the doubt, which, considering the consequences associated with falsely identifying schools, makes a great deal of sense and provides a method for holding all schools accountable.

Improvement vs. Increasing Status

The “safe harbor” provisions in Section 1111(b)(I) are not specific regarding whether or not a State must measure the reduction in percent non-proficient using a longitudinal or cohort framework (see also Appendix B). A longitudinal approach compares the performance of individually matched students from one grade to the next, whereas as a successive cohort approach compares the results from a particular grade one year to the results from that same grade the following year even though different students are being compared. If mobility is not a major issue and the State has a data system that allows for tracking individuals, a longitudinal approach has been shown to yield more consistent results (Hill, 2001) than a successive cohort approach. Carlson (2001) has demonstrated that even a quasi-longitudinal design produces more consistent results than a simple successive cohort model. A quasi-longitudinal model does not require matching individuals across years, rather it relies on the comparison of means (or other statistics of interest) from a particular grade in year one to the results from the next grade in year two. With highly mobile populations, the results from this approach will not differ very much from the results of a successive cohort approach.

[21] Thanks to J. P. Beaudoin from the Louisiana Department of Education for suggesting this approach.

When the NCLB Act became law in January 2002, most believed that an increasing status approach would be the only accountability approach allowed. Most argued that this is the only approach permitted under the law. However, even if one believed this were the case, a plethora of analytical and policy decisions remain. These issues and decision points, many of which are shown in Figures 8 and 9, illustrate the different approaches a State might choose. We discuss many of these issues below for both improvement and increasing status models.

Starting Points

Even though it is not mentioned in the law or draft accountability regulations, the Secretary's letter (Paige, 2002) and the Draft Pilot Peer Review Process Document[22] used by ED in the initial review of a few, selected State accountability systems during September 2002, appear to permit using variable starting points for different student subgroups as long as all are on a trajectory to reach 100% proficient in the required subject areas by 2014. If this flexibility were permitted in the final regulations on accountability systems, States would have to decide whether to use a single starting point for each academic content area (by grade span) or calculate different starting points for each student subgroup (by grade span). Different starting points allow the analysis to more accurately reflect where each subgroup is presently, but it also means that subgroups starting lower have to make a much steeper climb to reach 100% by 2014. Nevertheless, we believe that such an approach would still meet the spirit of the NCLB Act while at the same time presenting a more flexible approach for States. State analysts should model the effects of using common or different starting points to see which approach will help them correctly identify the schools most in need of improvement. Unfortunately, like many of the decisions that State leaders will need to make, this one will not simply be an analytical decision. Policy makers will ultimately have to decide whether the use of a common starting point will force more appropriate attention on student subgroups or whether the targets will be set so far out of reach that educators might get discouraged and give up. Additionally, policy makers may have to wrestle with potential fairness issues if constituents complain that one group is allowed to start at a lower point than another.

The law is quite specific about how to calculate starting points for an entire State [Section 1111(b)(2)(E)]. The law requires State officials to choose between the "20th percentile method" school and the lowest performing subgroup, yet every State that has analyzed existing assessment data has found the 20th percentile method yields the higher starting point and, as required by law, States are required to use the higher of the two starting points. The final regulations do not permit calculating the starting points separately for student subgroups.

One other issue or question related to setting starting points concerns whether States, that might want to, can average two or three years of data (e.g., 2000-01 and 2001-02) to set starting points. The final accountability regulations permit this option as long as 2001-02 data are included (see Section 200.16, Comments/Discussion, p. 71742). This issue is also addressed in Chapter 1 as a variable that could impact the decisions States must make in designing their accountability systems.

The authors are also aware that at least one State has proposed setting starting points and trajectories for each student subgroup by school. It is unclear whether this amount of flexibility will actually exist when final accountability regulations are promulgated, but we question this

[22] While the Draft Pilot Peer Review Process Document has not yet been disseminated, members of the Study Group have had an opportunity to review this document. We caution that the assumptions of flexibility are yet to be affirmed in related final regulations.

strategy on reliability grounds[23]. There is imprecision in every observed proportion and this imprecision is largely related to sample size. Setting the starting points, and therefore the twelve-year trajectories, on the basis of 30 or so students (a typical group size) will certainly be considerably less reliable than setting a starting point on the basis of thousands of students.

Intermediate Goals

The law affords States a choice in terms of how annual goals are established. States can either raise the status bar every year or they can hold the bar steady for up to three years at a time before raising it to a new level, but in both cases the trajectory must reach 100% proficient by 2013-2014.

Raising the bar annually is fairly easy; simply divide the difference between the starting point and 100% by 12 years to arrive at the increase in the status bar required each year. Raising the bar intermittently, but at least every three years with the first increase required by the 2004-2005 school year, affords some flexibility but leads to a series of decisions. There are several reasons why a State might choose one approach over another. In terms of communication, some have suggested that raising the bar annually will help keep the focus on regular yearly goals, while others have suggested that it will be easier to communicate three or four changes in the bar over twelve years by keeping the bar steady for three years at a time. There may be some technical advantages to maintaining the status bar at a consistent level for multiple years. As demonstrated throughout this document, sampling error, especially for those close to the bar, can have a noticeable impact on whether a school has been identified for improvement. By chance alone, several schools will "bounce" over the bar in one of the two or three years. For example, a simulation in one State (Wyoming) found that 25% of the schools in each of two years fell below the bar, but only 15% fell below the bar in both years. Obviously, if this simulation was carried to a third year when the bar would be raised significantly, many of the 25% below AYP in year-2 would have a harder time scoring above the bar in year-3 because of chance alone.

States choosing to focus on an improvement approach have a slightly different set of issues. These States are required to set an annual target for improvement (i.e., "safe harbor"), and they may choose to raise the status bar every year or every third year. State leaders should analyze the effect of raising the status bar yearly or every third year on the improvement approach prior to selecting their strategy. It could be argued that maintaining a steady target for three years would allow the State to "feature" the improvement methodology. On the other hand, knowing that a substantial status increase is looming in the third year of the intermediate goal could cause school and district personnel to focus on the status bar, rather than improvement. Obviously, State leaders must align the required yearly improvement targets with the status goals whether these are yearly or early third year.

Aggregating Data

The law is quite clear that States are responsible for classifying schools and school districts regarding their status relative to AYP each year. Some individuals have suggested that for very small units (e.g., n<10), States should be allowed to withhold these decisions so they are made

[23] Neither the Secretary's July letter (Paige, 2002) or the draft regulations on AYP (Federal Register, 2002, August 6) seem to suggest that setting starting points by school building would be consistent with NCLB Act requirements. Section 1111(b)(2)(E), requires States to establish starting points based on statewide student performance data and sets forth the manner in which the starting points are determined. Section 1111(b)(2) requires States to establish a single statewide accountability system. The former requirements do not provide for a manner in which starting points would be calculated for individual schools. Having individual school AYP requirements would serve to work against the requirement for a single, statewide accountability system.

only every other year. While this sounds very sensible, it is clearly not permitted under the NCLB Act. The law does permit, however, several approaches for aggregating data that can be used to increase the confidence in annual accountability decisions. The law refers to a "uniform averaging procedure," but this term has not been clearly defined in either statistical or policy terms. The authors, as a result of numerous conversations with ED officials, are convinced that methods such as multi-year averages (weighted or simple) and rolling averages are permitted, as long as each school or district is classified each year. It is important to point out that these data aggregation techniques can be used for either or both, status or/and improvement evaluations, but the choices play out somewhat differently for each approach.

Multiple-Year Averages.

States are clearly permitted under the NCLB Act to use multiple years of data to establish starting points as well as to use multiple years of data to establish the annual status measurements. Further, even though it is not clearly provided for in the law, the Study Group believes that State leaders could also use multiple years of data to establish starting points. For example, a State may average data from 2000, 2001, and 2002 to establish the baseline for 2002. The State could then average data from 2002 and 2003, for example, to calculate each school's status for 2003. Why would a State want to combine data across years? The most important reason is that it will allow States to base estimates of school performance on larger samples and thereby reduce the standard error of the observed proportions. For States that plan to use the minimum "n" approach, combining data across years will lead to more schools and student subgroups meeting the minimum "n" threshold so they can be held accountable for progress toward the goal of all students at or above proficient in reading or language arts and mathematics by 2014.

Simple or Weighted Averages?

But how should a State average these data? Should State analysts simply average the percent proficient from each unit across the two or three years or should they weight the average by the number of students enrolled in the given school? Arguments can be made for either choice. For example, if the State believes that each year is an unbiased estimate of a school's performance, then simple unweighted averages would be appropriate. On the other hand, if the State is considering aggregating performance across two years as a "single" estimate of a school's performance, then counting each individual student from each year (i.e., weighted averages by enrollment) would be the sensible approach. In this case, the State leaders would aggregate the two years of data into a single file and treat all students as belonging to a single year's estimate.

Rolling Averages or Multi-Year Comparisons?

