

Accountability
Systems and
Reporting

State Collaborative on Assessment and Student Standards

Statewide Educational Accountability Under NCLB

Central Issues Arising From An Examination Of
State Accountability Workbooks And
U.S. Department Of Education Reviews
Under The No Child Left Behind Act Of 2001

*THE STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS (SCASS)
ACCOUNTABILITY SYSTEMS AND REPORTING (ASR) CONSORTIUM*

William J. Erpenbach
Ellen Forte-Fast
Abigail Potts



THE 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.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Michael E. Ward (North Carolina), President

Ted Stilwill (Iowa) President-Elect

Suellen K. Reed (Indiana), Vice President

G. Thomas Houlihan, Executive Director

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

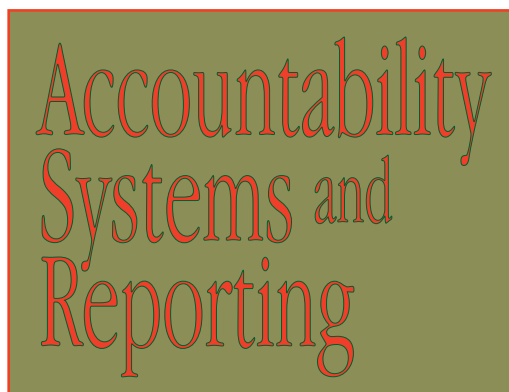
John Olson, Director of Assessments

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

Jan Sheinker, Coordinator
Comprehensive Assessment Systems for ESEA Title I SCASS

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

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



State Collaborative on Assessment and Student Standards

Statewide Educational Accountability Under NCLB

Central Issues Arising from an Examination of State Accountability Workbooks and U.S. Department of Education Reviews Under the No Child Left Behind Act of 2001

An Accountability Systems And Reporting State Collaborative On Assessment And Student Standards (ASR SCASS) Paper

July 2003

William J. Erpenbach
Ellen Forte-Fast
Abigail Potts

State Collaborative on Assessment and Student Standards
COUNCIL OF CHIEF STATE SCHOOL OFFICERS—WASHINGTON, DC

Acknowledgements

This paper resulted from the work of the Accountability Systems and Reporting State Collaborative on Assessment and Student Standards (ASR SCASS). Information included in this paper was collected from the State Consolidated Accountability Workbooks, a series of conference calls with state education agency staff, and several CCSSO meetings. The authors benefited tremendously from the feedback, comments, and reviews of the ASR and Comprehensive Assessment Systems (CAS) SCASS members and additional state education agency staff. In addition the authors would like to thank the following people for their assistance in editing and reviewing the paper:

Jan Sheinker, CAS SCASS
Arthur Halbrook, CCSSO
Frank Philip, CCSSO

This paper was supported entirely by funding from member States of the Accountability Systems and Reporting State Collaborative on Assessment and Student Standards (ASR-SCASS), through the Council of Chief State School Officers (CCSSO). Information about the ASR SCASS is available on the CCSSO web site, <http://www.ccsso.org>.

Table of Contents

| | |
|---|----|
| ACKNOWLEDGEMENTS | IV |
| TABLE OF CONTENTS | VI |
| PART I: | 1 |
| INTRODUCTION AND BACKGROUND | 1 |
| PART II..... | 5 |
| ISSUES IN STATES' ACCOUNTABILITY PLANS | 5 |
| Standards and Assessments in General | 6 |
| <i>Student Achievement Standards</i> | 7 |
| <i>Inclusion of Both Reading and Writing Assessment Results in Percent Proficient Calculation</i> | 8 |
| <i>Evolving Assessment Systems</i> | 9 |
| <i>First Administration Rule</i> | 11 |
| AYP Model..... | 13 |
| <i>AYP Indicators</i> | 14 |
| <i>Dual Accountability Systems</i> | 18 |
| <i>Strategies for (1) Protecting Confidentiality and (2) Enhancing Reliability</i> | 19 |
| Inclusion | 28 |
| <i>General Inclusion</i> | 28 |
| <i>Inclusion of Students with Disabilities</i> | 30 |
| <i>Limited-English Proficient Students</i> | 35 |
| Starting Points, Annual Measurable Objectives, and Intermediate Goals | 37 |
| <i>Starting Points, AMOs, and IGs Based on Other Than State Averages</i> | 38 |
| <i>A Novel Approach to Determining IGs</i> | 38 |
| <i>Establishing Starting Points, AMOs, and IGs in Timeline Waiver States</i> | 39 |
| Participation Rate and Other Academic Indicators..... | 39 |
| <i>Participation Rate</i> | 39 |
| <i>Graduation Rate</i> | 42 |
| <i>Other Academic Indicators</i> | 44 |
| Validity and Reliability..... | 45 |
| AYP Consequences and Reporting..... | 46 |
| PART III: CONCLUSIONS | 49 |
| Non-Negotiable Issues | 49 |
| Unanticipated Approvals | 50 |
| Approvals Not Likely to Have Long-Term Impacts on AYP Determinations | 52 |
| Approvals that May Have Long-Term Impacts on AYP Determinations | 52 |
| APPENDIX A – REFERENCES/RESOURCES | 53 |
| Others | 53 |
| APPENDIX B | 55 |

Part I:

Introduction and Background

This paper provides, in summary form, a discussion of the central issues arising from an examination of State Accountability Workbooks prepared for Peer Reviews through the U.S. Department of Education (ED) and subsequent approval decisions made by ED. These issues have their genesis in requirements set forth under the No Child Left Behind Act of 2001 (NCLB) and attendant regulations and policy. In large measure, they reflect areas where States have faced noteworthy challenges or have chosen to “push the envelope” in their development of statewide educational accountability systems. In addition, this paper focuses entirely on the Title I accountability requirements of NCLB and does not directly address the standards, assessments, program, or fiscal requirements of this law. The paper is based on information available through June 2003 and was finalized in cooperation with member States of both the Accountability Systems and Reporting (ASR) and Comprehensive Assessment Systems (CAS) State Collaboratives on Assessment and Student Standards.

When President Bush signed the No Child Left Behind Act of 2001 (NCLB)¹ into law on January 8, 2002, all 50 states, the District of Columbia, and Puerto Rico were presented with an unprecedented challenge: to implement a tightly prescribed accountability model with the goal of all students achieving grade-level proficiency in reading or language arts and mathematics within 12 years.² In the pages that follow, how States responded to this challenge—many in ways that could not have been anticipated by the legislators and policy makers whose vision this law represents—are described. Indeed, each State’s unique context meant that even the most narrowly defined accountability elements of the law would not play out in cookie-cutter fashion. Further, the process by which States’ plans took shape over the year preceding January 31, 2003, when their preliminary accountability plans were due to the U. S. Department of Education (ED), may have helped States to focus on—even to identify—the issues that were most critical to them as well as the philosophies that underlie their positions. Many States continue to refine their plans, even though “final” plans were due to ED by May 1, 2003, and ED was required to approve³ the plans within 120 days of January 31, 2003, unless a given plan clearly did not meet the NCLB requirements. At the end of June 2003, a great many States were still negotiating various aspects of their accountability designs with ED.

As it turned out, States did not have a full year in which to consider and develop their accountability plans. This reauthorization of the Elementary and Secondary Education Act (ESEA) carried the unusual provision of taking effect immediately upon signature by the President—a transition period was not authorized. In addition, although all States immediately recognized that NCLB had major ramifications for their accountability

¹ NCLB is the 2001 reauthorization of the groundbreaking 1965 Elementary and Secondary Education Act. The most recent previous reauthorization of this law was known as the Improving America’s Schools Act of 1994 (IASA).

² This was, of course, only one of the many challenges presented to States in NCLB.

³ Although many have used the term “final approval” to refer to the status of State plans, most State plans are conditionally approved (as of June 2003), meaning that any plan may still be subject to subsequent reviews and requests for additional information or modifications by ED.

systems, it was not immediately clear exactly what the specific requirements would be. In the months following enactment of NCLB as its policy positions and regulations evolved, ED issued a series of documents (including letters from Secretary Paige to Chief State School Officers) meant to clarify what was expected of States in terms of standards, assessments, and accountability and to specify how States were expected to demonstrate compliance with these requirements. Of particular interest to States were the accountability requirements. Although the requirements for standards and assessments under NCLB are indeed rigorous, they represent more an expansion of the previous requirements than they represent new territory. For most States, however, the accountability requirements would represent a new continent altogether. Further, States were faced with developing or modifying their accountability systems while ED was simultaneously developing regulations and making policy determinations, all without accompanying nonregulatory guidance. The final accountability regulations were not published until two months prior to the deadline for submitting accountability workbooks.

Background to the Reviews and Decisions

- NCLB Enacted (January 2002)
- Standards & Assessment Regulations Issued (July 2002)
- Accountability Regulations Issued (early December 2002)
- CCSSO's AYP publication Released (mid-December 2002)
- Accountability Workbooks Released to States by ED (late December 2002)
- State Meetings with ED Officials Begin (December 2002)
- First Five State Accountability Plans "Approved" (Early January 2003)
- CCSSO Workshop for States on Accountability Workbooks (mid-January 2003)
- State Accountability Workbooks due to ED (January 31, 2003)
- Peer Reviews of State Accountability Plans (January through April 2003)
- Consolidated State Application Materials due ED (May 1, 2003)
- ED "Approval" Decisions to States (January to June 2003)

In a July 24, 2002 letter to Chief State School Officers, Secretary of Education Rod Paige outlined a set of criteria that became known as ED's ten principles for accountability (see Appendix B for the complete list of these principles). In December 2002—a few weeks after promulgation of the final regulations on accountability—ED released a Consolidated Application Accountability Workbook that extended each of the ten principles into more specific Critical Elements with examples of situations that would and would not meet the underlying NCLB requirements. ED directed States to respond to each of the Critical Elements and submit their completed workbooks by January 31, 2003. In early January, CCSSO conducted the only national workshop offered to assist States in completing the workbook. These workbooks were then reviewed both onsite in each State by a team of three peers and ED staff who provided an analysis of whether each State's plan met the requirements of the law. Beginning in December 2002, ED also paid for State delegations to meet with department officials in Washington to discuss their plans prior to the Peer Reviews.

As part of a pilot for the workbook and review process, ED invited seven States (Colorado, Indiana, Louisiana, Massachusetts, Mississippi, New York, and Ohio) to submit their workbooks early and participate in a review during December 2002 and early January 2003. This pilot had two results. First, accountability plans for five of the initial seven States (Colorado, Indiana, Massachusetts, New York, and Ohio) were "approved" by Secretary Paige in an early January 2003 ceremony coinciding with the one-year anniversary of the NCLB signing (for some of these States, it would be several months before they received follow-up letters detailing the parts of their plans that needed modification). Second, ED used feedback from these States and from the Peers who took part in the pilot reviews to create a more detailed reporting template (Peer Review Report for Title IA Accountability Provisions of the No Child Left Behind Act of

2001) that would be used to capture key information in each of the subsequent Peer Reviews.⁴

The more central issues that emerged from this analysis of States' accountability plans and ED's approval decisions are described in Part II of this paper. It is the authors' intent here to provide a descriptive summary of information gathered directly from States. Thus, **the paper does not represent an evaluation of either the process or the outcomes associated with the accountability workbook reviews, nor does it preclude the need for such evaluations.** Further, readers are cautioned against assuming that the elements and strategies that other States are using can automatically be applied in their own States or would be effective in meeting a State's accountability goals. The former assumption would rely on ED's approval and the latter is a matter for empirical study.

In addition, States' accountability plans varied both within and across States in the extent to which specific strategies were made explicit. Not all States, for example, clearly described how they would calculate their AYP indicators, including how they define their numerators and denominators. Extensive follow-up work with each State would be necessary to capture all of these differences. Though beyond the purpose and scope of the present paper, such follow-up study would greatly enhance one's understanding of how States' accountability systems function and how they compare across States.

By the end of May 2003, more than half the State accountability plans had been approved. Then on June 10, the President announced that all State plans had been "approved." As indicated in an earlier footnote, it is important that readers understand that, technically, no State accountability plans have been fully "approved" by ED. In most (but not all) cases, States have received a letter from Secretary Paige stating that, "we have approved the basic elements of [State's name] accountability plan." This has customarily been followed by a statement later in the letter to the effect that, "Under Secretary Hickok will provide you a corresponding letter detailing the conditions of your approval." It is in this second letter from Under Secretary Hickok that the issues States must address to receive final approval are listed. Based on the information in the Hickok letter, States need to provide "updated information" in relation to the listed issues/concerns. Consistent with past practice regarding the release of Federal education funds, issues remaining unresolved could become conditions or stipulations to receipt of 2003-04 NCLB funds.

Neither the Paige nor Hickok letters nor any other related correspondence has been made public as of July 16, 2003⁵. The authors of this paper contacted States to obtain copies of these documents. In some cases, due to on-going negotiations with ED or within the State, States chose not to share some or all of their NCLB accountability plan documentation at this time. The Peer Review Reports have never been released to the public or to the States, and consequently could not be considered in this summary.