To this point, our discussion has focused on combining data across multiple years to improve status estimates, but States can combine data when using improvement-based approaches (i.e., "safe harbor"). When thinking about combining multiple years of achievement, most people immediately think of rolling averages, such as comparing the average scores of 2000, 2001, and 2002 with the averages of 2001, 2002, and 2003. However, a little simple algebra will show that rolling averages simply compares the first year to the last year, with the addition of constant, in this case 2000 compared to 2003. Therefore, the use of rolling averages does little to truly improve the reliability of the comparison. Successive multi-year comparisons provide a means for combining multiple years of data to help evaluate trends. Using the example above, we could compare the average of 2000 and 2001 with the average of 2002 and 2003. This would provide a State the real benefit of stabilizing the comparisons without the problems associated with simple rolling averages. Consider the following example to examine these two approaches:

- A school's results for these four years were as follows:

	2000	2001	2002	2003
# Tested	30	50	60	70
%Proficient	25%	35%	45%	40%

- If we used a simple average to characterize the 2000-2001 results, the average would be 30% proficient, but if we used weighted averages, the two-year average would be 31.25%. While this is not a big difference, one can imagine that with larger differences in either number of students and/or percent proficient, the difference between simple and weighted averages would be exacerbated.
- To demonstrate the difference between multi-year comparisons and rolling averages, consider the data in the table above. Using three-year rolling averages, we would compare the average scores (weighted or simple) for 2000, 2001, and 2002 with the average scores for 2001, 2002, and 2003. Using simple averages for the purposes of this discussion, the average of 2000, 2001, and 2002 is 35% proficient, and this would be compared with the average of 2001, 2002, and 2003 of 40.0% proficient, a 5% gain.
- Using multi-year comparisons (again using simple averages for this example), we would compare the average of 2000 and 2001 (30% proficient) with the average of 2002 and 2003 (42.5% proficient) or a 12.5% gain.

As seen in this example, the rolling average can mask changes by adding a constant to each side of the equation, and it does little to improve the reliability because once the same terms are subtracted from both sides of the equation (in this example, 2001 and 2002 scores), the reliability is based on a simple comparison (in reliability terms) of 2000 with 2003. On the other hand, the multi-year comparison can improve the reliability by essentially doubling the sample (assuming the same school size in each year), thereby reducing the standard error. Every doubling of the sample size leads to a reduction in standard error of approximately 30%. Not only does our ability to detect real change improve (i.e., narrowing the confidence intervals), avoiding having the constant falsely masking potential change improves our power to detect differences if they occur.

Summary

This chapter has attempted to provide a "roadmap" for the many decisions State leaders need to make as they build or refine their accountability systems for determining Adequate Yearly Progress under the NCLB Act. In doing so, we have tried to apply many of the principles of validity raised in Chapter 2. We attempted to remain particularly cognizant of the consequential aspects of construct validity when suggesting particular methodological approaches. This chapter was not intended to provide an exhaustive recipe for calculating AYP. Rather, we attempted to highlight various approaches for dealing with the many of the most crucial decisions required of State leaders. In doing so, we suggested methods that go beyond the simple intuitive solutions one might draw when first reading the law. For example, the law does not mention confidence intervals at all and many State leaders assumed that they had to search the statistical literature for a single, "magical" minimum "n." However, the law states quite strongly that AYP approaches need to be statistically valid and reliable, and the minimum "n" selected should also meet these same criteria. We considered this to be an impossible task within the bounds of practicality and therefore suggested using an approach that would clearly meet both the spirit and the intent of the law.

Finally, this chapter was intended to raise as many questions as it answered—perhaps more. However, these questions are designed to help State leaders focus their design discussions and weigh options prior to submitting accountability plans early in 2003. Again, the validity of the accountability inferences as well as the spirit of the law needs to serve as foci for these design discussion. We hope this chapter provided some tools to facilitate these discussions.

References

Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

Bailey, A. Y. (2002). *Incorporating federal requirements into state accountability systems—Lessons learned from four states.* Washington, DC: Council of Chief State School Officers.

Carlson, D. (2001). *Using cross-sectional comparisons to identify effective schools: Roulette anyone?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Crane, E. (2002, September). *What the New ESEA means for state accountability system, California's perspective.* Paper presented at the annual CRESST Conference, Los Angeles, CA.

Federal Title I Regulations for the U. S. Department of Education, 67 Fed. Reg. 45037-5047 (2002, July 5) (codified at 34 C.F.R. pt. 200).

Federal Title I Regulations for the U. S. Department of Education, 67 Fed. Reg. 50986-51027 (2002, August 6) (to be codified at 34 C.F.R. pt. 200).

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Hill, R. (2001, October). *Issues related to the reliability of school accountability scores.* Paper presented at the Reidy Interactive Lecture Series, Nashua, NH. [On-line]. Available: www.nciea.org

Hill, R. (2002). *Three suggestions for the immediate improvement of the NCLB design. Unpublished paper.* Portsmouth, NH: The National Center for the Improvement of Educational Assessment. [On-line]. Available: rhill@nciea.org

Kane, T. J., Staiger, D. O., & Geppert, J. (2001, July 15). *Assessing the definition of "adequate yearly progress" in the House and Senate education bills.* Los Angeles, CA: School of Public Policy and Social Research, University of California, Los Angeles.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). *Randomly accountable.* [On-line]. Available: http://www.educationnext.org/20021/56.html

Linn, R. L. (2002, September). *Measuring adequate yearly progress.* Paper presented at the Annual CRESST Conference, Los Angeles, CA.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (June 2002, June). *Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001* (CSE Technical Report

567). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Olson, L. (2002, August 7). Long-awaited ESEA rules are released. *Education Week,* XXI, 1, 36-38.

Paige, R. (July 24, 2002). *Dear colleague letter to education officials regarding implementation of the No Child Left Behind Act of 2001.* Washington, DC: U. S. Department of Education. [On-line]. Available: http://www.ed.go./News/Letters/020724.html

U. S. Department of Education. (2002). *Draft pilot peer review document.* Washington, DC: Author.

Chapter 4

Summary and Conclusions

This paper has addressed a set of issues raised in Title I of the NCLB Act related to making valid and reliable decisions in the calculation of adequate yearly progress (AYP) [see Section 1111(b)(2)(C)(ii) & (v)(dd)]. States are required to submit to the U. S. Department of Education (ED) descriptions of their statewide accountability systems by January 31, 2003. Mindful of the challenge to make valid and reliable decisions about schools and the prescriptive nature of the NCLB Act for making that decision, this paper has been framed by a comprehensive analysis of the validity of accountability systems prior to discussing the mechanics of calculating AYP. The determination of various aspects of AYP such as starting points, intermediate goals, and annual objectives are considered using a strict interpretation of the NCLB Act and, subsequently, relying on the limited flexibility already signaled as a possibility by the Secretary of Education. Considerations relevant to determining sample size "sufficient to yield statistically reliable information for each purpose for which disaggregated data are used and justify this determination" as prescribed by the law are also discussed.

This paper has emphasized, as the law does, the importance of building an accountability system that provides confidence in the validity of decisions made about districts and schools in determining AYP. The challenge presented to States, given the constraints imposed by the prescriptive nature of the law, is to make decisions about the elements of the system that maximize the reliability of the components and the validity of the decisions that emerge from that system. The primary construct validity questions that States must consider are:

- Is the system focusing on the "right" goals?
- Does the accountability system identify the schools that truly need to improve?
- Is the accountability system theoretically and logically related to improved student learning?

However, accountability systems under the NCLB Act requirements will lead to consequences and, therefore, States must consider the consequential aspects of validity when designing and evaluating their accountability systems. Some of the consequential questions a State must consider include:

- Is the system having the desired impact?
- Is the system leading to more or less equality of educational opportunity for all students?

- Is the accountability system leading to unintended negative consequences such as teachers leaving the profession early or has the curriculum become unacceptably narrowed?

Building a Case for the Validity of the System

The following steps would be involved in the process of building an argument supporting the validity of an accountability system:

1. Examine closely the intended outcomes, the first of which are the assessment results. In the NCLB Act, one must determine if a school, or what proportion of schools, met their AYP requirements. One then determines the number of subgroups in the school or district that failed to meet the AYP targets.
2. Corroborate the "official" findings through a disciplined search for additional information. Judging the impact and validity of the accountability system will require other data. These other data might include outcome measures such as other assessment data, process measures about the quantity of writing assignments, attitudes and opinion information about parent and client satisfaction, and teaching and learning information about the level of emphasis and time teachers devote to instruction of key academic content standards.
3. Check the design and implementation of each component in the system for any evidence of lack of reliability or other problems. Each component has to be defined and implemented in harmony with the functions of the other components in light of the system's purposes. Components of the system may be categorized as:
   - Setting purpose and focus or goals—standards, target for improvement, and theory of action for reform;
   - Selecting indicators—assessments aligned to standards, graduation rates, attendance rates;
   - Data collection, scoring, and analysis;
   - Drawing inferences and making decision indicators—rules for determining AYP status;
   - Implementing the decisions—determination of AYP; and
   - Evaluating the effects of the decisions—impact of the accountability system.
4. Examine the level and quality of the implementation of the reforms, from the actual classifications of the schools to the fidelity of the selected school reform efforts.
5. Conduct these analyses on several levels—at least statewide and for particular types of schools, perhaps selected on the basis of size, geographic area, and student population.