⁴ In each of the subsequent reviews, the three Peers consolidated their comments into a single report using this reporting template. This single report was then submitted to ED, usually within one week of the Peer Review meeting. Beyond the submission of this report, Peers had no further knowledge of or input into the decision and approval process. Following each Peer Review, States received follow-up contacts from an ED representative to discuss areas of concern identified during the review and, typically, to request that the State submit additional clarifying or supporting information. These initial follow-ups do not appear to have been documented in a formal record of which the authors are aware; therefore, no public record exists for review. Further, since the Peer reports have not been made available to the general public, there is no way to determine how the Peers' input has been related to the specific issues ED has raised with States or to the approval decisions in general. Because ED has not publicly released any information about the review and plan determinations for any of the States, the writers have relied on the individual States for the information presented in this paper.

⁵On July 18, the State Accountability Plan Decision Letters were released on the U.S. Department of Education website at www.ed.gov/offices/OESE/CFP/al/index.html for half the states. Additional letters were to be posted as they became available.

Finally, readers should be aware that some of the information presented in this paper might change as the result of on-going negotiations between some States and ED over various accountability workbook issues. Lynn Olson, in an *Education Week* article, “All States Get Federal Nod on Key Plans,” (June 18, 2003) observed that some State representatives are wondering “exactly what approval means at this point.” Olson quoted one State official who noted that; “It’s interesting because there are still lots of items in our state accountability workbook that we are working on, that we have still not reached a decision about, that we are still negotiating with the U. S. Department of Education. ... There are still a lot of unanswered questions.” Another individual interviewed by Olson for the article observed, “Since the plans themselves, and the basis for approving them, are not yet widely available or publicly available, it’s hard to know what to make of it....”

In Part II, many of the substantive issues that arose during the Peer Reviews are identified and discussed. Specific examples of how ED’s approval decisions evolved over the course of the Peer Review process are provided in Part III of this paper. It is likely that additional examples will yet emerge as a result of the continuing plan approval negotiations in spite of the fact that ED has reported that all plans have been “approved.”

Part II

Issues in States' Accountability Plans

As noted earlier, States were required to submit a Consolidated Application Accountability Workbook to ED by January 31, 2003, in which they presented, at a minimum, their preliminary accountability system designs. In this workbook, States were to address a number of "Critical Elements" related to the ten principles ED set forth for the design and implementation of statewide accountability systems. During the ensuing months, ED conducted an onsite Peer Review of each State's proposed accountability systems and began to release approval determinations. ED required States to finalize their accountability systems by May 1, 2003, addressing issues raised through the Peer Reviews and, specifically, the issues noted by ED in the negotiations process that followed the Peer Reviews. Of course, States can always amend their plans at any time, although these amendments would need to be approved by ED. Although many Peer Reviews were completed just prior to May 1, ED was still negotiating various aspects of their accountability plans with approximately 75% of the States at that time. Under sec. 1111(e)(1)(C), the Secretary is required to "approve a State plan within 120 days of its submission unless the Secretary determines that the plan does not meet the requirements of this section." ED did meet this requirement.

As evidenced in the examination of State Accountability Workbooks and ED's approval decisions, the final accountability system designs vary markedly, reflecting the uniqueness of each State's approach to public education, attendant State laws, assessment and accountability system designs, and political influences. Further, States did not interpret all of the NCLB requirements in the same manner and some have continued to pursue system components that ED has deemed as not being consistent with the NCLB statute and regulations. Across the States, accountability system components vary in complexity. States' existing systems and their capacities for implementing these systems differed considerably prior to NCLB and influenced their plans for incorporating NCLB requirements into their own contextual situations.

ORGANIZATION OF PART II

The central issues presented in Part II are organized into several categories:

- Standards and Assessments in General
- AYP Model
- Inclusion
- Starting Points, Annual Measurable Objectives, and Intermediate Goals
- Participation Rate and Other Academic Indicators
- Validity and Reliability
- AYP Consequences and Reporting

Each section includes an overview followed by more specific information about the details of some States' approaches. Certainly, several of the issues could appear under more than a single heading. The authors hope readers find the current organization useful for understanding the issues. Readers may obtain more information at CCSSO's website

(www.ccsso.org/nclb) or ED's website (www.ed.gov/offices/OESE/cfp/csas/index.html). Readers should also review the approved State plans available at either website to obtain greater detail regarding the State context (and rationale) for each of these issues.

Standards and Assessments in General

Although this paper focuses on State accountability systems, these systems are dependent upon a State's academic content and student achievement (called "performance" under IASA) standards and its assessment system to generate the data necessary to make accountability determinations. The critical information that feeds into the accountability system comes from the assessments, which are to be based on the standards. In addition, the perspectives that underlie each State's accountability system presumably also underlie its approach to assessment. So, it seems appropriate to consider a few assessment issues here and to do so before moving onto the accountability issues, *per se*, keeping in mind that ED has repeatedly said that it does not consider "approval" of a State's accountability plan to indicate approval of its standards and assessments (which may be subject to a separate review process).

By January 2002, when NCLB took effect, only about one-third of the States had fully met the standards and assessment requirements for NCLB's predecessor, the Improving America's Schools Act of 1994 (IASA). Many were still working toward completion of academic content standards and student performance standards (called "achievement standards" under NCLB) and assessments, aligned with these standards, to be administered at least once annually in each of grades 3 through 5, 6 through 9, and 10 through 12. As of June 2002, 20 States were operating under a Waiver of Timeline Agreement with ED and five were operating under a Compliance Agreement to meet these requirements. In other words, about one-half of the States did not yet have systems with assessments in both reading or language arts and mathematics, aligned with their academic content and student achievement standards, in place in each of the 3-5, 6-9, and 10-12 grade spans—let alone in each grade, 3 through 8.

Under NCLB, States have until the 2005-06 school year to expand their standards to reflect grade-level (rather than grade-range) expectations and to implement aligned, annual reading or language arts and mathematics assessments in each grade, 3 through 8, and at the high school level (at least once annually in grades 10 through 12). Science assessments must be implemented at least once annually in each of the 3 through 5, 6 through 9, and 10 through 12 grade spans by 2007-08.

In their examinations of States' accountability plans, Peer Reviewers did not address the specifics of States' standards and assessment systems. (ED has consistently signaled to various State representatives that these will be reviewed, as necessary, at a later date under a separate review process.) However, as noted above, it is really not possible to think about or consider the systems separately. For example, without a clear understanding of how a State determines whether a student is proficient in reading or language arts, especially when results from two or more tests contribute to that rating, one cannot grasp the meaning of Proficient at the student level, or of the aggregate Percent Proficient indicator at the school or district level. It also logically follows because of the interdependence between assessments and accountability that it might also be necessary for ED to revisit some aspects of States' accountability plans after review of

State standards and assessments as described in the section below on student achievement standards.

For the present purposes, the primary issues with regard to accountability systems are States' student achievement standards and the consideration of student achievement results in reading or language arts and mathematics in each of the required grade levels.

STUDENT ACHIEVEMENT STANDARDS

States were also required to submit to ED by May 1, 2003, as part of the Consolidated State Application process, detailed information related to timelines for developing and implementing the additional standards and assessments required under NCLB. How these will be reviewed with respect to the NCLB requirements is unknown at this time. ED representatives have indicated to some States that systems of standards and assessments are likely to be reviewed in a separate process later this year in a follow-up to the accountability system reviews. The additional standards and assessments could also be reviewed at this time. For the accountability plan reviews, however, Peers were asked to consider only how the results on any alternate assessments were to be combined with results on the regular assessments. This generally involved a superficial review of the alignment between achievement standards on the two types of assessments, achieved through questioning of State staff during the Peer Review.

Even though States' achievement standards were not directly reviewed in the accountability plan approval process, it is worth noting here that NCLB introduced a new accountability framework for States, thus changing the context in which achievement standards will be applied from this point forward. Since annual performance targets and ultimate accountability goals are based on the percent of students achieving proficiency, where a State sets the proficient bar has major ramifications for how its AYP model will play out for schools and districts.

Understandably, some States have seen NCLB's passage as a time to revisit/review their achievement standards. This has not always been seen in a positive light. In an *Education Week* article ("States Revise the Meaning of 'Proficient'," October 9, 2002), author David J. Hoff reported on three States (Colorado, Connecticut, and Louisiana) that decided to modify their definitions of what students need to know and be able to do to demonstrate proficiency; that is, they had changed or redeveloped their definition of proficiency or had changed the label used for one or more levels since NCLB was signed into law⁶. In a more recent *New York Times* article, "States Cut Test Standards to Avoid Sanctions (May 22, 2003)," author Sam Dillon concludes that many States are "Quietly...doing their best to avoid costly sanctions [for schools and districts]." Dillon reports that in addition to Colorado's inclusion of "partially proficient" students with "proficient" students in the group considered proficient for NCLB AYP purposes, Texas has reduced the number of items students must pass on the State's assessments while Michigan has lowered the percentage of students who must pass the statewide tests in order to assert that a school has made adequate yearly progress (AYP).

Although an ED spokesperson "rejected the argument that states won't set and keep high standards," Dillon points out that "the law leaves it up to the states to establish their

⁶ Contrary to the information in the Hoff (2002) article, Louisiana did not set a new proficiency standard; rather, the State renamed its Proficient level, changing its name to Mastery (personal communication, J.P. Beaudoin, May 2003).

own standards of success.” It is important to keep in mind that, as noted above, States set their academic standards under the 1994 ESEA reauthorization based on a very different accountability construct. Given the different approach to accountability under NCLB, it should not surprise many that States might chose to revisit their standards to ensure alignment with the new construct.

In addition to considering how States’ achievement standards may change over time under NCLB, the Peer Review process did include discussion of National Assessment of Educational Progress (NAEP) State-level scores as a point of comparison with States’ achievement standards. ED has not announced any specific plans for conducting such comparisons⁷.

INCLUSION OF BOTH READING AND WRITING ASSESSMENT RESULTS IN PERCENT PROFICIENT CALCULATION

States’ academic content standards (often called “frameworks”) are always structured around basic content areas—though the specific areas may vary across States. In the area of language arts, some States have separate standards in reading and writing while others have a single set of standards that cover both reading and writing. In the latter cases, reading and writing may be addressed in different strands, but sometimes single strands cover both reading and writing content.

At this point, nearly all States have systems yielding separate scores for reading and writing, usually because these skills are assessed with separate tests and, especially in the case of writing, assessed only at two or three grade levels. NCLB specifically requires the inclusion of reading or language arts results in AYP. Following the requirements of the law, many States proposed AYP models that included only reading (and mathematics) scores. Some (e.g., **Florida**) included writing results as their other academic indicator for the elementary and middle school levels. However, it appears that ED has required some States (e.g., **Delaware**) to combine reading and writing results for use in the primary Percent Proficient AYP calculations. Other States that have combined standards, such as **Wisconsin**, have been allowed to use only reading results in AYP.

As the Peer Reviews began, ED was advising States with *language arts* content standards, including reading and writing components, that assessments addressing the full range of these standards must be part of AYP determinations. Thus, if a State intended to assess only a portion of these standards, such as only the reading strands, that decision represented a change in its standards for making AYP determinations, and would be subject to a “re-review” by ED. Changes or additions to a State’s assessments used for AYP determinations would also likely require a similar re-review. However, as the Peer Reviews progressed, it became clear that more and more States with language arts standards including reading and writing components appeared to be opting to use only reading for AYP determinations, and ED began to accept these proposals without mention of a need for a follow-up review. Thus, **Delaware**, for example, which was reviewed early in the process, was required to include both reading and writing results in the AYP Percent Proficient indicator but **Florida** and **Wisconsin**, which were reviewed

⁷ The Education Trust’s *Education Watch 2003 State Summary Reports* (www.edtrust.org) include State assessment results and comparisons with NAEP results by state, although only limited guidance is provided for understanding score differences and comparisons. The CCSSO series *State Education Indicators with a Focus on Title I* (www.ccsso.org) reports state assessment results and trends and NAEP state-level results.

later, were not. However, **Florida** did elect to use writing as its other academic indicator at the elementary and middle school levels, effectively making writing part of its AYP determinations (although without the same requirements for annual measurable objectives, intermediate goals, or eventual 100% proficiency). It should be noted that **Florida** is considering some changes in its State assessments and anticipates it will need to clarify these as part of its final accountability system approval.

Although ED has emphasized this is a State-by-State decision hinging on how reading and writing are represented in States' content standards, this did not seem consistent with the pattern of approvals as they evolved over time.

In addition, States' achievement standards are typically set separately for reading and writing and ED has not addressed how States are to determine the Percent Proficient for the combined reading and writing scores. For example, it is not clear whether these combined scores can be compensatory or whether reading proficiency should be given greater weight. In the absence of clear expectations, States have taken several approaches. Notably, **Delaware** received approval for weighting reading scores more heavily than writing scores in their overall language arts index, arguing that the writing scores tend to be less reliable than the reading scores. This suggests that States would not need to ensure that the combined score reflects the proportions apparent in the academic standards, at least for NCLB purposes.

Finally, it should be noted that most States administer writing assessments only in a subset of the grades in which reading must be assessed. Whether this will change over time as States develop new assessments to fulfill NCLB requirements is unknown. It is also unclear how inclusion of writing only at certain grade levels will eventually affect alignment of standards and assessments in States at those grade levels where writing is not assessed.

EVOLVING ASSESSMENT SYSTEMS

States such as **Alabama, Idaho, Michigan, Montana, New Mexico, South Carolina, and West Virginia** as well as the **District of Columbia** have not finalized their assessment systems and are working on agreements with ED for this purpose. In many instances, these and other States are in the process of phasing out norm-referenced tests (NRTs) and phasing in new criterion-referenced tests (CRTs) or are changing over to augmented NRTs. For the most part, several of these States have been using a mixed, somewhat transitional system of NRTs and CRTs for AYP purposes. It is probable that this will necessitate further review of several aspects of their AYP models once the final assessments are on line. Readers are also reminded NCLB requires in sec. 1111(b)(3) that States implement "a set of high-quality, yearly student assessments," further setting forth the related requirements but not specifically addressing types of assessments such as NRTs. The latter is addressed, however, in §200.3(ii)(A) of the standards and assessments regulations (July 2002). States opting to use NRTs for AYP purposes are required to assure that they are "augmented with additional items as necessary to measure accurately the depth and breadth of the State's academic standards..." In the analysis of comments and changes appendix to those regulations, the Secretary noted "student results from an augmented nationally normed assessment must be expressed in terms of the State's achievement standards, not relative to other students in the nation [p. 45045]."

Use of up to three sets of assessments—(1) old system, e.g., NRT; (2) transitional system, e.g., NRT some grades, CRT others; and (3) new system, e.g., CRT all required grades—to make AYP determinations results in an accountability system that is unwieldy at best. The scores on different tests carry different meaning and many States lack the capacity to monitor and evaluate the impact of these differences on the resulting accountability inferences. Thus, in some States, the scores on which AYP are based will vary over time, yet schools and districts will be required to continue making steady improvements in their achievement scores. NCLB makes no concessions for changing assessment systems, requiring in all cases that an AYP decision be made every year for every school while progressing toward the target of all students at the proficient level in reading or language arts and mathematics by 2013-14.

STATE-LOCAL ASSESSMENT SYSTEMS

Under NCLB (and also its predecessor, IASA), States are allowed to use results from only statewide assessments, a combination of State and local assessments, or only local assessments for accountability purposes. States that are well-known for their use of locally-selected and/or locally-developed assessments, such as **Maine**, **Nebraska**, and **Iowa**, have only been recently approved under NCLB and had to make their cases for approval of accountability systems based on data derived from these assessments.

In **Nebraska**, districts are required to use the School-based Teacher-led Assessments and Reporting System (STARS) or “Rule 10” or administer NRTs that, together, cover the academic content standards (although not all assessments required under NCLB will be administered until 2003-04). The State has prescribed four achievement levels—basic, progressing, proficient, and advanced—and each district defines the cut scores that correspond with these achievement levels on its assessment, using criteria established under “Quality Indicators.” Thus, although the achievement level descriptors do not vary across districts, the meaning of Proficient can vary across districts. However, the State does employ an annual evaluation of each district’s standards and assessments. Each district submits an assessment portfolio to the State and an expert panel evaluates the assessments and processes established by school districts for determining student achievement levels. After each assessment cycle, districts report the number of students scoring at each achievement level to the State. For the NRTs, the proficient level is defined as a national percentile rank of 50 to 74. Nebraska has set the starting points and intermediate goals based on either the local assessments or the required norm-referenced tests if a local assessment is not available. The State has also determined a statewide trajectory for NCLB AYP decisions. Nebraska has State academic content standards.

In its AYP model, **Iowa** will use the results from the Iowa Tests of Basic Skills (ITBS) or the Iowa Tests of Educational Development (ITED). Iowa argues that these assessments are “common comparable measures across all schools, thus ensuring fairness, validity, and reliability when making unbiased, rational, and consistent determinations” and has no plans to augment or otherwise modify these standardized norm-referenced tests for NCLB AYP purposes. For AYP, the State defines proficiency as the 41st percentile or higher (2002 National norms—spring standardization study) and plans to report results based on the 2000 national norms (spring 2000 standardization study) through 2013-14. School districts determine from three windows—fall, winter, or spring—when the tests will be given. It should be noted that **Iowa** has also not developed State academic content standards.

In **Maine**, an advisory committee will recommend to the Commissioner the AYP starting points for reading and mathematics based on the State's performance on NAEP by "equating"⁸ performance on **Maine's** comprehensive assessment system with average NAEP performance for the content area and grade span. **Maine's** AYP starting points will be no less than the NAEP national average. Six starting points will be established for reading and mathematics at grades 4, 8, and 11.

FIRST ADMINISTRATION RULE

Some States offer students the opportunity to retake a required test they did not pass. This practice is especially prevalent at the high school level when the test is an end-of-course or graduation measure, but it does occur at the lower grades as well. Sometimes, students are allowed additional attempts within the same school year. At the high school level, many States allow the first attempt to take place in grade 9 or grade 10—even though the tests typically assess knowledge and skills required for graduation at the end of grade 12—with subsequent attempts throughout high school. While approximately 20 States now have high school graduation or exit examinations, not all States addressed in their workbook plans how multiple test attempts would be accounted for in terms of AYP and Participation Rate calculations.

In these multi-attempt situations, NCLB regulations, §200.20(c)(3), require States to use the first score a student obtains in their AYP calculations; something not required under the NCLB statutes. After that rule was published, at least one State wrote to ED requesting an agency review of "three regulatory decisions [that] were published without any period of required review...." One of those rules was the section cited in this paragraph. ED has invited States to comment on whether this regulation should be amended in its March 20, 2003, Notice of Proposed Rule Making (NPRM) pertaining to the academic achievement of students with the most significant cognitive disabilities.

So far, the trend is mixed with regard to strategies for including results of multiple administrations of high school course exit or graduation exams in AYP calculations. **New York** received approval for its plan, which gives credit for students passing the graduation exam prior to grade 12 but does not penalize schools for non-passing scores achieved prior to grade 12. For example, a student's first attempt may take place in grade 11, but that student's score will not count for AYP unless the student passes. If that student fails and reattempts in grade 12, the grade 12 score will count regardless of whether she or he passes or fails. The rationale here is that, because the test is considered a grade 12 assessment, attempts in earlier grades are considered to be "accelerated."

New Jersey's plan permits students up to three attempts on the State's High School Proficiency Assessment, but the State will count only the spring grade 11 administration for accountability purposes. In **Michigan**, high school assessments are governed by State law and include the opportunity for students to "dual enroll" in college classes while in high school based on exhausting the high school curriculum. Students now seeking to qualify for dual enrollment in grade 11 are allowed to take the assessments in grade 10. **Michigan** received ED's approval to recognize a 10th grader's score of proficient on an early assessment and a grade 11 score of proficient for those students in dual enrollment who test in grade 10 but who do not score proficient or better at that time.

⁸ The details of this strategy are not clear; Maine does intend to apply the NAEP-based starting points at the State, district, and school levels.

Nevada will use cumulative pass rates up to and including its grade 11 April administration of the high school exit exam for a given graduating class as the numerator in the percent proficient for AYP determinations. The denominator will include all students in the numerator plus all students who participate in the grade 11 April test administrations. Participation rate will be calculated based on the ratio of 10th graders taking the high school exam divided by the total grade 10 enrollment. In 2003-04, the State will move to tracking cohorts from fall grade 10 to the April administration in grade 11.

Alabama's High School Graduation Test allows students to “pretest” in the grade 10. If a student scores at the Proficient level, the score is “banked” for graduation requirements. The grade 11 assessment, considered the “official administration,” will be used for making AYP decisions. With regard to participation rate, **Alabama** will use the following definition: “number of grade 11 students enrolled according to the 120-day enrollment report who either have previously passed the *Alabama High School Graduation Exam* or who attempted a state assessment in the spring of grade 11 divided by the number of grade 11 students enrolled according to the 120-day enrollment report.”

Additional examples illustrate the complexity of this issue. **Ohio** currently administers a few assessments more than once during the school year including one in reading at the fourth grade level. The State argued that it administers these assessments more than once annually for diagnostic purposes and that combining results from several assessments of one test within a year is a better reflection of student and school performance. ED originally indicated in its approval letter that, “Ohio can continue its practice of offering students multiple opportunities to take an assessment, yet, for NCLB accountability, students' results from the first assessment must be the results used in AYP decisions...” The ED letter continued, “the Ohio fourth grade assessment...is designed to measure what students know at the end of the year. In particular, while giving the fourth grade assessment early may provide insightful diagnostic information, it does not seem like an early administration of this assessment would be a good reflection of what fourth graders should know and be able to do at the end of the year. As such, the results for AYP purposes must come from the first official administration of these assessments and not assessments given for diagnostic purposes.” Thus, it seemed that **Ohio** would be required to use the results from only the final administration and not allowed to consider the cumulative percent proficient over a school year for AYP. However, as this paper was being finalized, ED has indicated (but not yet confirmed) to the State that for its elementary school assessments where multiple administrations are given, cumulative results can be counted.

Oregon was also initially advised by ED that their Technology Enhanced Student Assessment (TESA) system might not meet NCLB requirements for accountability purposes. (TESA was approved under the IASA standards and assessments review) because not all schools yet had access to this system and the State was also using another assessment for AYP purposes. TESA is an on-line system of adaptive tests that students take several times a year to assess their progressing levels of proficiency; the adaptive format means that no matter how often a student accesses the tests (up to three times annually) that student will see a fresh form because the items are dynamically drawn from an item bank for each administration. Even though the scores are based on different samples of items, they carry comparable meaning across administrations and students because the items have been calibrated to a common scale. The State uses the immediate feedback from the on-demand results of this system to inform instruction. For AYP

purposes, Oregon proposed to use the percent of students who, over the year, met relevant benchmarks. ED initially rejected this proposal.

At issue was (1) how the Participation Rate is determined and (2) how **Oregon's** practice fails to meet the "first test/first score" regulation. ED asked the State to impose a common testing window for determining AYP. Thus, the State put this procedure into operation by counting the results for the test(s) taken closest to May 1. Whether students who had already demonstrated proficiency would have to sit for this test is unknown at this time although follow-up conversations suggest that early-testing students demonstrating proficiency (something that not many are able to do) might have these results recognized for AYP determinations, (This would be more consistent with ED's recent decision regarding a similar practice in **Ohio**).

The practical effect of ED's regulation and related policy at the elementary and middle school levels is that a State's use of diagnostic assessments throughout the school year to help measure students' subject mastery may be permissible depending on supporting arguments and rationale. The State would be required to designate a single point in time at which assessment results are used for AYP purposes. Students demonstrating proficiency through the diagnostic assessments or other forms of "early" testing would be able to have their scores recognized and not have to sit for further testing. At the high school level, the key as to what ED approves seems to be the point at which students are expected to have taken the courses that contain the content standards assessed in a normal sequence (on track for graduation on time). So as in **New York**, if a student takes the high school assessment before the grade 12, but all of the standards are not covered until that grade, the scores do not count until grade 12 unless the student "passes." If, in another State, the standards that are included on the assessment are covered by the grade 11, a student's scores taken at grade 11 are the ones that count for AYP even if he or she takes it again at grade 12 before "passing."

AYP Model

This section addresses the performance variables used in AYP calculations, the integration of NCLB AYP with States' other accountability systems, and the strategies States have proposed to enhance the reliability—and sometimes also the validity—of AYP decisions. In developing this section of the paper, the authors observed that more "sophisticated" accountability systems seemed closely linked to a State's capacity—staffing levels, resources, and rich data bases. AYP models employing multiple tests for reliability and validity in decision-making appeared to be much more reflective of the extent to which a State had a wealth of data and the ability to commit staff, technical assistance, and other resources to conduct research and analyses. These States were also typically more able to involve a wider array of stakeholders in building their systems.

It should also be noted that under Critical Elements 3.1 through 3.2b (see also Question A7 in the Peer Review Report) States were required to describe in their accountability workbooks the methodologies/criteria/procedures they intended to use to determine whether each student subgroup, public school, and LEA makes AYP. However, no examples of acceptable models were provided nor has ED yet issued related guidance to assist States or reviewers in making judgments related to this matter. No examples were provided in the "Examples for Meeting Requirements" column of Critical

Element 3.2 of the State accountability workbook either; instead a portion of the accountability regulations are reiterated.

Clearly, States put forth a wide variety of models for determining how schools and districts will be identified under the law. In some instances, they reported being questioned at length during the Peer Review process and ED insisted on changes such as those described below under *Independence of AYP Indicators* for **Delaware** and **Wyoming** at the end of this section. In other cases, how a State proposed to calculate AYP was not the subject of much discussion during the Peer Review nor addressed to any significant degree in follow-ups from ED. Although ED did develop a related internal policy (see References/Resources at the end of this paper), that policy covers only the option of States basing AYP determinations on missing AMOs in the same subject for two consecutive years or missing the AMOs in either subject for two consecutive years⁹. It does not address the impact of Participation Rates or Other Academic Indicators. That policy (and six others) has not been made available to States or the general public.