An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that

- The components of the system are aligned to the purposes and are working in harmony to help the system accomplish those purposes; and
- The system is accomplishing what was intended (and did not accomplish what was not intended).

Whether the State accountability system is in the early stages of implementation or is well established affects the focus of the analysis. In the early stages of implementation, mainline outcome information such as the impact of the accountability system on schools and whether learning actually improves may not be available. Therefore, the focus must center on the clarity of the descriptions of the separate components, the completeness of the definitions, the coherence among the different components of the plans, and the feasibility of the plans. As systems are implemented, the focus can expand to the effects of the implementation efforts. Examining, reporting, and making a case for the validity of the system are important elements of meeting the NCLB Act requirements as well as building public confidence in the decisions made as a consequence of that system.

Making the Case for Validity: Searching for Both Positive and Negative Evidence

In judging the validity of the system, the State must also examine ways in which the system can be rendered invalid. The sources of a system's lack of validity (or lack of reliability that leads to a lack of validity) are of two general types:

a. The definition or implementation of any one stage or component is defective in some way.
b. Two or more of the stages are in conflict or out of alignment, either conceptually or computationally.

While these two sources of invalidity result from parts of the system being invalid, in some cases, the entire system may be in conflict with approaches most likely to lead to improvements in student learning, rendering it invalid. The two general types of problems described here can be separated into two categories—intra-component problems and inter-component problems.

Intra-Component Problems.

1. Errors of various kinds. Look for the most frequently occurring errors in analysis, reporting, or in the application of decision rules that are usually a result of human or technological error. These may result from incorrect definition of variables, input and keyboarding errors, use of the wrong data files, use of wrong scoring keys, or wrong formulas are applied are all among the possibilities.

2. Conflict between design and implementation. It is crucial that the definitions and specifications for each component and subcomponent be inspected. Sometimes changes occur in definitions that are not carried forward into data analysis or procedures.

3. Improper definition or formulation of a component. With the seriousness of identifying a school for improvement, it is important that schools be accurately identified, and that process requires a proper conceptualization of a school as a place which educates not just the students enrolled in a given year, but one that is responsible for—and must be judged on the basis of—how well it educates different groups of students over the years. One of the largest sources of unreliability in an accountability system is "student variance"—the variability of students from one year to the next. If the reliability of the results and decisions made by an accountability system do not take this into account, the decisions about the quality of a school's program are more random than real resulting in some schools being identified for improvement one year and not the very next year, with no changes in the school's program as may be expected. One way to compensate for the random variation might be in the computation of proper standard errors. Infinite formulas that produce larger standard errors decrease the chances of making false judgments about a school.

Inter-Component Problems: Mal-Alignment Among a System's

The first inter-component conflict, (a), is between the purposes and the selection of indicators. The second, (b), is between the purposes and the decision rules for identifying schools for various kinds of intervention. Illustrative descriptions of potential mismatches follow:

(a) Mal-alignment between purposes and indicator selection. This can occur when States have academic content standards in many areas but judge schools based on only one or two areas such as reading. This naturally restricts the focus of instruction since teachers will spend more time on those areas for which they are held accountable. If this is spelled out in the purposes for the accountability system as an acceptable outcome, then the system is valid. If not, the technical definition of validity is violated; the system is not measuring or focusing on the purposes that it purports to.

(b) Mal-alignment between purposes and the rules for interpreting the results and making school classification decisions. Accountability systems need to be dealt with as a whole, meaning that the parts must be in alignment. The model should be selected on the basis of the purposes of the system, and the views that it presents about the nature of schools and the best ways to improve them. The model itself will usually dictate the rules for interpretation and decision-making based on assumptions about the definition of a "good school" and a "bad school." Problems arise when the decision rules are not compatible with the purposes of the system. An analysis of some likely sources of invalidity due to mal-alignment between purposes of an accountability system and its design and decision rules is provided in Exhibit 1 at the end of Chapter 2 and included in the following discussion of Consequences and Policy Issues for the decision points that follow.

Toward a Framework for State AYP Plan

In posing a decision framework and a series of steps that can be followed when developing systems for calculating AYP, this first paper in the series addresses such topics as the *nature of data sources, number of starting points, sample size issues, aggregation issues, and setting intermediate goals*. Two factors drive the discussion for each issue: (1) adherence to the NCLB Act provisions and expectations and (2) technical defensibility. States need to build systems that honor the call for valid and reliable AYP decisions in the law, and clearly set forth how the chosen system design proposes to adhere to the provisions and expectations of the NCLB Act in light of these considerations.

Essential to choices made in the design of the system at each decision point, States must consider the alignment of the components in the system to purpose. The first is between the purposes and the selection of indicators. The second is between the purposes and the decision rules for identifying schools for various kinds of intervention (although this latter point is not discussed in this paper).

There is no question that the NCLB Act allows for a combination of a status and an improvement approach as reflected in the "safe" harbor provisions in making accountability decisions. There is a question concerning whether an improvement-based model alone such as a value added model or improvement judgment first procedure such as determining "safe harbor" before status would be permitted under the law. The “safe harbor” provision clearly allows schools, if they first have not met the change in status requirement, to meet AYP if they reduce from one year to the next—for the subgroup that did not meet the status target—the percent non-proficient by 10%. However, a school finding itself using “safe harbor” every year to meet AYP for all 12 years is not likely to end up with 100% proficient by 2014. Although this outcome seems to be an unintended consequence of the law, States are not likely to be permitted to propose a model that leads to significantly fewer than 100% of all students reaching proficiency. The authors suggest that it may be possible that, by adjusting the goal so that it reflects all students reaching proficiency in 2014, an improvement model might meet both the letter and intent of the law under a more flexible interpretation of the NCLB Act.

Several factors will influence initial choices of approaches in designing accountability systems:

- States that were fully compliant with provisions of the 1994 Reauthorized Elementary and Secondary Education Act may have only assessed student achievement once in each grade span (3-5, 6-9, and 10-12) making the use of a longitudinal system virtually impossible.
- Availability of assessment data at each grade 3 through 8 allows for more options, especially data with unique student identifiers.
- The size of schools and student subgroups will also affect the decision. Small schools with data collected only three times over the K-12 span (consistent with 1994 requirements) will have a difficult time using an improvement model.
- Mobility rates of students could also have a considerable impact on a State’s choice of approaches, because longitudinal approaches would exclude large numbers of students in schools with high mobility.
- Status approaches might better meet the intent of the law because they will tend to include more students than longitudinal methods that require matching of students across years. In particular, States with high mobility rates may need to define “full academic year” as something shorter than a full calendar year to maximize students included.

Still, student variability across years may have very real implications for the validity of accountability decisions. Differences in groups of students from one class to the next, that is, from one year to the next, may result in a school meeting AYP one year, not the next and meeting AYP the third even though the school’s instructional program has not changed and students are learning better each year they are in the school. The State should consider whether using longitudinal or successive group frameworks may increase the validity of accountability decisions. The degree of mobility in a State or school may also affect this choice:

- If the State has a data system that allows for tracking individuals, a longitudinal or quasi-longitudinal approach will yield more consistent results than a successive group approach.
- A quasi-longitudinal design does not lose as many students from the system and produces more consistent results than a simple successive group model.

With this established, the paper then provides the analytical and policy decisions related to a status model as prescribed by the NCLB Act. The following summary presents the decisions to be considered, some possible consequences of those decisions, and the policy implications of those decisions and consequences.

Minimum “n”

Minimum “n” is incorporated in the NCLB Act provisions to ensure that State accountability decisions meet a certain threshold with respect to the validity and reliability of their decisions such that they will not be undermined by a sample size too small to be reliable. States must explain to ED the minimum “n” they will use in their accountability system and justify the related decisions they have made.

As discussed in Chapter 3, it is also very important to recognize, and distinguish between, the difference between minimum “n” used, for example by the National Center for Education Statistics (NCES), for reporting and the minimum “n” States might use for school and district AYP decisions. While the minimum “n” used by NCES for NAEP purposes appears to be a reporting minimum, it is clearly more than that, but it is still not designed for making accountability decisions. The decisions States must make regarding minimum “n” considerations for AYP are different than those that NCES, for example, needed to make for NAEP purposes.

- **Decision:** What minimum “n” yields the most reliable decisions, but also does not lead to negative consequences by under-identifying schools that should be identified for improvement and over-identifying those that should not be so identified?
    - While a minimum “n” of 200 to 1,000 might be needed to make highly reliable decisions, those decisions would have little validity for making AYP decisions about a school under the NCLB Act.
    - While a minimum “n” of 25 to 30 might strike a reasonable balance between decision consistency and practicality, this may still result in many potentially unreliable decisions.
- **Consequences:** Assuming the NCLB Act had been effect earlier, several States examined student performance data from past years to conduct simulations to estimate the impact of the law’s new accountability requirements on their districts and schools. The following likely consequences of setting various minimum “n’s” were found:
    - Increasing or decreasing the minimum number of students required to make performance determinations has an impact on both the number of schools identified for improvement and the timing of that identification.
    - Some schools with small numbers of low-performing students in subgroups will not be identified that should be, because the number of students in the subgroup is insufficient to make a reliable judgment about AYP status.
    - Raising the minimum “n” to levels high enough to have a noticeable effect on reliability would require samples so large that it would be impractical for many States to set such high thresholds.
    - Regardless of minimum “n” used, virtually all schools would be identified for improvement at given points in time, with high proportions failing to meet AYP within two to three years ranging from 49% to 88% according to State simulations and analysis of existing data.
- **Policy Implications:** Based on an examination of results from the simulation studies described above, the following conclusions about the likely consequences of setting various minimum “n’s” were drawn:
    - Minimum sample sizes that could yield somewhat reliable results (e.g., 25 to 30 students) also avoid identifying as many of the “right” schools that perhaps should be identified. Small schools and small student subgroups may still be judged to have met

AYP due to falling below the minimum “n” threshold even if all of the students in the group fall well below the target.

- With minimum “n” sizes large enough to make more reliable decisions (e.g., 200 students or more), many schools would “meet” AYP, but school districts might not make AYP because student achievement data are aggregated across schools up to the district level. The small schools are still judged to have met AYP due to falling below the minimum “n” for reliability. The district, on the other hand, is responsible for all students in all the schools, resulting in a number above the large minimum “n” and therefore, may be identified for failing to meet AYP even though the schools those students attend were deemed to have met AYP.
- Excluding rural and small schools due to a high minimum “n” shifted the accountability burden to large schools.
- Allowing schools to avoid serving their subgroups simply because those subgroups are relatively small is inconsistent with the intent and goals of the NCLB Act.

Statistically-Based Approaches

Using statistically-based approaches, such as confidence intervals, is based on the idea that modeling and considering sampling error can help us understand the reliability and certainty of decisions made in the accountability system. Confidence intervals or z-tests recognize that the observed proportion of students scoring proficient in any one year is an estimate of that schools’ performance. Therefore, the confidence intervals describe the probability that the “true” score occurs within a range of scores rather than a precise number.

- **Decision:**
  - Should a State use a fixed minimum “n” or a statistically-based approach to maximize the reliability and minimize the negative consequences associated with making AYP decisions?
- **Consequences:**
  - If the State chooses to use a statistically-based approach, it will have to explain this to stakeholders in ways that they understand and find credible.
  - Two schools with the same proportion of students scoring proficient may have different AYP results due to larger (for a small school) or smaller (for a large school) confidence intervals.
  - States that rely on a fixed minimum “n” may under-identify for improvement small versus large schools.
- **Policy Implications:**
  - Use of confidence interval allows a State to hold all schools, large and small, accountable.
  - The use of statistically-based procedures allows State policy-makers to understand the certainty with which they are classifying/identifying schools for improvement.
  - Publicly identifying one school for improvement, while not identifying another school with similar results but a different confidence interval, will present a challenge for AYP reporting.
  - Strategies for communicating the impact of confidence intervals on AYP decisions to the public and schools will need to be developed.

- District and State officials may find the calculation of confidence intervals time consuming and complex.

Starting Points

Starting points must be set by States either by ranking schools (by grade spans) and selecting the percent proficient in a content area for the school that falls at the 20th percentile for enrollment or by using the score of the lowest performing subgroup statewide to set the starting point if it results in a higher percent proficient than the first option. (According to current data, the second option rarely occurs.)

1. Select data to be used in selecting starting points
   - **Decision:** How many years should be used as the basis for calculating starting point?
     - States may choose a single year of data from 2001-2002 as the basis for setting starting points.
     - States may use a simple or weighted average of up to three consecutive years ending with 2001-2002 to set starting points.
   - **Consequences:**
     - States that have excluded significant proportions of any subgroup in the past but are now including all students will find that using a single year to set starting points provides a more accurate picture of where schools are starting.
     - States that have a consistent data collection system that includes all students in the system every year may find that averaging across two or three years increases the reliability of the data and stabilizes information about where schools are starting.
   - **Policy Implications:**
     - Schools performing well below the starting point may find that whatever starting point is set, it is so high that the targets seem unattainable, and they will become discouraged and give up.
     - Schools well above the starting point may become apathetic, believing they have nothing to worry about and no improvement is needed.
2. Calculate starting points for all schools (by grade span)
   - **Decision:** How are starting points calculated for the entire school and all schools as the basis for making valid AYP decisions about a school?
     - States may use the "20th percentile method" or lowest performing subgroup to set starting point.
     - Using the "20th percentile method" school yields the higher starting point as required by law.
   - **Consequences:**
     - Some higher achieving schools will not have to show any positive change in status for several years.
     - The lowest performing student subgroups will have to show dramatic positive changes in status the first few years.

- **Policy Implications:**
    - Schools may find it advantageous—in effect "gaming" the system—to move students out of programs or fail to put students in programs based on their educational needs due to a greater impact on a particular student subgroup than on the total group results.

3. Calculation of starting points for content areas
    - **Decision:** How should starting point be calculated for academic content areas as the basis for making valid AYP decisions about a subgroup?
        - Single starting point for each academic content area.
    - **Consequences:**
        - Regardless of different or uniform starting points, several States have predicted, based on data simulations using previous years' student performance results, that nearly all schools in their State will be identified for improvement within 3 to 5 years, because only one student subgroup below its respective target in any one year is defined as the school making AYP.
    - **Policy Implications** for setting school-wide starting points
        - It is possible the targets may be so far out of reach that educators will get discouraged.
        - If nearly all schools are identified for improvement, it seems inevitable that the public will question the reliability of the accountability system and lose confidence in it.
        - If nearly all schools are identified for improvement, even those perceived by the public and educators as successful, the public and educators may become indifferent/apathetic to the fact that any school has been identified for improvement.

Aggregating Data

Aggregating data can be used by States to increase the sample size upon which accountability decisions about a school are based, thus increasing the reliability of those decisions. There are several methods of aggregating data that might be used:

1. Multi-year averaging
    - **Decisions:** How can multi-year averages (weighted or simple), using multiple years of data, be useful in establishing starting points and establishing annual status measurements?
        - Simply average the percent proficient from each unit across the two or three years, if the State believes that each year is an unbiased estimate of a school's performance.
        - Weight the average by the number of students enrolled in the given school, if the State is considering aggregate performance across two years as a "single" estimate of a school's performance.
    - **Consequences:**
        - Using multiple years of data to establish baseline allows the use of larger samples, thus reducing sampling error and, therefore, improving the reliability of status estimates.
    - **Policy Implications:**
        - More schools and subgroups will meet a minimum "n" threshold and be accountable for the progress of their own schools and student subgroups.

2. Rolling averages
   - **Decision:** How can rolling averages be used to determine AYP using an improvement-based approach ("safe harbor") for each school or district each year?
     - Compare, for example, the average scores of 2000, 2001, and 2002 with the averages of 2001, 2002, and 2003.
   - **Consequences:**
     - Rolling averages of 3 years simply serves to compare the first year to the last year and, in spite of the intuitive appeal of rolling averages, do little to truly improve the reliability of the comparison.
   - **Policy Implications:**
     - Apparent year-to-year change, or lack of change, in a school's performance is an appearance only, giving inaccurate impressions to the public about the school.
3. Successive multi-year comparisons
   - **Decision:** How can successive multi-year comparisons be used by those who want to evaluate trends and also want/need to aggregate data?
     - Compare, for example, the average of 2000 and 2001 with the average of 2002 and 2003.
   - **Consequences:**
     - This would provide a State the real benefit of stabilizing the comparisons without the problems associated with simple rolling averages.
   - **Policy Implications:**
     - States will need sufficient data in the early years of the 2001 ESEA Reauthorization to use this approach so they can meet the requirements of making AYP decisions about schools each year.

Intermediate Goals

1. Raising the status bar every year on a trajectory to reach all students (100%) proficient by 2013-2014
   - **Decision:** How can intermediate goals be set?
     - Divide the difference between the starting point and 100% by 12 years to arrive at the increase in the status bar required each year.
   - **Consequences:**
     - This approach assumes that school improvement is a perfectly linear process, while research has clearly documented that it is not.
   - **Policy Implications:**
     - Raising the bar annually will help keep the focus on regular yearly goals.
2. Holding the status bar steady for up to three years at a time before raising it to a new level on a trajectory to reach 100% proficient by 2013-2014
   - **Decision:** How can intermediate goals be set?
     - Raise the bar intermittently, but at least every three years with the first increase required by the 2003-2004 school year.