AYP INDICATORS

The range of options available to States in the selection of indicators for NCLB AYP calculations is limited. States are required to use five kinds of indicators for AYP:

- Separate summary indicators for proficiency in reading or language arts;
- Separate summary indicators for proficiency in mathematics;
- Separate indicators of participation in reading or language arts assessments;
- Separate indicators of participation in mathematics assessments; and
- At least one other academic indicator at the elementary and middle school levels and at least graduation rate at the high school level.

The graduation rate at the high school level was intended to be narrowly defined (see §200.19 of the accountability regulations) although States can also submit another definition for the Secretary's consideration. The other academic indicator was left to States' choosing at the elementary and middle school levels. States could choose to include additional indicators, but these indicators would have to operate conjunctively with the five required ones, meaning that they could have the effect of maintaining or increasing the number of schools identified for improvement but could never decrease this number. For obvious reasons, few States added extra indicators to their AYP model. This section considers the performance indicators; participation rate and the other indicators are discussed in a subsequent section.

Percent Proficient

With regard to calculating the indicators used to make determinations regarding proficiency, all States chose to either use a straight percent proficient or an index in which a value is attached giving at least some credit toward proficiency for student achievement scores falling below that level. Most States decided to use a simple percent proficient in their AYP calculations; this is the statistic described in the law and regulations and is generally simpler to calculate than an index.

⁹ Neither NCLB nor the related accountability regulations specify exactly how AYP is to be calculated.

average across grades or groups, or a composite combining multiple types of indicators. In the weighted performance indices, less credit is given for performance below proficient than above. At its simplest, such an index would equal the percent proficient by representing each score below proficient with a zero and each score above with a one. Or, a State could give, for example, zero credit for performance in a Below Basic level, .5 credit for each score in the Basic level, and 1 credit for each score in either the Proficient or Advanced level

As the Peer Reviews progressed, ED took the position that States could include weighted performance level indices in their AYP models provided that (1) reading or language arts and mathematics are treated separately and (2) additional points are not allocated for an advanced level of performance that could mask or compensate for the performance of students below proficient. **Delaware** and **Oregon**, for example, were advised by ED that their weighted index scores would not be allowed for NCLB purposes because higher weights were given to score levels above proficient. In putting forward its State Board approved index, **Oregon** proposed to assign 33 points to a low score, 67 to a “partially meets” score, 100 points to a proficient score, and 133 to an advanced score. The State set its 2014 target at 115 points—halfway between proficient and advanced. A scatter plot was presented based on actual data from the State’s schools demonstrating a correlation of $r=.96$ between percent proficient and the index. Oregon concluded and argued unsuccessfully that while it is theoretically possible that a school with many advanced students could compensate for some students below proficient, the effective difference between looking at the percent proficient and their index in practice is negligible.

Mississippi received approval for its AYP model, which includes a weighted average of performance across grades as an index. In **Mississippi’s** index, the school-level percent proficient for a given group, such as Hispanic students, is calculated by first comparing the percent proficient at each grade level with the target and then weighting these by the proportion of the total school “n” for Hispanic students represented at each grade level. The index is a sum of the weighted differences. The index appropriately represents each student’s score in proportion to the total number of scores; simply averaging the percents from each grade level would give disproportionately higher weights to scores in grades with smaller enrollments.

Delaware initially proposed the use of an index in which each student’s representation was apportioned across subgroups rather than repeated across subgroups. Every student’s score would be included in the total student category; each student would also be represented proportionately in the summaries for each student’s appropriate subgroups. Scores for Sally, who is white, eligible for free lunch, is LEP, and receives special education services would be apportioned 25% in each of these four subgroup summaries; scores for Sally’s classmate, Ron, who is African-American and qualifies for no other category would be represented 100% in the African-American category. **Delaware** had to remove this model from their AYP system prior to its approval. ED indicated that apportionment was unacceptable and students would have to count multiple times, stating that the weighted method “diminishes the impact on school accountability of any subgroup in which most students count 1.0.” The reality, at least for students served in Title I programs, however, is that they are likely to count in at least two subgroups, and often in three (race/ethnicity, economically disadvantaged, and LEP or SWDs).

Delaware did win approval for its Language Arts index, which weights writing 10% and reading 90%. The State argued that the writing test is considerably less reliable than the reading test and, therefore, should contribute less to the total score. **Oregon's** AYP determinations will be based on a combination of results from a reading knowledge and skills test and a writing performance assessment. **Louisiana** will use an index with several components, one of which is a growth indicator, to identify schools for rewards above and beyond the AYP system but will not use an index for AYP itself.

Independence of AYP Indicators

In mid-June, it became clear from the review of accountability workbook approvals and conversations with State Education Agency (SEA) staff that some States considered each of the five AYP indicators to be independent while others did not. That is, many States plan to identify schools and districts for improvement only if they miss their AYP target for the same indicator two years in a row. For example, **West Virginia** groups the academic indicators (percent meeting the standard in reading or language arts and mathematics), the participation rate in each subject area, and the other academic indicator of graduation and attendance. Other States will identify schools and districts that miss either Percent Proficient or Participation Rate within one of the content areas (reading or mathematics) in each of two consecutive years.

As an example, State A considers Percent Proficient and Participation Rate to be independent, meaning that a school or district would need to miss its AYP target in Percent Proficient in each of two consecutive years to be identified for improvement. Missing the target for Percent Proficient only in year 1 and in Participation Rate only in year 2 would not result in being identified for improvement. State B pairs Percent Proficient with Participation Rate, so a miss in Percent Proficient only in year 1, followed by a miss in Participation Rate only in year 2 (Pattern 2 in the figure below) would result in being identified for improvement. These two cases are illustrated below (an X indicates that the AYP target was missed and the gray shading indicates a pattern that results in identification for improvement).

| | | Pattern 1: The 2 indicators within a content area are independent | | Pattern 2: The 2 indicators within a content area are paired | | | |
|---|-----------|---|----------------------------------|--|-------------------------------|--------------------------------|--|
| | | Reading % Proficient | Reading Participation Rate | Math % Proficient | Math Participation Rate | Other Academic Indicator | AYP Outcome |
| State A – only identifies for improvement using pattern 1 | Year 1 | X | | X | | | In need of improvement: Reading only |
| | Year 2 | X | | | X | X | |
| State B – identifies for improvement using patterns 1 and 2 | Year 1 | X | | X | | | In need of improvement: Both Reading and Math |
| | Year 2 | X | | | X | X | |

These issues did not seem to emerge earlier in the review process because many States' plans did not explicitly describe the pattern of performance that would result in identification for improvement. Two States, **Wyoming** and **Delaware**, brought this issue up themselves during their review process and were subsequently required to "pair" the

indicators within each content area (like State B in the illustration above). In this instance, the Other Academic Indicator (applied only to “All Students” for initial accountability determinations) acts independently or some what like a “wild card.”

DUAL ACCOUNTABILITY SYSTEMS

Under sec. 1111(b)(2) of NCLB, States are required to develop and implement a single, statewide State accountability system. Through most of the early Peer Reviews, ED appeared to insist that States do just that—present a single system of accountability applicable to all schools and districts regardless of whether they received Title I funds. The only exception “on the table” was the one authorized in NCLB legislation—a different set of rewards and sanctions could be applied in schools and districts not receiving Title I funds. However, States would still have to provide for rewards and sanctions applicable to schools identified for improvement but not receiving Title I funds.

In later reviews, ED signaled a softening of its position on dual accountability systems and no longer challenged these. As a general rule, ED’s position now seems to be that as long as the very top and very bottom school/district classifications and Title I school/district identification for improvement requirements are “in sync,” then dual accountability systems are acceptable. Based on discussions with SEA staff, being “in sync” appears to mean that at the very top, a State system may not recognize a school/district as high performing that is identified for improvement under Title I and a school/district identified as very low performing under that State’s system would also have to be identified for improvement under Title I. However, there appear to some exceptions to this “general rule.”

For example, in **Florida**, the existing A+ Plan for Education features measurement of academic growth for individual students. Schools earn points for students in the lowest 25% who earn achievement gains comparable to those of the norm group for the State. This value-added model is possible using **Florida’s** vertically-scaled assessments in grades 3 through 8 and its student identifier system. **Florida** proposed to bring the A+ Plan for Education into alignment with the requirements of NCLB’s unitary accountability system by offering that no school will be designated as meeting AYP if it has been graded “D” or “F” under the A+ school grading system. **Florida** asserts this two-tiered system is more challenging than the NCLB requirements.

Schools in **Virginia** will be able to achieve the highest accreditation rating even if they are identified for improvement under NCLB. The State uses four accreditation ratings to report school performance—Fully Accredited, Provisionally Accredited/Meets State Standards, Provisionally Accredited/Needs Improvement, and Accredited with Warning. In a June 9, 2003 letter to Under Secretary Hickok, **Virginia** Board of Education President Mark Christie expressed the concern that “Virginians should understand that many Virginia schools will achieve full accreditation—our highest rating—and other acceptable ratings under Virginia’s own successful Standards of Learning (‘SOL’) ratings system, yet be viewed as ‘failing’ in some respect under the federal AYP formula because of retroactive application of future policies.”

ED also approved **Arizona’s** plan for a dual statewide accountability system—a plan that can result in different “labels” for the same schools. The plan establishes five labels for **Arizona** schools for State purposes, from excelling to failing, but it is silent on the issue of consistency in reporting school performance for NCLB and State purposes.

The **Arizona** accountability plan contains the following components:

- Rewards schools for the academic gains of students who still may not meet State standards but show significant progress (schools receive credit based on overall improvement of test scores instead of improvement by one or more subgroups of students);
- Tracks the growth of specific students in the same school year over year to best assess the school environment—not other factors affecting a child’s education; and
- Is an annual method for tracking school progress—not a one-time “hit or miss.”

In **Louisiana’s** three-tiered model, schools are identified for improvement if they fail to make AYP either from the subgroup/NCLB analysis or the total school analysis (3rd tier). In addition, a school only attains the highest school designation, “Exemplary Academic Growth,” by meeting both the NCLB requirements and the School Improvement requirements. In **Ohio**, a school at the State’s second highest performing level could also be a school identified for improvement under Title I.

In **Michigan**, another State with an approved dual statewide accountability system, the State will use, in addition to NCLB, a school accountability/accreditation system framework that gives schools and districts a “report card” with A, B, C, D/Alert, and Unaccredited letter grades in six areas. After computation of a school’s (or district’s) composite grade for the six areas, a final “filter” will be applied to determine whether or not the AYP standards have been met. A school that makes AYP will not be listed as Unaccredited. A school’s composite grade will be use to establish priorities for assistance to “underperforming” schools and interventions to improve student achievement.

Iowa also received approval of an accountability system that it refers to as the “Relative Contribution Model.” Under this model, an LEA must first meet the statewide trajectory for NCLB AYP for all subgroups, and then meet its own trajectory for **Iowa** regulations. Local education agencies then may, for schools that are above the State's trajectory, apply the LEA's trajectory to all schools within the LEA, or calculate the “relative contribution” of each school building toward the LEA's trajectory. As such, uniform application of the trajectory formula will continue to expect lower performing schools to “make up” more ground (in order to reach the State's trajectory) than higher achieving schools.

STRATEGIES FOR (1) PROTECTING CONFIDENTIALITY AND (2) ENHANCING RELIABILITY

In the NCLB law and regulations, States are required to establish specific conditions under which their AYP indicators can be reported without (1) breaching confidentiality for any individual student and, separately, (2) the conditions under which AYP models are considered reliable (note that this is different from actually evaluating the reliability of AYP decisions). The key variables here are the decisions States will make with respect to¹⁰:

- Minimum “n” for reporting and protecting confidentiality;

¹⁰ Most, if not all, of these are discussed in CCSSO's recent publication, *Making Valid and Reliable Decisions in Determining Adequate Yearly Progress*.

- Minimum “n” for accountability determinations;
- Uniform averaging procedures under sec. 1111(b)(2)(J);
- Use of confidence intervals; and
- Use of standard errors of measurement.

Protecting Confidentiality in Reporting

To address the protection of confidentiality, all States identified a minimum number (n) of students/scores/data points necessary for reporting. Among the accountability plans “approved” to date, these minimum reporting “n’s” range from 5 to 30, with a mode of 10. Several States also suppress reporting of proportions nearing 0 or 100 as a further protection of students’ privacy.

Enhancing Reliability—Minimum “n” and Confidence Intervals

In developing the soundness of their theoretical bases and approaches to reliability of system design, States chose a minimum “n” of data points necessary for the calculation of a particular statistic such as Percent Proficient or the Participation Rate. In addition, several States will also apply some form of confidence interval (CI) to their AYP calculations (assuming the minimum “n” requirement has been met as a “first test”¹¹), but, for the most part, will generally do so only for their Percent Proficient indicators. **Maryland** and **Louisiana** are a notable exceptions in that they will apply a CI for Percent Proficient and when invoking “safe harbor,” an approach similar to those other States will use as reported in the section that follows on “safe harbor” determinations. **Maryland** also applies a 95% confidence interval to “safe harbor” determinations. **Louisiana** chose a 99% CI and **Mississippi** chose a 95% CI and only applies this test for Percent Proficient. **Kansas** and **Massachusetts** also elected to use a 95% CI. **Iowa** will utilize a 98% (one-tailed) confidence band as a significance test for its AYP calculations. **Georgia** has also indicated that it “will apply a confidence interval approach to determine AYP for small schools whose overall population is below the minimum number of 40.”