- **Consequences:**
  - For States choosing to focus on an improvement approach, maintaining a steady target for three years would allow the State to "feature" improvement rather than focusing on status.
  - Knowing that a substantial status increase is looming in the third year of the intermediate goal could cause school and district personnel to focus on the status bar, rather than on improvement.
  - By chance alone, several schools will "bounce" over the bar in one of the two or three years.
- **Policy Implications:**
  - It may be easier to communicate three or four changes in the bar over twelve years by keeping the bar steady for three years at a time.

Conclusions and Remaining Questions

There are several dilemmas faced by most State Educational Agencies attempting to implement the new NCLB Act accountability requirements without losing public confidence in educational accountability that has been gained over the last several years at substantial financial and political expense. Those States that implemented State and federal accountability requirements in seamless systems under the 1994 ESEA Reauthorization may well find themselves in the greatest dilemma, having only recently gained public trust in the results of building and implementing these accountability efforts (with both State and federal resources). What will be the consequences for substantially altering current systems to accommodate the considerably more prescriptive aspects of the NCLB Act?

Further, the NCLB Act challenges States to make multiple accountability decisions (as many as 37 as in the example illustrated in Chapter 1 for each school in each grade span) about schools and districts. Yet, the validity and reliability of these decisions can clearly be compromised by the multiple ways (different student subgroups each year, separately for reading or language arts and mathematics, separately for participation in the assessments, and separately for other academic indicators) in which a school or district is evaluated and, potentially, identified for improvement under the law. A school can be identified for failure to meet the target in any one of these categories. There is error associated with each decision. The error in the accountability system is therefore multiplied by the number of decisions made within it.

While a confidence interval approach may provide a sound methodology for making a statistically reliable decision, as called for in the law, it presents some communication challenges for States, districts, and schools, as noted earlier. Further, the compounding of error due to multiple decisions magnifies the reliability problem. For example, assume a State/school has a starting point of 40% proficient and so has an annual measurable objective increase of 5%. According to a simplified procedure using Table 4 (see Chapter 3) from this paper for confidence intervals, a school size of 25 would need to allow for a 10% error rate. In this situation, an observed performance of 30% proficient could not be rejected; the error roughly is equal to about two years' of increase (8-10%). If the school has 25 total students assessed at a given grade level and a subgroup has 10 students, an error rate of 13-16% would be allowed for the subgroup confidence interval. Therefore, the subgroup would pass if the goal were 40%, with an observed Percent Proficient of approximately 27%, effectively lagging the goal by two to three years. However, while using a statistically-based approach illuminates these issues, using a minimum “n” does not eliminate this concern—it simply hides it from easy examination.

Finally, the conjunctive nature of the NCLB Act prescriptive accountability model further compounds the reliability dilemma. As shown in this paper, there is error around each of the accountability decisions to be made in a conjunctive model. As the number of decisions increases so does the probability of compounding the overall error, thus impacting the resultant accountability decisions. As prescribed by the law, State analysts must determine what the outcome would be for at least 37 decisions that contribute to AYP determinations. For independent (which these are not) decisions at the .05 alpha level (95% confidence interval), nine conjunctive decisions (comparisons) result in an actual alpha of approximately .4 to .5. In other words, this means that a school will be identified for improvement on chance alone.

Given these realities for decision-makers and the seriousness of sanctions imposed under the NCLB Act, how can States best implement the NCLB Act? This paper has discussed the options available under a strict interpretation of the law and, in a very limited way, the “flexibility options” that have been signaled so far through letters released by the U. S. Department of Education.

Several areas of flexibility have potential to mitigate a few of the validity and reliability concerns discussed in this paper. First, states and districts could use two consecutive years below the target in the same student subgroup before school identification for improvement occurs. This will stabilize the reliability of identification based on subgroup and provide a sustainable basis for school improvement efforts. Second, States can decide to apply a method of matching the scope of the problem with AYP in a school (e.g. as determined by high vs. low number of groups not meeting the AYP targets and indicators). In effect, States could prioritize their efforts for improvement and sanctions based on the severity and breadth of problems with the AYP indicators. This approach would allow States and districts to target on the specific problems of student performance for specific groups of students while supporting the successful aspects of the school program.

Further Issues for Analysis

There are many additional issues that, as they are considered and implemented, have significant consequences for the validity and reliability of accountability system based decisions. Some of these issues require flexibility in interpreting the intent of the NCLB Act (through regulations or non-statutory guidance) not yet signaled by the DE. These issues require further research and investigation and are appropriate topics for future papers in this series:

- Would allowing for growth-based accountability systems (other than the specific provisions for “safe harbor”) improve the validity and reliability of the decisions made about schools?
- What affect does the identification for improvement of a school based on a different student subgroup each year due to fluctuations of scores above and below the status or due to fluctuations in the “n” in subgroups have on the validity of decisions and confidence of the public in accountability identification, and on the ability of schools to sustain school improvement efforts?
- What is the effect on reliability and validity of using a conjunctive decision-making model as required under the NCLB Act compared with a compensatory system?

- What is the impact on the validity and reliability of accountability decisions when a high percentage of individual students are represented in multiple cells (duplication), in effect weighting the decision based on a small group of students?
- What are adequate levels of validity and reliability given the seriousness of the sanctions assigned to schools under the NCLB Act?
- How might qualitative reviews of small schools be used to augment decisions when group size limits the reliability of AYP decisions?
- How do we focus on the schools that everyone (consensus) agrees are really in need of improvement, given the limitations on capacity of SEAs and school districts? What process of determining true levels of need for technical assistance would be acceptable to deliver technical assistance based on some form of tiered levels? When schools are identified one year because they are low in mathematics and the next because they are low in reading or language arts or one year because their African-American student subgroup is low and the next year because their LEP student subgroup is low, but in no two consecutive years are the same subgroup or the same content area low, how would their level of need be determined?
- What is the difference between a school identified for improvement due to the failure of several subgroups that account for a small percentage of the students (due to multiple subgroup memberships of a few students) versus a school identified due to the failure of a single subgroup that accounts for a large percentage of the school? How does this difference affect the appropriate level of assistance and sanction?
- How will the technical issues related to transitions and additions of assessments required by the NCLB Act affect the validity and reliability of AYP decisions? What design changes and transitional steps will be required as tests change, new tests come on line, and grades and academic content areas not previously assessed are added to the AYP decisions?
- What procedures should be established to allow schools identified for improvement to review the data and to present evidence if they believe that the identification is in error, as required by the NCLB Act?
- What is the probability of correctly classifying a school identified for improvement?
- How do we clearly define LEP students and SWDs for purposes of AYP and as a basis for future research on actual improvement rates for LEP students? What is the impact of different LEP students and SWDs goals on results? What do we know from research on the improvement rates of LEP students and SWDs in States and districts? How accurate is it if successes do not continue to be counted on annual accountability determinations because they are exited from these programs?
- What affect would the use of statistically-base approaches (e.g., confidence intervals) have on public confidence in the accountability system, particularly given the reality that confidence intervals would likely be different for different subgroups, some schools with higher scores would fail to meet AYP while schools with lower scores do not?
- What are the unintended consequences of the high stakes sanctions under the NCLB Act to student achievement that are likely to result from lower expectations under revised academic content and student achievement standards?

- What are the unintended consequences of the high stakes sanctions under the NCLB Act to student achievement that are a likely result of lower expectations for depth, breadth, and challenge in assessment design?

The importance of building accountability systems from which valid and reliable decisions about schools can be made is essential to the credibility of efforts to hold schools accountable for student learning. Because identification under a strict interpretation of the NCLB Act occurs when any student subgroup or academic content area falls below the required trajectory rather than as the result of a pattern of or trend in low performance by a subgroup or in a content area, schools may be identified for improvement based on a random occurrence rather than a reliable result. The goal of a valid and reliable accountability system is to separate successful schools from those in need of assistance. A more flexible interpretation of the adequate yearly progress provisions of the NCLB Act may actually increase the likely of creating valid and reliable systems.

Ensuring the validity of accountability decisions requires an extended analysis of comparable evidence and consideration to competing interpretations of the meaning of the results. While validity demands what may appear to be a more complicated system than policy makers or the public might find readily accessible to immediate understanding, serious consequences of misidentifying schools for improvement will have far-reaching repercussions. The misidentification of schools or districts for improvement or, conversely, worthy of reward leads to an inevitable diffusion of limited resources, confusion over what programs are working, and a loss of public confidence in the public schools and in our ability to hold them properly accountable for student learning. It is of utmost importance that States carefully and deliberately approach the decisions that will be necessary to ensure that, in the final analysis, they develop a valid and reliable accountability system—one that will engender confidence in the decisions whenever schools or districts are identified for improvement under the NCLB Act.

Appendix

A

Excerpts from The *No Child Left Behind Act* of 2001

This selection of brief excerpts from the 2001 ESEA reauthorization is intended to support the search for answers to questions regarding Adequate Yearly Progress and State accountability determinations.

Among the 2001 ESEA Reauthorization requirements, States must make important decisions regarding the:

1. *Development of a single statewide accountability system (conjunctive model) to ensure that (a) at least 95% of enrolled students are assessed as required, (b) all student groups reach the State's proficient level of academic achievement in all required subjects by the end of the 2013-14 school year, and (c) objectives for increasing English proficiency are met;*
2. *Identification for improvement of schools and districts receiving grants under Title I that do not meet the State's adequate yearly progress requirements; and*
3. *Determination of what constitutes "statistically reliable information" when disaggregating student performance data for accountability purposes.*

As noted previously, the purpose of this paper is to address a particular provision of the 2001 Reauthorized Elementary and Secondary Education Act (ESEA). This concerns how States will define the Adequate Yearly Progress of schools and local educational agencies (school districts) consistent with section 1111(b)(2)(C)(v)(II) of the 2001 ESEA Reauthorization [also known as the *No Child Left Behind* Act of 2001 (NCLB)]." This provision is:

> …except that disaggregation of data under subclause II shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student; ….

Further guidance is provided in regard to this matter in recently published final Federal regulations.[24] The regulations make it clear that each State must determine and justify the minimum number of students sufficient to yield statistically reliable information. Specifically:

> **§200.7 Disaggregation of data.**
>
> "(a) Statistically reliable information. (1) A State may not use disaggregated data for one or more subgroups under §200.2(b)(10) to report achievement results under section 1111(h) of the Act (report cards) or to identify schools in need of improvement, corrective action, or restructuring under section 1116 of the Act **if the number of students in those subgroups is insufficient to yield statistically reliable information** [emphasis added].
>
> "(2) **Based on sound statistical methodology, a State must determine and justify in its State plan the minimum number of students sufficient to yield statistically reliable information for each purpose for which disaggregated data are used** [emphasis added].
>
> "(b) Personally identifiable information. (1) A State may not use disaggregated data for one or more subgroups under §200.2(b)(10) to report achievement results under section 1111(h) of the Act [report cards] if the results would reveal personally identifiable information about an individual student.
>
> "(2) To determine whether disaggregated results would reveal personally identifiable information about an individual student, a State must apply the requirements under section 444(b) of the General Education Provisions Act (the Family Educational Rights and Privacy Act of 1974).
>
> "(3) Nothing in paragraph (b)(1) or (b)(2) of this section shall be construed to abrogate the responsibility of States to implement the requirements of section 1116(a) of the Act for determining whether States, LEAs, and schools are making adequate yearly progress on the basis of the performance of each group listed under section 1111(b)(2)(C)(v) of the Act.
>
> "(4) Each State shall include in its State plan, and each State and LEA shall implement, appropriate strategies to protect the privacy of individual students in reporting achievement results under section 1111(h) of the Act and in determining whether schools and LEAs are making adequate yearly progress on the basis of disaggregated subgroups."

Other Related Accountability Requirements

Section 1111(b)(2)(A): "Each State plan shall demonstrate that the State has developed and is implementing a single, statewide State accountability system…. Each State accountability system shall—

> "(ii) be the same accountability system the State uses for all public elementary schools and secondary schools or all local educational agencies in the State, **except that** [emphasis added] public elementary schools, secondary schools, and local educational agencies not participating under this part are not subject to the requirements of section 1116 [school improvement].…"

[24] U.S. Department of Education (2002, July 5). Title I—Improving the Academic Achievement of the Disadvantaged; Final Regulations (34 CFR Part 200). Washington, DC.

Sec. 1111(b)(2)(A)(i): Accountability must be based on academic standards and academic assessments adopted under (1) and (3), etc. (without mention of subject areas).

Sec. 1111(b)(2)(B): "Each State plan shall demonstrate, based on academic assessments described in paragraph (3), and in accordance with this paragraph, what constitutes adequate yearly progress of the State, and all public elementary schools, secondary schools, and local educational agencies in the State, toward enabling all students to meet the State's student academic achievement standards, while working toward the goal of narrowing the achievement gaps in the State, local educational agencies, and schools."

Sec. 1111(b)(2)(C): "'Adequate yearly progress' shall be defined by the State in a manner that—

> "(v) includes separate measurable annual objectives for continuous and substantial improvement for each of the following:
>
> "(I) The achievement of all public elementary school and secondary school students.
>
> "(II) The achievement of—
>
> "(aa) economically disadvantaged students;
>
> "(bb) students from major racial and ethnic groups;
>
> "(cc) students with disabilities; and
>
> "(dd) students with limited English proficiency; ….
>
> "(vi) in accordance with subparagraph (D), includes graduation rates for public secondary school students and at least 1 other academic indicator, as determined by the State for all public elementary school students [and may include other academic indicators at the State's discretion but must be measured separately for each student sub-group];…."

Sec. 1111(b)(2)(D)—

> "(ii) except as provided in subparagraph (I)(I) [a limited exemption], may not use those indicators to reduce the number of, or change, the schools that would otherwise be subject to school improvement, corrective action, or restructuring under section 1116 if those additional indicators were not used, but may use them to identify additional schools for school improvement or in need of corrective action or restructuring."

Sec. 1111(b)(2)(G): Each State shall establish statewide annual measurable objectives, pursuant to subparagraph (C)(v) for meeting the requirements of this paragraph and which—

> "(i) shall be set separately for the assessments of mathematics and reading or language arts under subsection (a)(3)….

Sec. 1111(b)(2)(I): ANNUAL IMPROVEMENT FOR SCHOOLS.—Each year, for a school to make adequate yearly progress under this paragraph—

> "(i) each group of students described in subparagraph (C)(v) must meet or exceed the objectives set by the State under subparagraph (G) except that….
>
> "(ii) not less than 95 percent of each group of students described in subparagraph (C)(v) who are enrolled in the school are required to take the assessments, consistent with paragraph (3)(C)(xi) and with accommodations, guidelines, and alternative assessments….

Accountability Requirement for English Proficiency

(Subpart 2—Accountability and Administration, Part A, Title III):

Sec. 3121(a): "Each eligible entity that receives a subgrant from a State educational agency under subpart 1 shall provide such agency, at the conclusion of every second fiscal year during which the subgrant is received with an evaluation…that includes—

"(3) the number and percentage of children in the program and activities attaining English proficiency by the end of each school year, as determined by a valid and reliable assessment of English proficiency; and

"(4) a description of the progress made by children in meeting challenging State academic content and student achievement standards for each of the 2 years after such children are not longer receiving services under this part.

Sec. 3122(a) ACHIEVEMENT OBJECTIVES AND ACCOUNTABILITY:

"(1) IN GENERAL.—Each State educational agency or specially qualified agency receiving a grant under subpart 1 shall develop annual measurable achievement objectives for limited English proficient children served under this part that relate to such children's development and attainment of English proficiency while meeting challenging State academic content and student academic achievement standards as required by section 1111(b)(1).

"(3) CONTENTS.—Such annual measurable achievement objectives—

"(A) shall include—

"(i) at a minimum, annual increases in the number or percentage of children making progress in learning English;

"(ii) at a minimum, annual increases in the number or percentage of children attaining English proficiency by the end of each school year, as determined by a valid and reliable assessment of English proficiency consistent with section 1111(b)(7); and

"(iii) making adequate yearly progress for limited English proficient children as described in section 1111(b)(2)(B)….

Sec. 3122(b) ACCOUNTABILITY.—

"(1) FOR STATES.—Each State educational agency receiving a grant under subpart 1 shall hold eligible entities receiving a subgrant under such subpart accountable for meeting the annual measurable objectives under subsection (a) including making adequate annual yearly progress for limited English proficient children.

Appendix

B

The NCLB Act "Safe Harbor" and "Opportunity to Review" Provisions in Final Adequate Yearly Progress Determinations

Background

The *No Child Left Behind* Act (NCLB) requires States to develop a single, statewide accountability system that will identify schools and districts[25] "for improvement" if the percentage of their students scoring at or above the "proficient" level on annual assessments in reading or language arts and mathematics is unacceptable or students do not meet the minimum requirements for participation in the assessments or the other academic indicators prescribed in the law. Under the NCLB Act, schools and districts are identified for improvement if (1) the percentage of students scoring at or above "proficient" is lower than required in any year, regardless of the students' (at the school or district level) performance in prior years, (2) the student participation rate on the State assessments is lower than required (95%), or (3) students do not meet an "other academic indicator" requirement (fail to make progress). This is a "status" accountability design in which accountability decisions are based solely on whether each AYP targets is met or not. (See the subsection, "Accountability requirements Under the NCLB Act—Shifting Emphases and New Challenges," in Chapter 1 for additional information.)

However, the NCLB Act also includes two provisions for further review of the student performance results prior to a final determination of failure to make AYP and identification for improvement. The first of these is the "safe harbor" provision and the second is the opportunity for a school or district to present additional evidence when it believes that the proposed identification is in error.

[25] The NCLB Act provisions addressed here are also applicable to requirements pertaining to the identification of local educational agencies (school districts) for improvement by State Educational Agencies [see Section 1116(c)(1)]. The final regulations on accountability (December 2002) extended "safe harbor" provisions to LEAs (§ 200.20).