¹¹ It is not clear from reading a number of States’ plans whether or not a minimum “n” will be explicitly applied to indicators other than Percent Proficient; it is assumed in these cases that if the minimum “n” stated for Percent Proficient is not met, the standard AYP calculations are disrupted entirely and the State would have to employ other methods for determining AYP.

Table 1: Approaches to Enhancing Reliability in 50 Approved State Plans, the District of Columbia, and Puerto Rico

| State | Min. N to Report | Approach by Indicator | | | | |
|------------------|------------------|-----------------------------------|---|-----------------|--------------------------|--------------------------|
| | | Percent Proficient/ Index | Participation Rate | Graduation Rate | Other Academic Indicator | Safe Harbor |
| Alabama | *10 | N > 40 | N > 40 | | | |
| Alaska | 5 | N > 20 and 99% CI | N > 41 | | | |
| Arkansas | 10 | N > 25 over three yrs | | | | |
| Arizona | 10 | N > 30 and CI | N > 30 | | | |
| California | 11 | 50/15%/100 | | | 95% CI | |
| Colorado | 16 | N > 30 and 95% CI | N > 30 | | | |
| Connecticut | *20 | Subgroups: N > 40 and 99% CI | N > 40 | | | |
| Delaware | 15 | N > 40 | N > 40 | | | |
| District of Col. | 10 | N > 25 | N > 40 | | | |
| Florida | 10 | N > 30 | N > 30 | | | |
| Georgia | 10 | N > 40 | N > 40 | N > 40 | N > 40 | |
| Hawaii | 10 | N > 30 | N > 40 | | | |
| Idaho | *10 | N > 34 | N > 34, Sliding scale N < 34 | | | |
| Illinois | 10 | N > 40, +/-3% | N > 40 | | | |
| Indiana | *10 | N > 30 and 99% CI | N > 40 | | | |
| Iowa | 10 | N > 30 and 98% CI | N > 40 | N > 30 | N > 30 | |
| Kansas | 10 | N > 30 and SEM and 95% CI | N > 30 | | | |
| Kentucky | 10 | 10 per grade/30 per school and CI | 10 per grade/30 per school | | | |
| Louisiana | 10 | N > 10 and 99% CI | N > 40 | N > 10, 99% CI | N > 10, 99% CI | N > 10 and 99% CI |
| Maine | *10 | N > 20 and 95% CI | N > 41 | | | |
| Maryland | 5 | N > 5 and 95 % CI | N > 30, 1 gr. tested; N > 60, 2+ grs. tested, districts | | | N > 5 and 95% CI |
| Massachusetts** | 10 | N > 20 and SEM and 95% CI | | | | |
| Michigan | *10 | N > 30 | N > 30 | | | |
| Minnesota | 9 | N > 20 Sliding CI 95% to 99% | N > 40 | | | |
| Mississippi | *10 | N > 40 and 95% CI | N > 40 | N > 40 | N > 10 | N > 40 current year only |
| Missouri | 30 | N > 30 | N > 30 | | | |
| Montana | 10 | 95% CI | N > 40 | | | |
| Nebraska | 10 | N > 30, N > 45 SWD | | | | |
| Nevada | 10 | N > 25 and 95% CI | N > 20, N < 20: N-1 | | | N > 25, 75% CI |
| New Hampshire | 11 | N > 11 and 95% CI | N > 40 | N > 40 | N > 40 | N > 11 |

Table 1 continued.

| State | Min. N to Report | Approach by Indicator | | | | |
|----------------|------------------|--|--|---|--|-------------------------------------|
| | | Percent Proficient/ Index | Participation Rate | Graduation Rate | Other Academic Indicator | Safe Harbor |
| New York | 5 | $N \geq 40$ | $N \geq 40$ | | | |
| North Carolina | 5 | $N \geq 40$ | $N \geq 40$ | $N \geq 40$ | $N \geq 40$ | . |
| North Dakota | *10 | alpha=.01 | alpha= .01 | alpha=0.01 | alpha=0.01 | alpha=0.01*** |
| Oklahoma | *5 | $N > 30$ and 99% CI, $N > 52$ for subgroups | | | | |
| Ohio | *10 | $N \geq 30$ $N \geq 45$ SWD | $N \geq 40$ | | | |
| Oregon | *6 | $N \geq 42$ scores and 99% CI | | | | |
| Pennsylvania | 10 | $N \geq 40$ | | | | |
| Rhode Island | 10 | $N > 45$ and 95% CI | | | | |
| South Carolina | 10 | $N \geq 40$ | $N \geq 40$ | | | |
| South Dakota | 10 | $N \geq 10$ and 99% CI | $N \geq 40$ | | | |
| Tennessee | 10 | $N > 45$ | $N \geq 45$ | | | |
| Texas | 5 | $N \geq 30$ for all Students $N \geq 50/10\%/200$ for subgroups | $N \geq 40$ for all students $N \geq 50/10\%/200$ for subgroups | $N \geq 40$ for all Students $N > 50/10\%/200$ for subgroups | $N \geq 40$ for all Students $N \geq 50/10\%/200$ for subgroups | |
| Utah | *10 | 10 per year and 99% CI | $N \geq 40$ | | | Statistical test, 2003 alpha=.25 |
| Vermont | 10 | $N > 40$ and 99% CI | | | 99% CI | |
| Virginia | *10 | $N \geq 50$ | $N \geq 50$ | | | |
| Washington | 10 | $N \geq 30$ | $N \geq 30$ | $N \geq 30$ | $N \geq 30$ | |
| West Virginia | *10 | $N \geq 50$ | $N \geq 50$ | $N \geq 50$ | $N \geq 50$ | |
| Wisconsin | 6 | $N \geq 40$ $N \geq 50$ SWD and SEM | $N \geq 40$ $N \geq 50$ SWD | $N \geq 40$ $N \geq 50$ SWD | $N \geq 40$ $N \geq 50$ SWD | |
| Wyoming | 6 | $N \geq 30$ and CI | $N \geq 40$ | | | |

* This State suppresses results in cells with fewer than a specified number of students and also for cell proportions nearing 0 or 100.

** Massachusetts reports results for cells with 40 or more students over two years and no fewer than 15 students in either of these years. The State issues its improvement ratings for schools with an average of at least 20 students per year over two years, but fewer than 50 in either year, using "a custom determined error-band of up to 4.5 points" (MA-Consolidated State Application Accountability Workbook, p. 31) as well as a 95% CI. For schools averaging 50 or more students across two years and no fewer than 40 students in either year, the State uses an error band of 2.5 points.

***The alpha=0.01 will apply to safe harbor only after the state conducts a study of its effects and reaches agreement with USED on its application. Until the study is complete the safe harbor will be as prescribed in NCLB.

Initially, it seemed clear from the Peer Reviews and State "approvals" that ED would not allow the use of a CI for the Participation Rate or any other indicator considered a "count." However, as noted later in the section of this paper addressing Participation Rate

and Other Academic Indicators, ED did approve in late determinations at least two State plans employing the use of CIs with “count” indicators. These approvals were for **North Dakota’s** model (albeit with the caveat that other States proposing a statistical test on a “count” indicator would have to provide the supporting impact data) and **Louisiana’s** application of a 99% CI to calculations of percent proficient, reduction of non-proficient students, and status of attendance and graduation rates.

Minimum “n’s” also vary across subgroups in some cases. As has been widely noted, **Ohio** applies a minimum “n” of 30 for the total school or district as well as for all but one other subgroup. For Students with Disabilities, **Ohio** set a minimum “n” of 45 for calculation of Percent Proficient. Similarly, **Wisconsin** will use a minimum “n” of 50 for the SWDs subgroup and 40 for all other subgroups.

Oklahoma received approval for a minimum “n” of 52 for each individual subgroup and 30 for the all students group. The State’s rationale for a larger sample size for subgroups is based on the fact that multiple comparisons are made for each school. In other words, schools will be identified as failing if they fall below the standard for *any* of the relevant subgroups of students. Therefore, in consultation with their Technical Assistance Committee, the State adopted a more reliable 99 percent confidence interval for AYP decisions on subgroups, rather than the 95 percent confidence interval that it will apply to the all students group. The State arrived at a minimum “n” size of 52 by considering that schools will be identified as failing if they fall below standard in, on average, five to six subgroups. The probability of at least one error in five comparisons can be estimated as $5 * .01 = .05$ (assuming errors to be independent), which is the same as the probability of an error in the overall comparison using a 95 percent confidence band. Therefore, the minimum “n” for subgroup comparisons that is equivalent to a sample size of 30 for the overall comparison can be computed as follows:

- Overall Confidence Bound = $1.96 * SE = 1.96 * SD / \sqrt{30}$
- Subgroup Confidence Bound = $2.58 * SE = 2.58 * SD / \sqrt{N2}$
- Setting these two equations to be equal and solving for N2 results in a minimum “n” size of 52 for subgroup comparisons.

Texas proposed a different approach to applying minimum “n’s”—one the State has used in its accountability system for many years. For the “all students” group, Texas will use a minimum “n” of 30. However, for all subgroups, the State will do the following: if the subgroup has 200 or more students, it will be considered for AYP. If the subgroup has between 50 and 199 students, it will be considered for AYP only if it represents at least 10% of the entire student body. Subgroups with fewer than 50 members will not be considered for AYP. Texas refers to this as the “50/10%/200” rule. Similarly, **California** will require a minimum “n” of 50 students in a subgroup and these 50 students must represent at least 15% of the students at the school. If either of these conditions is not met, the subgroup minimum rises to 100.

Wyoming put forward an interesting variation of minimum “n” for accountability in its many small schools and districts. The State will adopt a rule whereby schools with fewer than 30 students, but at least 6 students with assessment scores, will be evaluated using a combination of AYP and Body of Evidence data. For an interim period, schools with fewer than 6 will be reviewed based on average data over the previous 2 to 3 years which is intended to reach at least 6 scores. **Montana** will use a 95% CI and no minimum “n” size. **Alaska** will use a minimum “n” size of 20 and a 99% CI. **South Dakota** will

use a minimum “n” of 10 plus a CI of 95%. **North Dakota** will use an alpha equal to 0.01 and no minimum subgroup size (exact probabilities as opposed to normal approximations will be used). There are an “overwhelming” number of small schools in that State; 58% of their 4th grade schools would not meet a minimum “n” of 25.

Enhancing Reliability—Uniform Averaging

In most States, data will be combined across grade levels within schools and districts for AYP purposes. When States’ full assessment systems are in place, this will usually increase the number of data points on which the Percent Proficient statistic will be based. Until then, this has little real impact on AYP determinations in most jurisdictions.

A number of States will also consider multiple years of data in their Percent Proficient calculations. Some, like **West Virginia**, will always (when available) consider three years of data. Others (e.g., **Ohio** and **Tennessee**) will either use the single current year or the average of the current year and the previous one or two years, whichever score results in the best standing for the school or district. This option is applied independently for each school and district and is intended to account for unreliability of data when it may result in a questionable identification of a school yet not penalize the school when it would not result in identification. Of course, the benefit is not long-term since a low score one year may be offset when averaged with previous higher scores but that same low score will depress subsequent averages. It is not clear from most States’ plans whether these averages will be weighted by the number of scores for each year as it would be most appropriate to do (student enrollment typically varies from year-to-year).

This allowed variation within a State in the data used for AYP does reflect greater flexibility than what may have been assumed earlier. Section 1111(b)(2)(J) of NCLB specifically permits States to establish uniform procedures for averaging data and ED had indicated to some States during Peer Reviews that States exercising these provisions must apply them uniformly across all schools and districts.

Another question concerns whether States are under any obligation to “roll up” data over two or three years in order to create a minimum “n” sufficient to make an AYP determination for any subgroup otherwise too small for a determination to be made at the school or district level. While some States have provided for this in their AYP models, others have not and ED has not issued any related guidance or policy decisions.

Enhancing Reliability—“Safe Harbor” Determinations

NCLB includes at sec. 1111(b)(2)(I)(i) provisions providing for a further review of any group or subgroup’s progress when it appears that they may not have met the State’s AYP requirements. This review, commonly referred to as a “safe harbor” review, is based on recognizing decreases of at least 10% in the number of the group or subgroup’s students who fail to meet or exceed the proficient level of academic achievement on a State’s assessments

ED rejected **North Carolina’s** proposal to apply its minimum “n” to each of the two years considered under “safe harbor.” That is, the minimum “n” would apply only to the current year and results would be compared across the two years regardless of the size of the previous year group. Although “safe harbor” will no doubt prevent some schools from

being identified as being in need of improvement for a time, States may need to exercise particular care in how the “gains” are being interpreted, keeping in mind not only the inherent instability of gain scores but also that the two sets of compared scores are based on different students and perhaps on groups that differ widely in size.