“Safe Harbor”

Safe harbor” provisions apply to reviewing student performance on State assessments in mathematics and reading or language arts at the school building level as well as at the district level. If the performance of one or more student subgroups on one or both of these assessments fails to meet AYP targets, then “safe harbor” provisions can be applied to further review the performance of the student subgroup(s) in question provided that the subgroup(s) met the participation rate requirement and the State’s other grade level academic indicator progress requirement. If either of these latter two requirements were not met, the school or district cannot benefit from a “safe harbor” review. Section 1111(b)(2)((I) provides, “ANNUAL IMPROVEMENT FOR SCHOOLS.—Each year, for a school to make adequate yearly progress under this paragraph—

(i) each group of students described in subparagraph (C)(v) must meet or exceed the objectives set by the State under subparagraph (G) [measurable objectives for the assessments of mathematics and reading or language arts], except that if any group described in subparagraph (C)(v) does not meet those objectives in any particular year, the school shall be considered to have made adequate yearly progress if the percentage of students in that group who did not meet or exceed the proficient level of academic achievement on the State assessments under paragraph (3) for that year decreased by 10 percent of that percentage from the preceding school year and that group made progress[26] on one or more of the academic indicators described in subparagraph (C)(vi) or (vii).

The final regulations on Title I accountability requirements (December 5, 2002) clarified in §200.20(d) that the law’s “safe harbor” provisions extend to school districts (referred to as LEAs in the law). The law appears to be quite clear on the application of a “safe harbor” review. It is important to remember that, consistent with the NCLB Act and December 5, 2002, final regulations on accountability, the AYP determinations must be applied as set forth under Section 1111(b)(2)(I). The first level of review is “status”—did the school or district meet all of the AYP targets (at least 37 measures)? The second is “improvement”—did the school meet the “safe harbor” provisions for the student subgroup(s) in question? It is only necessary to apply the “safe harbor” measures when a school or district fails to make AYP requirements in one or both of the required subject area assessments. The final level is the opportunity for schools and districts to request a review and present additional evidence after being identified for improvement.

Following are three examples of applying the “safe harbor” test at the school building level:

1. Assume same starting point for all subgroups in School District Z—40% proficient in reading at the 4th grade. In this case, assume the LEP group is 10% proficient (90% of the group are below proficient). The annual measurable objective for the next year for all subgroups is 5% (60% below proficient divided by 12). Assume the LEP group in School A increases 8%. The school has missed its AYP target for this group because (1) the % proficient did not increase to a total of 45% (needed to make up the difference between 10% and 40% plus “grow” by 5%) and did not make “safe harbor” as a subgroup because

[26] States must define “progress.” If a State sets escalating annual targets, then the “safe harbor” subgroup must meet the target for the current year. If a State defines “progress” as a minimum value (e.g., attendance must be 90% or higher), the subgroup must meet that value. Whatever the target is for ALL students on a given academic indicator, the “safe harbor” subgroup must meet or exceed it as a condition of applying that provision under the law.

they would have had to have had at least a 9% increase from 10% to 19% to meet the 10% test.

2. Now assume that it is the following school year and there is a new cohort of 4th grade LEP students. Assume that 26.5% of this group scores at the proficient or higher level on the reading test (an increase of 8% over the previous 4th grade LEP students). By now, the district's measurable objective for 4th grade reading has increased to 50% proficient or better. So, this newest group of 4th grade LEP students would still have failed to meet the AYP target. However, they would meet a "safe harbor" test because of their 8.5% gain. Since the increase from 18% to 26.5% over the previous year is more than 10% of the 82% difference between those proficient the previous year and 100%.

"Opportunity to Review"

If a school or district misses an AYP target, the NCLB Act also includes provisions permitting an opportunity to present additional evidence if either has reason to believe that the identification has been made in error. The relevant provisions are:

1. Section 1116(b)(2): "OPPORTUNITY TO REVIEW AND PRESENT EVIDENCE; TIME LIMIT.—

   (A) IDENTIFICATION.—Before identifying an elementary school or a secondary school for school improvement under paragraphs (1) or (5)(A), for corrective action under paragraph (7), or for restructuring under paragraph (8), the local educational agency shall provide the school with an opportunity to review the school-level data including academic assessment data, on which the proposed identification is based.

   (B) EVIDENCE.—If the principal of a school proposed for identification under paragraph (1), (5)(A), (7), or (8) believes, or a majority of the parents of the students enrolled in such school believe, that the proposed identification is in error for statistical or other substantive reasons, the principal may provide supporting evidence to the local educational, agency which shall consider that evidence before making a final determination.

   (C) FINAL DETERMINATION.—Not later than 30 days after a local educational agency provides the school with an opportunity to review such school-level data, the local educational agency shall make public a final determination on the status of the school with respect to the identification.

2. Section 1116(c)(5): "OPPORTUNITY TO REVIEW AND PRESENT EVIDENCE.—

   (A) IDENTIFICATION.—Before identifying a local educational agency for improvement under paragraph (3) or for corrective action under paragraph (10), a State educational agency shall provide the local educational agency with an opportunity to review the data including academic assessment data, on which the proposed identification is based.

   (B) EVIDENCE.—If the local educational agency believes proposed for identification under paragraph (1), (5)(A), (7), or (8) believes that the proposed identification is in error for statistical or other substantive reasons, the agency may provide supporting evidence to the State educational agency, which shall consider that evidence before making a final determination not later than 30 days after the State educational agency provides the local educational agency with the opportunity to review such data under subparagraph (A).

This review process provides “one last check” upon which to consider a determination that a school or district may be identified for improvement. Although the law addresses a process that may occur after a school or district has been identified for improvement (failing to meet an AYP target for two consecutive years), the process also seems lend itself to a review after the first year in which a school or district fails to make an AYP target.

It is important to note that the evidence necessary is founded on “error for statistical or other substantive reasons.” Also, it seems to follow that the evidence cannot be based on the results of academic assessments beyond those required in Section 1111(b)(3) or other academic indicators the State may require since those may only be used to identify schools or districts for improvement [Section 1111(b)(2)(D)(ii)].

To implement the above provisions related to the “opportunity to review,” States may want to consider especially:

- Assurance that review determinations will be completed within 30 days after the LEA provides schools an opportunity to review the assessment results and within 30 days after the State provides LEAs with the results.
- Assurance that local and State reviews will be based solely on “error for statistical or other substantive reasons” including what evidence or arguments would constitute “other substantive reasons.”
- Criteria, procedures, and timelines the SEA will establish for its review process.
- Criteria, procedures, and timelines for the LEA review process.
    - Will each LEA establish its own process?
    - Will the SEA establish a uniform process for all LEAs?
    - How would this be done?
    - Would a school be able to “appeal” to the SEA an LEA’s determination if the school were not satisfied with that determination?
- How will the SEA “track” the number of local reviews annually, the nature and basis for the review requests, and the results of these reviews?

Appendix

C

Critical Questions and Related Major Considerations in Building State Accountability Systems—An Excerpt

Excerpted from: Incorporating Multiple Measures of Student Performance into State Accountability Systems—A Compendium of Resources

Introduction

The questions which follow have been excerpted from Chapter 2 of a recent publication by the Council of Chief State School Officers (Erpenbach et al., 2002). The chapter was authored by Paul M. LaMarca of the Nevada Department of Education. In the full chapter, each of the questions and underlying considerations is followed by a discussion intended to provide additional background for readers.

The purpose of these questions is to provide a beginning point for State teams responsible for building (and refining) statewide systems of educational accountability for the improvement of teaching and student achievement. They are also intended to help guide the thinking and to focus the discussions of policymakers and planners by raising issues and related major considerations for which decisions are likely to be needed in order to accomplish this challenging and important task. State teams will undoubtedly approach these questions and the major considerations in widely varying, frequently iterative, patterns.

The questions are not presented in any particular order of importance. It is believed, however, that these are the "critical" questions that States should address first as they build and implement educational accountability systems. The questions are also not intended to represent an exhaustive listing. There will be other questions that any one State may ask in the process of defining and building its system. The way a State answers one question will affect the answers to other questions, and States may have to revisit questions as they work on and with their accountability systems. The "major considerations" are intended to highlight important issues and variables that will likely have significant impact on the resultant accountability system. Like the critical questions, the related major considerations posed are not intended to be exhaustive in their scope nor should any weight be given to the order in which they are presented.

Critical Questions and Related Major Considerations

1. Legislative and Policy Influences – What legislative and policy initiatives will determine how the accountability system is designed and implemented?