However, **Ohio** received approval to use three-year averages for “safe harbor” reviews. The State will average the most recent three years of test scores (including current year scores) and compare the results to the current year’s assessment results. The higher score will be used to determine whether the school, district, or subgroup achieved the necessary ten percent reduction from the previous year in order to satisfy AYP requirements. For some subgroups, schools, or districts the three-year average will be applied; for others, the most recent results will be applied.

In **North Dakota**, “following a study of the effects of statistical reliability on safe harbor and with the joint concurrence of the State and the U. S. Department of Education, the State will employ the binomial distribution statistical method within the calculation of safe-harbor status for subgroups.” The State notes that a statistical reliability model for safe harbor has not been tested or validated. **North Dakota** has proposed a study, which will take several months, on the effects of adopting the binomial distribution for safe harbor reviews.

In **Nevada**, the impact on 2002-03 AYP classifications using CIs for relative growth comparisons (“safe harbor” determinations) will be jointly studied by ED and the **Nevada** Department of Education. The State will be using a one-tailed 75% CI in “safe harbor” reviews for the 2002-03 school year and a 95% CI interval for percent proficient determinations.

Utah was also approved to use multiple years of data for “safe harbor” determinations (see Critical Element 3.2 of the State’s workbook). For the 2003 school year only, **Utah** will use a statistical test employing a one-tailed alpha of 0.25. According to the State’s workbook, “Data and results will be submitted to the U.S. Department of Education for further review and discussion. Based on that discussion and Department approval, it is Utah’s intention to employ a test of statistical significance using a one-tailed alpha of 0.01 for groups with $N > 10$ for two consecutive years.” The details of the State’s multi-year plan follow below:

- In the first year of NCLB implementation, reduction in percent not proficient (improvement) will be compared to the baseline year. The LEA, school, or student subgroup will make AYP if the null hypothesis is *not rejected*.
- For the second year of NCLB implementation, improvement will be measured from the previous year and from two years previous. Any school or subgroup will make AYP if (a) the null hypothesis is *not rejected* at the 0.01 level that the portion of students not proficient has been reduced by 19 percent over two years OR (b) the observed portion of students not proficient over the past year has been reduced by 10 percent. The test of statistical significance will be calculated on the two-year data only.
- For the third and *all subsequent* years of NCLB implementation, improvement will be measured from the previous year, from two years previous and previous three years. The LEA, school, or student subgroup will make AYP if (a) one *does not reject* the null hypothesis at the 0.01 level that the portion of students not proficient has been reduced by 27.1 percent over three years, OR (b) the

observed portion of students not proficient over the past two years has been reduced by 19 percent, OR (c) the observed portion of students not proficient over the past year has been reduced by 10 percent. Note that the test of statistical significance will be calculated on the three-year data only.

Oregon received approval to use a 99% CI in making percent proficient determination for AYP. **Oregon's** system acknowledges that while the reliability of measuring improvement over one year is low except for large subgroups, it becomes substantially higher if improvement is examined over two years and even higher over three years. At the same time, the system recognizes that an LEA, school, or student subgroup will make AYP if it can show that performance has substantially improved in the most current year(s). Therefore, the first test in each case is one of statistical significance for improvement over the longest period of time. If the LEA, school, or subgroup fails that test, it still can make AYP by showing substantial growth, but it no longer has the advantage of statistical uncertainty—the observed results must have increased by the required amount or it fails to make AYP.

Texas has proposed an alternative definition to safe harbor and is awaiting final approval from ED contingent upon submission of data showing a high correlation between the two methods. The State's' accountability plan contains the NCLB "safe harbor" definition and, "alternatively, for all students and each student group that fails to meet the performance standard on the assessment measure, AYP performance requirements are met if there is (1) improvement on the assessment measure at a rate that, projected forward, puts the school or district on target to reach proficiency standards by 2013-14 and (2) improvement on the other performance measure. This alternative to 'safe harbor' is intended to address the importance of each school and district making sufficient gains on all measures to reach 100% proficiency by the year 2013-14. Like 'safe harbor,' it requires that improvement be made on the other performance measure in order for the provision to take effect."

Massachusetts and **Pennsylvania** made similar proposals regarding the use of individual school improvement targets for "safe harbor" reviews. While the former's proposal has been approved by ED, the latter's is still in negotiation. In **Massachusetts**, in order to make AYP, schools and subgroups must demonstrate student performance above State targets for the time period in question or show improvement at a rate that, projected forward, puts the school "on target" for getting all students to proficiency or above by 2014. School and district performance is assessed using a proficiency index that measures the extent to which students have achieved or are progressing toward proficiency in ELA/reading and mathematics. The proficiency index (a measure of students' proximity to achieving the proficient level on the MCAS tests) was designed to assume the role scaled scores previously played in the State's school and district accountability system.

Pennsylvania is working to develop a methodology that can be mutually satisfactory to both the State and ED that will permit incorporation of its performance index in "safe harbor" reviews in conjunction with 2002-03 AYP determinations. **Pennsylvania's** process would permit use of a performance index in which each school sets its own trajectory for growth (similar to Massachusetts and Iowa) using their own 2002 baseline towards 100% proficient in 2014. This trajectory would be applied for "safe harbor" determinations when schools did not meet the State AMOs for accountability measures. The State is committed to measuring both absolute achievement levels and growth

arguing it is crucial for progress to be determined in a way that is sensitive to academic growth all along an achievement scale. **Pennsylvania** currently uses the performance index approach in making awards and informing technical assistance strategies.

Enhancing Reliability—Standard Error of Measurement (SEM)

Some States also chose to consider the standard error of measurement (SEM) in their AYP determinations. Although theoretically appropriate, it should be noted that in most cases the SEM will be much smaller than the 95% CI, not to mention the 99% CI, so its application may not have significant implications when the two strategies are combined.

Massachusetts applies a 2.5 point error band for relatively large groups and calculates a custom error band, which may be as large as 4.5 points, for smaller groups. **Wisconsin** also received approval for its plan to examine various cell sizes to maximize the percent of schools considered under regular AYP conditions while minimizing the error in making AYP decisions using a one-tailed z-test¹². A minimum “n” cell size of 40 includes over 66% of the schools (based on one year’s results) and maintains an error of less than 10 percentage points in that State. **Wisconsin** applies a student-level application of SEM associated with the proficient scale score for each content area. To obtain a comparable error band for schools below the minimum “n” size, student level records will be used to calculate the standard error for the school.

In a variation on the use of SEM, **Illinois** received approval of its methodology to use a fixed minimum “n” of 40 for AYP determinations with a 3% error band. The State indicated that it believed that its “n” size of 40 would produce reliable decisions but it also presented evidence regarding how the fixed “n” is biased against schools with percentages of students closest to AMOs. For example, when the NCLB AMO required 55% of all students proficient, a value of 52% proficient would be used to judge AYP in each subgroup.

Enhancing Validity—Opportunity to Review and Present Evidence

NCLB includes in Section 1116 provisions for schools and districts to request further review and the opportunity to present additional evidence whenever they believe that a determination of failing or make AYP or identification for improvement has been made in error or is otherwise unwarranted. For the Peer Reviews, States were asked to describe, in Critical Element 9.2 of their accountability workbook, their plans and procedures for schools and districts to appeal accountability decisions.

There was a wide variation among State responses to Critical Element 9.2. Some States provided very little detail on how requests to review and present evidence from school districts would be handled at the SEA level or even whether they intended to provide structure and guidance to LEAs on the handling of reviews from schools (as provided for in the law). Other States have developed sophisticated, systematic procedures for handing of all requests—whether from schools or districts. While this is an area of the accountability workbooks that did not seem to receive a great deal of attention from ED, it is an area that could prove troublesome for States and school districts. The need for systematic, uniform, and objective processes/procedures for

¹² Wisconsin chose the minimum “n” size based on both including as many schools as possible and reducing the percentage of error in making determinations—hence a CI type approach. However, only the SEM is used in determining AYP.

receiving and acting on evidence submitted by a school or district questioning an identification for improvement seems quite self-evident.

Inclusion

NCLB requires the inclusion of all enrolled students in both the assessment system and the accountability system. However as reflected in the earlier reviews of States' standards and assessments by ED, how students participate in assessments varies greatly across States. In completing their accountability systems designs, States also adopted varying strategies for how to include some groups of students in their AYP determinations. This section first considers how students in general are included via States' definitions of full academic year (FAY) and student tracking systems. It then discusses the inclusion of students with disabilities and students whose English proficiency is limited in State accountability systems.

GENERAL INCLUSION

Full Academic Year

Under NCLB, Percent Proficient calculations are to be based on the performance of students enrolled (a discussion of enrolled versus tested follows later in this paper) for a full academic year (FAY); however, all students enrolled at the time of testing are required to take the assessments regardless of how long they have been in the school or district. Each State defines its own FAY, with the restriction that the definition cannot encompass a period of more than 365 calendar days (see Critical Element 2.2 of the accountability workbook).

Defining FAY proved to be somewhat complicated for several States. To begin with, the definition cannot be considered in isolation of other factors, such as multiple testing windows, widely varying school year beginning and ending dates, and year-round schools. The definition is also affected in practice by a State's capacity to track and report student enrollments.

With regard to the issue of variable testing windows, many States plan to use enrollment counts at two points, generally a "snapshot" about September 30 or October 1 as the first and the testing window or a date near to it as the second. With this type of definition, FAY could be inconsistent across districts in States with non-standard testing windows. In these situations, a State could impose a specific date for the collection of the second enrollment data point that is not related to the testing window and could, therefore, be consistent across districts. However, in most States this would mean that the second point would likely impose a new data collection burden and would not be associated with the same compliance incentives as the first point (in most States, the fall snapshot is the one on which fiscal allocations are based).

Hawaii, Iowa, and Colorado are among the States that have defined FAY as enrollment from one test administration to the next. Depending on the exact definition of FAY, a student would have to be present for each of two consecutive annual test administrations and, in some cases, also have been continuously enrolled during the interim.

- **Iowa** defines FAY for each individual student based on enrollment the first day of the testing period for ITBS or ITED in the previous school year and enrolled through the academic year to the first day of the testing period for current year. In this State, districts may select from three test windows annually—fall, winter, and spring.
- **Hawaii** and **Colorado** both require continuous enrollment between assessment administrations. While **Hawaii** specifically notes that the FAY will comprise no more than 365 days, **Colorado** notes that the definition may require students to be enrolled for over 12 months.

In **Wisconsin**, a State that changed its assessment window from late winter to fall, FAY has been defined (since 1999) for State and Federal accountability purposes as follows:

- School—students continually enrolled in a school during the annual fall census of the prior year to the current year (12 months). For students that move together from one school to the next at transitional grades (often beginning of 3, 5, and 9), they are enrolled for a full academic year if they have been in the district FAY.
- District—students continuously enrolled in the district from the fall census of the prior year to the current year (12 months).

Tracking Back of Students

The final regulations on accountability (December 2002) clarified in §200.13(c) that States need to make annual accountability determination for each school. Of particular concern here is how States will make accountability determinations in schools containing grades not covered by the State assessments (e.g., K-2 schools) and small schools where minimum “n” applications result in the schools falling outside of the AYP model. These regulations also indicated that ED would provide non-regulatory guidance to assist States in this matter; that guidance has not been issued to date.

One of the options open to States is “tracking back of students.” **Delaware** received approval for a tracking model that relies on apportioning student results. In this State’s model, “when students take the grade 3 assessment, provided that the student was in the school for a full academic year, then: the school that provided Kindergarten services gets 10% of the score; the school that provided first grade services gets 30% of the score; the school that provided second grade services gets 30% of the score; and the school that provided third grade service gets 30% of the score. For the grades 4 and 5 content standards, 50% of the score goes to the school that provided fourth grade service and 50% to the fifth grade school. For the grades 6 through 8 content standards, one-third of the score goes to each of grades 6, 7, and 8. For the grades 9 and 10 content standards, half of the score goes to each of grades 9 and 10.” This model will change in 2005-06 when reading, language arts, and mathematics in grades 4, 6, and 7 become part of the State’s AYP model. At that point, the grade 3 scores will still be apportioned back to K through 3 at the ratio stated above. For grades 4 through 8, 100% of the score will be apportioned to the single grade and at the high school level, AYP will be based on grade 10 assessments. This approach meets the requirements for annual accountability determinations for all schools described at the introduction to this subsection.

INCLUSION OF STUDENTS WITH DISABILITIES

Out-of-level Testing

Several States allow students to take assessments for grades below the ones in which they are enrolled, usually based on the specific recommendation of a student's Individualized Education Plan (IEP) committee. In the early Peer Reviews, ED rejected this practice, saying that all students had to be assessed on the standards for the grade in which they are enrolled. However, in later reviews, ED did not close the door on out-of-level testing, provided that the results are counted as "non-proficient." But, on June 27, in a letter to Chief State School Officers, Secretary of Education Rod Paige reversed all earlier ED decisions related to out-of-level testing for the 2002-03 year only. Referring to out-of-level assessments as "Instructional Level Assessments or ILAs," the Secretary stated that

...if a State permitted the use of ILAs during the 2002-2003 school year to measure the progress of other students with disabilities [other than those with the most severe cognitive disabilities for whom proposed regulations would permit the use of alternate assessments provided that the percentage of students held to related alternate standards does not exceed 1% of all students in the grade assessed] based on their IEPs, the State may hold schools and districts accountable for the achievement of these students against instructional-level standards rather than grade-level standards. This policy only applies to assessments that were administered during the 2002-2003 school year, which will be used to make AYP determinations for the 2003-2004 school year. The 1.0 percent limit...does not apply to these test scores.

The Secretary's letter did not address related issues for those States that had been told they could not use out-of-level assessments to meet NCLB assessment/accountability requirements or that decided not to provide such assessments believing they were prohibited by regulation. The decision seems to have clearly, at least for 2002-03 AYP determinations, resulted in the following ways that SWDs can participate in a State's AYP system:

- Participate in a State's "regular" assessments with or without accommodations;
- Participate in an alternate assessment that is aligned with the State's academic content and student achievement standards for the student's grade level;
- Participate in an instructional-level assessment (or out-of-level assessments as they are more commonly known) based on instructional-level standards (a term not common to NCLB); or
- Participate in an alternate assessment reserved for the most severely cognitively disabled students based on State academic content and student achievement standards reflective of professional judgment of the highest possible learning standards possible for these students. Not more than one percent of all students in the grade assessed may be scored at proficient or better against these standards.

States affected by the Secretary's decision may wish to study carefully related data before deciding whether to apply ILA proficiency determinations to 2002-03 AYP decisions. On the one hand, applying these may have the effect of identifying fewer schools for improvement. On the other hand, applying these may serve only to postpone identification for improvement by one year and create what might appear to be a substantial regression in this subgroup's performance in many schools when comparing 2003-04 assessment results to those from 2002-03.

With respect to those States that proposed out-of-level testing in their accountability workbook plans, **Delaware** (an early Peer Review State) proposed limited out-of-level testing for SWDs at the end of grades 5, 8, and 10 with a policy that students participating in these tests would receive only a scale score and instructional needs comments. This meant that the students would be rated as "well below" the standards for accountability purposes. ED responded in mid-March that out-of-level testing is "inconsistent with the statute and regulations [out-of-level testing for SWDs was a contentious issue during the negotiated rule-making for the standards and assessment regulations and is prohibited under those regulations]." **Washington**, another early-review State, was also advised by ED that they could not use out-of-level testing to meet State assessment requirements and make AYP determinations.

South Carolina and **Louisiana**, on the other hand, have approved plans allowing out-of-level testing as long as the student receives a proficiency score based on the grade level achievement standards, likely resulting in a non-proficient score. Further, students participating in the out-of-level assessments will be counted as having participated in the statewide assessments, thus including them in the AYP participation rate calculation.

In **South Carolina**, State law mandates the opportunity for "off-grade" assessments and, according to State officials, approximately 5% of all students take "off-grade assessments" that are aligned to State standards and offer useful information to teachers. **Iowa** argued that the ITBS is vertically scaled and has linked standards, so grade-level interpretations can be made for students tested out-of-level. ED rejected out-of-level testing below grade for **Iowa** but approved it for above grade level. The State indicated that 2 to 3% of students are tested of out-of-level.

Other States, such as **Utah**, have decided to discontinue the use of out-of-level testing since the practice does not yield results for the grade in which the student is enrolled.

Alternate Assessments

Since 1994, ED has allowed States to use an alternate assessment for SWDs unable to take a State's regular assessments with or without accommodations although neither NCLB nor IASA statutes specifically described such tests. ED's guidance for standards and assessments under IASA did provide for alternate assessments as follows: "For a small number of students with disabilities, the severity of their physical or cognitive limitations prevents them from participating meaningfully in exactly the same assessments as other students, even with the availability of appropriate accommodations. For this small population of students, appropriated alternatives should be used to assess their educational progress [1996, p. 43]." In implementing related provisions of IASA, ED required States to ensure that guidelines were developed for the participation of SWDs in alternate assessments (for those who could not participate in the regular

assessments with or without accommodations). IDEA also required States to develop alternate assessments by July 1, 2000.

The final regulations for standards and assessments under NCLB (July, 2002) address alternate assessments for SWDs in §200.6(a)(2)(i) providing that, “the State’s academic assessment system must provide for one or more alternate assessments for a student with disabilities as defined under section 602(3) of the IDEA who the student’s IEP team determines cannot participate in all or part of the State assessments under paragraph (a)(1) of this section, even with appropriate accommodations.” While the IASA guidance reflects the position that alternate assessment of SWDs be restricted to a very small number of students with severe cognitive or physical disabilities, the new regulations appear to suggest that States could have more than a single alternate assessment for SWDs (which would be consistent with the current proposed regulation covering a limited number of the most severely cognitively disabled students).

In August of 2002, ED issued proposed regulations allowing States to also offer an alternate assessment directed solely at students with the most severe cognitive disabilities but limited participation in this alternate to not more than 0.5% of the total student enrollment. States could continue to provide, without limitation, an additional alternate assessment to other SWDs unable to participate in their regular assessments with or without accommodations (consistent with applicable law and regulations). This proposed regulation was dropped prior to final promulgation of the accountability regulations in December 2002.

In a March 20, 2003, Notice of Proposed Rule Making (NPRM), ED proposed a new rule that would permit States to measure the performance of “students with the most significant cognitive disabilities” against **alternate achievement standards** with a limit of 1% of all students assessed in a State or school district whose results may be included in accountability measures. The alternate assessment standards must be aligned with the State’s academic content standards and reflect professional judgment of the highest learning standards possible for those students.

The final accountability regulations also indicated that ED would provide nonregulatory guidance to assist States in this matter; to date, that guidance has not been issued. The rule included criteria defining students eligible to participate in this particular alternate assessment. In responding to questions about this proposed regulation, ED has consistently stated that there is no limit on the number of SWDs who may take this alternate (assuming eligibility criteria are met). However, the participating students are counted in AYP in two different ways—(1) all students taking the alternate are counted as participating (FAY cannot be applied to Participation Rates) and (2) the State can choose to calculate the number proficient as either the number proficient and advanced divided by the number enrolled FAY or the number tested FAY. The number counted as proficient or advanced cannot exceed 1% of a district’s or State’s total public school enrollment (although the proposed regulations do provide at §200.13(c) for States to request from the Secretary and LEAs to request from States an exemption to this limit). States that have received ED approval of their accountability systems have been advised that they can follow the NPRM language for 2002-03 only—as a transitional year—while the regulation is being finalized.

A related issue that does not seem to have received much attention yet is how to handle a situation in which more than 1% both take the alternate and score proficient or

advanced. Would the “excess” scores have to be reported as below proficient? Assuming that these students’ scores would have to be reported below proficient, would it be at the most basic level of proficiency or could it be at a level just below proficient assuming the State has two or more levels below proficient (this can be an issue for States with weighted or indexed formulas for calculating AYP)? How would the “excess” participants be included in the determination of Participation Rates? In **California**, the scores of SWDs above the 1% cap will be considered “far below basic.”

There also appears to be a consensus now among States and ED that the definition of “students with the most severe cognitive disabilities” was set too stringently. The original requirement was that only students who score at least three standard deviations below the mean on an IQ test could be considered “severely cognitively disabled”—the result being that a much smaller percentage than 1% of all SWDs would meet the definition. Further, the definition seems based on an “IQ like” standard deviation theory as opposed to functionality, a perspective that predominates among special educators. With regard to this issue and to out-of-level testing, States have questioned whether ED can require States to disregard the recommendations of IEP teams when they involve the use of alternate assessments for students not meeting the three standard deviation criterion or out-of-level assessments when considered appropriate.

In addition to an alternate assessment for the most severely cognitively disabled students, some States also use a separate alternate assessment for less severely disabled students for whom the regular assessment is still considered inappropriate. ED initially rejected this concept (although that does seem somewhat contradictory to the July 2002 regulations on standards and assessments). However, ED has given varying responses since publishing the March 2003 NPRM. For example, **Oklahoma** currently has an alternate performance-based assessment for SWDs with severe cognitive disabilities. Participation in this assessment is about 3/10 of 1%, although the State does anticipate a slight increase with full implementation. **Oklahoma** was informed that it may be possible to set different achievement levels for a “second alternate” it wants to develop for SWDs with “moderate” cognitive deficiencies as long as this assessment is based on the same academic standards as the regular assessment. The State is developing their second alternate assessment based on the same standards but with items requiring less extensive knowledge.

In **Texas**, SWDs unable to participate in the State’s regular assessments with or without accommodations take either the State-Developed Alternative Assessment (SDAA) or Locally-Developed/Determined Alternative Assessments (LDAAs) if they are unable to participate in the SDAA. The local Admission, Review, and Dismissal Committee determine which assessment is appropriate for SWDs based on a student’s LEP. ED approved the State’s proposal to count SWDs taking either the SDAA or LDAAs as non-participants and to exclude them from the calculation of percent proficient. How this decision will affect the determination of AYP in the State’s schools for 2002-03 is not yet known. Further, consequences for the participation rate calculations are likely mitigated as well by the State’s 50/10%/200 minimum “n” size for participation (see details earlier in this paper). **Texas** indicated that it intends to further study this subject in 2004.

Continuing to Include “Exited” SWDs in AYP Determinations

ED’s position has generally been that students can only be considered within the SWDs subgroup for Title I accountability and reporting purposes if they are receiving services under Section 602 of the Individuals with Disabilities Act (IDEA); students who once received special education services, but who no longer require these services, could not be included in the SWDs group. Some have argued that exclusion of these exited students fails to recognize program “successes,” negatively affecting AYP determinations by excluding those who benefit the most from the programs and services. **Mississippi** originally submitted a plan for continuing to include “exited” SWDs in the subgroup beyond the point of service; this was dropped from the State’s approved plan.

However, **Georgia** later won approval for a proposal in which students continued to be included in the SWDs subgroup provided they are still receiving services either in the form of monitoring or support in the transition to the regular classroom. Monitoring and support have historically been considered an SWDs service in some States who have consulting teachers, a model that could be revived under these conditions. The consulting teachers are special educators who assist regular classroom teachers in smoothing the transition of students into the regular classroom setting. With the impact of “highly qualified teachers” on special education, especially at secondary level, most States may have no other choice but to revive this practice, if allowed under the new IDEA reauthorization.

Further, for **Georgia** students who transition from special education and return to the regular education classroom, a student support team plan/Section 504 plan is developed to assist in monitoring student progress and to identify needed changes in the educational support plan as necessary. The use of the student support team plan/504 plan process provides oversight to support continued academic progress. While students are in this process, their State assessment accountability scores will continue to be included with the scores of the special education subgroup. Currently, plans for students with disabilities and LEP students in Georgia are being refined to clearly specify monitoring conditions. Also, program exit criteria and definitions incorporating the monitoring segment have yet to be developed. These plans will be in effect beginning 2003-04 pending appropriate in-State approvals.

Larger Minimum “n” for SWDs

Some States (e.g., **Nebraska, Oklahoma, Ohio, Wisconsin**) have established higher minimum “n’s” for accountability determinations with SWDs subgroups. **Ohio** was the first State to propose this strategy. This State set a minimum “n” of 30 for all proficiency determinations with the exception of the SWDs subgroup, for which the State will use a minimum “n” of 45. The State argued that there are measurement issues unique to the students with disabilities group. The larger subgroup size is designed to compensate for the heterogeneity of this subgroup, the extensive use of accommodations in assessing students with disabilities, and the substantial variation in identification rates for this population. **Ohio** provided significant supporting literature and rationale for the larger subgroup size. **Nebraska** will also use 30 for all students and all subgroups except for SWDs where the State will use 45 for minimum “n” determinations. As noted earlier, **Oklahoma** will use a minimum “n” of 30 for all students and 52 for all subgroup AYP determinations including SWDs. Similarly, **Wisconsin** will use a minimum “n” of 40 for

all students and a minimum “n” of 50 for SWDs, arguing in a paper included in its workbook, as **Ohio** did, that the subgroup presents unique measurement issues.

LIMITED-ENGLISH PROFICIENT STUDENTS

Accommodations and Alternate Assessments

Review of the State accountability workbooks also underscored that inclusion of students with limited English proficiency (LEP) in State assessment systems continues to be a challenge for most States. Many struggled with ways to ensure the meaningful assessment of LEP students’ proficiency in reading or language arts and mathematics. Very few States yet provide assessments in one or more languages other than English. States with English-only laws, such as **Indiana**, are using modified alternate assessments developed originally for SWDs. Even when a State offers an alternate assessment, combining the results from this assessment with the results from the regular assessment can be a challenge; several States proposed a “rubric” to compare and combine the results of LEP assessments to results from their standard assessments. More recently, some States have received funding to develop assessments more appropriate to the needs of LEP students.

Illinois is using its Illinois Measure of Annual Growth in English (IMAGE) as both an LEP alternate test of State academic content standards and an English language proficiency assessment (something IMAGE was originally designed to do) under Title III. ED approved this for the State because the upper level of IMAGE is aligned to **Illinois’** content standards and the lower levels of English language acquisition proficiency. The key to using an assessment for both purposes is that the test has to completely measure both the standards and language acquisition proficiency.