Consider:

- State and federal legislative initiatives
- State Board and Department accountability policies and oversight responsibilities
- Role of accountability in State Board and Department policy
- Purposes of laws and policies in the broader context of the state's educational system, initiators of accountability strategies, influence of public and policy makers, historical influences, etc.
- Relationship of the accountability system to variables the school can influence

2. Purpose – What are the purposes of the accountability system?

Consider:

- Classifications of schools and LEA
- Effects on school/LEA improvement plans, including staff orientation and development
- Rewards and sanctions such as those applied to teachers, schools, students, and districts (for example, see *Quality Counts 2001*, January 11, 2001, pp. 82-84)
- Assistance provided to schools and LEAs such as funds for remedial instruction and technical assistance from SEAs, professional organizations, intermediate education agencies, and consortia of various forms
- Validity considerations, including how the state will define the concept of a quality or improving school

3. Accountability Model – What characteristics should the accountability model have?

Consider:

- Defining and constructing the model according to purpose and use
- Index vs. profile
- Status, longitudinal, or successive groups
- Compensatory or conjunctive
- Targets for schools — interim and long-term, annual improvement requirements, for school as a whole, grade levels, and/or groups of students
- Applying weights to measures and indicators
- Special procedures or considerations for small schools

4. Multiple Measures and Indicators – What multiple measures of academic content and non-cognitive indicators (e.g., attendance) will be used in the system and why?

Consider:

- Logic and method for determining how the measures reflect the construct of a good or improving school
- Criteria for selecting measures and indicators
- Role of academic content and performance standards in selecting/developing multiple measures for accountability
- Characteristics of multiple measures and indicators (e.g., types of assessment items, when tests are administered, definition of non-cognitive indicators, matrix sampling vs. census testing)
- Relationship between multiple measures/indicators and the concept of a how well a school is performing

5. Non-Standard Measures and Exemptions – How will non-standard measures be used?

Consider:

- Accommodated or modified assessments, criteria for eligibility
- Alternate assessments, criteria for eligibility
- Exemptions, criteria

6. Combining Data – How will data from multiple measures and indicators be combined to categorize schools?

Consider:

- Validity and reliability considerations for student-level data
- Combining categorical vs. continuous data
- Combining rules and justification
- Setting cut scores on individual tests
- Accounting for students not tested
- Using results from non-standard assessments
- Developing criteria to determine school and LEA categorization

7. Technical Issues – What technical issues and additional analyses will need to be explored in order to develop and evaluate the system?

Consider:

- Reliability of individual assessment results, combined results at the student level, decision consistency
- Reliability of accountability results, decision consistency
- Evaluation of using differential weights for indicators
- Analysis of impact on subgroups (including, but not limited to, situations in which subgroup improvement is a component of school improvement criteria)

- Simulation studies to evaluate potential impact
- Comparisons of different models

8. Reporting – How are accountability results reported?

Consider:

- Audiences
- Formats
- Levels of aggregation
- Keeping reports understandable and useful
- Disaggregated results
- Minimizing misinterpretation and misuse
- Additional information, other than the results used specifically for accountability decisions, to provide to schools and LEAs

9. Impact – What is the potential impact of the accountability system?

Consider:

- Number of schools and LEAs identified as needing improvement
- Intended effects on education (e.g., instruction focused on standards; efficient use of funds for staff development; better-targeted school improvement plans) and potential unintended effects (e.g., instruction focused on only the skills that are tested especially if not all standards are assessed, increased drop-out rates)
- Finances, resources, instructional time, etc.
- Strategies to monitor impact and effects and address potential unintended negative consequences

10. Evaluating and Validating the System – How will the system design incorporate the need for revisions over time?

Consider:

- Monitoring impact for fine-tuning the system
- Adding or deleting measures and indicators
- Ongoing evaluation of impact, effects on instruction, curriculum, professional development, teacher recruitment and retention, etc.
- Revising measures and indicators
- Policy and legislative changes

Reference

Erpenbach, W. J., Carlson, D., LaMarca, P. M., & Winter, P. W. (2002). *Incorporating multiple measures of student performance into state accountability systems; A compendium of resources.* Washington, DC: Council of Chief State School Officers.

Appendix D

Glossary

The following Glossary was adapted from *Critical Issues in Large Scale Assessment: A Resource Guide* (2002) by Doris Redfield and is also available from CCSSO. The definitions provided are applicable to their use in this paper.

Accommodations: Changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To be appropriate, assessment accommodations must be accommodations that are also used in instruction and they must not change the construct intended to be measured by the assessment or the meaning of the resulting scores (for extended discussion, see Redfield, *Critical Issues*, 2002).

Accountability: The systematic use of assessment data and other information to assure to those inside and outside of the educational system that schools are moving in desired directions. Commonly included elements are goals, indicators of progress toward meeting those goals, analysis of data, reporting procedures, and consequences or sanctions. Accountability often includes the use of assessment results and other data to determine program effectiveness and to make decisions about resources, rewards, and consequences [see Redfield, *Critical Issues* (2002) for an extended discussion].

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

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

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

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

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

Assessment System (i.e., an aligned assessment system): A series of assessments of student performance at different grade levels which are based on publicly adopted standards of what is to be taught coupled with high expectations of student mastery. A standards-based assessment system is designed to hold schools publicly accountable for each student meeting those high standards.

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

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

Breadth: Refers to the comprehensiveness of the content and skills embodied in the standards, curriculum, and assessments.

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

Consequential Validity Evidence: Data that illuminates the extent to which the assessment has the desired effects (e.g., on students, teachers, administrators, the curriculum, instruction and/or other entities0.

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

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

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

Content Validity Evidence: Data that illuminate the extent to which

(1) The knowledge, skills, and cognitive demands of the learning objectives underlying an assessment are accurately reflected in the assessment; and

(2) The assessment adequately covers the **domain** of knowledge, skills, and cognitive demands represented in the learning objectives.

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

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

Curriculum: Refers to what is taught.

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

Defensibility: Refers to the technical properties of an assessment that makes its use for a particular purpose just. Such properties include validity, reliability, fairness, and lack of bias.

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

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

Face Validity Evidence: Tests that are face valid look like they measure what they purport to measure. For example, a "writing" test that relies solely upon multiple-choice questions about the conventions of writing such as grammar, punctuation, and spelling, is lacking in face validity

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

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

High-Stakes Sanctions: Sanctions that have important, direct, or lasting consequences for programs or institutions.

Instruction: Refers to the teaching methods used to deliver the curriculum to students.

Inter-Component Alignment of the Accountability System: Refers to consistency among the components of an accountability system, such as purposes of an accountability system, the selection of indicators it includes, and the decision rules for identifying schools for various kinds of intervention or sanctions.

Intra-Component Integrity of the Accountability System: Refers to the characteristics of each element or indicator in the accountability system. To achieve intra-component integrity, each element needs to have a proper definition of each component, alignment between design, and implementation of each part of the system, and degree of error.

Large-Scale Assessments: Those assessments that are administered to relatively large numbers of students. State testing programs and local school district testing programs are examples. Large-scale programs are in contrast to tests and other assessments administered on a smaller scale (e.g., by classroom teachers for instructional purposes).

Laws: Refers to legislative mandates. Violations carry negative legal consequences.

Longitudinal Methods: In "true longitudinal" methods, the focus in on the difference between different measurements of the same students—collected at two or more points in time. "Quasi-longitudinal" approaches, on the other hand, focus on the difference between the results for all the third-grade students one year and all the fourth-grade results the next year, for example—only some of whom were in that school for both assessments. The results for the two methods have been found to be quite similar for most schools; however—as might be expected—they can be different for schools with high student mobility.

Matrix Sampling: A measurement technique whereby a large set of test items is organized into a number of relatively short items sets. Each subset is then administered to a sub-sample of test takers, thereby avoiding the need to administer all items to all examinees (e.g., for program evaluation purposes).

Norm-Referenced Test Interpretations: Score interpretations based on a comparison of a test taker's performance to the performance of other people in a specified **reference population**.

Performance Standards: Standards that specify how well students must perform in order to meet certain levels of proficiency. Performance standards consist of four components:

(1) Performance levels which provide descriptive labels for student performance (e.g., advanced, proficient);

(2) Descriptions of what students at each performance level must demonstrate relative to the test;

(3) Examples of student work that illustrate the range of performance for each performance level; and

(4) Cut scores which separate one level of performance from another.

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

Reliability of Accountability Systems: Refers to the degree to which the data from indicators used in an accountability system are consistent over repeated applications of the decision rules and hence are dependable, and repeatable; the degree to which data from indicators are free of sampling error. Because each indicator has a degree of error, the accountability system's degree of error is a function of the combined error of the indicators.

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

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

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

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

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

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

Standard Assessment: Refers to the administration of an assessment in the prescribed, standard way, without the use of **accommodations** or **modifications.**

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

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

Technically Sound Accountability Systems: Systems that are defensible, reliable, and valid for the purposes for which they are used, fair, and unbiased.

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

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

Valid: Refers to the degree to which a test measures what it purports to measure. See **Validity**.

Validity of a Test:

(1) An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores

(2) The extent to which a test measures what its authors or users claim it measures

(3) The appropriateness of the inferences that can be made on the basis of test results

Validity of the Accountability System: An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that:

- The components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and
- The system is accomplishing what was intended (and did not accomplish what was not intended.)

Accountability Systems & Reporting

CAS
COMPREHENSIVE ASSESSMENT SYSTEMS for TITLE I

SCASS
State Collaborative on Assessment
and Student Standards