Continuing to Include “Exited” LEP Students in AYP Determinations¹³

Classification of students into the LEP subgroup in a way that makes sense under NCLB AYP rules raises a conundrum. The clear goals of NCLB AYP are to get all students to proficiency in reading or language arts and mathematics by 2013-14; progress toward these goals is how schools are judged each year. Yet, by definition, LEP students are not proficient in reading on the State’s assessments; sec. 9101(25) of NCLB defines an LEP student as one—

- “(D) whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to **deny the individual** —
- (i) **the ability to meet the State’s proficient level of achievement on State assessments described in section 1111(b)(3);** [emphasis added]
 - (ii) the ability to successfully achieve in classrooms where the language of instruction is English; or
 - (iii) the opportunity to participate fully in society.”

So, it is impossible for 100% of this subgroup to reach proficiency. For schools and districts that serve a substantial number of LEP students, this imposes a ceiling on the performance of this group; at some point it will be impossible to make AYP simply due

¹³ This paper uses the term “exited” to describe students who are no longer receiving structured LEP services directed at their English language acquisition. Technically, students are not “exited” from LEP programs until they meet all of a State’s criteria for that status.

to how this subgroup is defined under law. Thus, several States have taken a fairly unique approach to the notion of continuing to include “exited” LEP students in the subgroup by carefully defining “exit” criteria. As noted below, in order for LEP students to continue to be included in AYP determinations for this subgroup, they must not have met the State’s criteria to be exited from the program.

Some advocates have suggested that States use a “once LEP, always LEP” classification rule for this subgroup in making AYP determinations under NCLB. This would allow AYP calculations for the LEP subgroup to also account for the success of instructional programs for LEP students. The authors are not aware of any State that actually proposed this and how ED might respond to such a plan is unknown. However, a number of States did successfully advance strategies that will permit the continued inclusion of “exited” LEP students in LEP subgroup AYP determinations for at least a few years.

Indiana and **Delaware** will include “exited” LEP students in the LEP subgroup for two consecutive years beyond attainment of English language proficiency. These States refer to Title III, sec. 3121, requiring States to monitor LEP students for two years following LEP services and cite the need for a student to achieve a proficient score on State assessments more than once to enhance classification accuracy. Since ED has said that a student must still be receiving the LEP services to be considered within that subgroup, these two States have also successfully argued that monitoring is a service. **Indiana** will include students in the LEP subgroup until they score proficient for two consecutive years on the assessment of English language proficiency, arguing that a single score is not reliable enough to support a high stakes decision, like program eligibility, for individual students.

Ohio also intends to include, within that subgroup’s AYP determinations, LEP students for whom the school or district is “monitoring” English language acquisition. The State must provide ED, as a part of its final approval process, the criteria it will use to determine whether a student is no longer LEP and what it means to be monitored and for how long the monitoring will occur.

Georgia plans to continue including the scores of LEP students in that subgroup even after these students meet the exit criteria for the English for Speakers of Other Languages (ESOL) program as long as they are receiving monitoring and/or direct services through the ESOL program.

California will include LEP students within the LEP subgroup until they score proficient on the California Standards Test (CST) in reading/language arts for three consecutive years; these students will not have to take the State’s English language proficiency test during this period. This unique allowance is possible because the State specifically worked to align its CST in reading/language arts with its California English Language Development Test (CELDT); LEP students can be re-designated as English proficient when they reach either the basic level on the reading/language arts CST or the proficient level on the CELDT. However, should a single score in the target level on either test be enough to re-designate students as English proficient, then all LEP students would, by definition, achieve below the proficient level on the achievement test: the AYP target of 100% proficiency could never be achieved for this group.

In its argument, **California** makes specific reference to the Title IX definition of LEP students, “difficulties in speaking, reading, writing or understanding the English language [that] may be sufficient to deny the individual the ability to meet the State’s proficient level of achievement on State assessments described in section 1111(b)(3.) Since **California** re-designated formerly LEP students have already achieved proficiency on the CELDT, the State will not continue to assess these students in English language proficiency. The State has received written confirmation from ED of this practice. **California’s** definitions of LEP are consistent across both Title I and Title III.

South Carolina defines as LEP for Title I and Title III purposes as a student “who has a primary language other than English and is not proficient in listening, speaking, reading, writing, or comprehension in the English speaking classroom as determined by a language assessment instrument” (testing proficient for three years is required to exit LEP status). **South Carolina** has set the criteria to exit LEP status as students:

- No longer meeting the definition of LEP;
- No longer participating in ESOL classes nor receiving mainstreamed services (one to four hours of instruction per week of supplemental English language services);
- Who have tested proficient on the language proficiency test for three years consecutively; and
- Who have tested proficient once, at minimum, on the State's PACT assessment.

A few other States made similar proposals. ED has not issued a policy paper on this but now appears to accept this practice depending on a State’s rationale and given arguments about monitoring and providing some form of continuing services. States considering this approach should carefully review the definition of LEP provided in sec. 9101(H) of the law, develop program “exit criteria” and provide for on-going monitoring of the progress made by “exited” LEP students, which should include the provision of any additional assistance or support they may need. States should also consider related factors such as the increased subgroup size, continued English language acquisition assessment that may or may not be required under Title III, and the possible impact on related categorical program funding if they are thinking about a modification in their LEP definition.

Starting Points, Annual Measurable Objectives, and Intermediate Goals

Annually, a State must determine whether each school and district met the yearly Annual Measurable Objectives (AMOs). The school or district’s AYP status (met or did not meet the AMOs) for a given year must be compared with its status for the previous year to determine whether the school or district will be identified for improvement.

Some States proposed to make this comparison specific to a subject and subgroup. That is, in order for a school to be identified for improvement, the same subgroup would have to miss the AMOs twice in the same subject. However, ED consistently took the position that schools and districts not meeting the AMOs in the same subject area for two consecutive years must be identified for improvement, regardless of the patterns at the subgroup level. So, if all subgroups except students with disabilities meet the reading AMO in 2003-04, and all subgroups except limited English proficient students meet the reading AMO in 2004-05, the school is identified for improvement. What is not yet clear

from the reviews and “approvals” is what happens when a school (a) meets its AMOs and goals for the other academic indicator, but fails to make the Participation Rate goal, one year and then (b) meets its AMOs and the Participation Rate goal, but fails to meet the goal for its other academic indicator, for the subsequent year. Is this school identified for improvement? Both the statute and the regulations are silent on these matters and, as noted earlier, ED has not issued any related guidance.

STARTING POINTS, AMOs, AND IGS BASED ON OTHER THAN STATE AVERAGES

Most States set starting points, AMOs, and Intermediate Goals (IGs) that will apply at the State, district, and school levels, statewide. **Virginia** wanted to set separate starting points and trajectories for the different subgroups under NCLB. ED rejected this proposal. **Illinois** proposed to vary the size of the increments for their IGs (referred to as the “Illini Plan”). ED has been consistent in requiring that the size of the increase (“rise”) must be constant, although the length of the “run” can vary from one to three years; thus, ED did not approve this aspect of **Illinois**’ AYP model. **Illinois** modified its AYP model and the final submission includes the Illinois Equal steps plan, which was approved by ED.

Iowa will permit each LEA to set the starting points, AMOs, and IGs for itself and its schools, although the State will still determine State starting points, AMOs, and IGs. According to Iowa officials, “the state’s trajectory is the point of comparison for all organizational levels, grade levels, and subgroups. Schools that are achieving above the state’s trajectory, and whose rate of change is on track with their own trajectory are okay. Schools that are achieving above the state’s trajectory, but whose rate of change is not on track with their trajectory, will be required to submit a corrective action plan with the [Iowa Department of Education]. If a school does not achieve at the level of the state’s trajectory, but their rate of change is on track for 100% in 2014, they do not make AYP.” Elsewhere, **Iowa**’s plan indicates, “For any school or subgroup that falls below the state’s trajectory, the NCLB process will take precedence. In this way, an LEA must first meet the state’s trajectory for NCLB for all subgroups, then meet their own trajectory for Iowa regulations” (Appendix A, *The Iowa Model*, p. 40 of the Iowa Accountability Workbook).

A NOVEL APPROACH TO DETERMINING IGS

New Jersey received approval for a somewhat novel approach to determining IGs. The State successfully advanced a methodology employing equal increments of growth calculated on a percentage rate. Under this approach, a growth model will be employed based on the Compound Annual Growth Rate (CAGR), a formula often used to calculate interest rates for investments. CAGR is used when given a present value, how much would the investment have to grow in a given amount of time to achieve a set future value.

$$\text{CAGR} = (\text{FV}/\text{PV})^{1/n} - 1$$

In this formula, FV is the future value, PV is the present value, and “n” is the number of years. The interest rate, or growth rate, is constant over the number of years. The result for AYP purposes are AMOs that fall along an accelerating curve, similar to the patterns

resulting from using 3-year increments in the beginning of a cycle and single-year increments at the end (e.g., **Ohio's** pattern).

ESTABLISHING STARTING POINTS, AMOS, AND IGS IN TIMELINE WAIVER STATES

During the accountability plan reviews, ED signaled that States may use data from the first year of testing under a final assessment system rather than data from 2001-02 if those data are based on “old” assessments. **Ohio** was advised that they may set starting points for their new grade 10 assessment on the basis of the first testing in 2002-03 or set the high school starting points on the basis of the 2001-02 ninth grade assessment. In either case, the State must ensure that schools are identified for improvement prior to the beginning of the 2003-04, school year, and the requirement for all students proficient by 2013-14 cannot be delayed.

At least one State (**Wisconsin**) has already chosen to establish these points/requirements based on 2001-02 data, arguing that the “old” and “new” assessments are closely linked.

Participation Rate and Other Academic Indicators

NCLB expanded previous ESEA AYP indicators to include Participation Rate and Other Academic Indicators.¹⁴ As they approached these requirements in designing accountability systems, States posed a wide range of methods aimed at calculating the participation rates and other academic indicators. Many of these are described below.

PARTICIPATION RATE

Although the NCLB requirements related to calculating Participation Rate seem fairly clear on initial reading, a variety of issues surfaced as States addressed this part of school and district accountability determinations. Among the issues were

- Whether the rate could be calculated based on only students enrolled for a full academic year;
- How to calculate the rate when multiple assessments are involved;
- How to calculate the rate when the assessments are offered at multiple points;
- Including in the numerator enrolled students who do not take the assessments; and
- Determining a Participation Rate minimum “n”.

Use of Enrollment for a Full Academic Year

Some States proposed calculating Participation Rate on the basis of their definition of students enrolled for a full academic year (FAY). In this approach, the number of students who took the test and have been enrolled for a full academic year would be divided by the number of students enrolled for a full academic year. This appears to conflict with the provisions of sec. 1111(b)(I)(ii) and conflicts with an internal ED Policy

¹⁴ The accountability regulations require only that schools and districts meet, or make progress toward meeting, the State's targets for Other Academic Indicators. States have a number of options in this arena ranging from recognizing any progress toward the targets to meeting or exceeding the targets (as well as raising the targets over time).

Paper regarding participation in States' academic assessments. However, although **Delaware** originally proposed this definition and later removed it from their final plan, a few States continued to seek approval for this approach, arguing that sec. 1111(b)(I)(ii) includes a reference to paragraph (3)(C)(xi) that, in turn, limits consideration for AYP decisions to students who have attended schools within the LEA for a full academic year.

Calculating the Rate When Multiple Assessments Are Administered within a Content Area

Questions involving the calculation of Participation Rate play out in interesting ways in States that use multiple assessments within a content area. In most of these cases, local assessments are administered at different times during the year and comprise multiple components. In States with a single assessment for reading, students need only attempt that one test in order to count as having participated; an attempt generally means that the student has responded to at least a specific minimum number of items, (e.g., 15 or 20). An attempt does not mean that the student must have answered enough questions to earn a valid score, even in a sub domain. So, in States with multiple components, would a student who took—or maybe just attempted—only one of the three components of the reading assessment count as having participated? If that student were required to attempt all three components, this State would be imposing a higher standard for participation than most other States. To date, ED has not issued related policy determinations or guidance.

Calculating the Rate When Assessments Are Offered at Multiple Points

There are at least two issues related to differential timing of assessments across districts—(1) calculating the Participation Rate when States have two or more testing windows during the same school year and (2) calculating the rate when States test multiple components at different times of the school year.

At least one State, **Iowa**, provides three testing windows for school districts—fall, winter, and spring. In **Iowa**, annual participation rate information is collected for the building-wide or district-wide assessment program. **Nebraska** does not set dates for conducting its STARS assessments and districts offer assessments multiple times over the course of the year. To ensure that all districts have the opportunity to include all assessments, Nebraska defines full academic year as enrollment from the last Friday in September until the end of district assessments or end of the school year. For determining Participation Rate, school districts there report the number of students who took at least 75% of the assessments or were assessed on at least 75% of the academic content standards (whether, or how, the students not included in these reports are accounted for is unknown).

Some States, such as **New York**, use high school end-of-course exams or graduation exams for AYP determinations. In these States, students typically have several opportunities to take the assessments, sometimes beginning as early as grade 9. The questions with respect to calculating Participation Rate are self-evident. ED has not given specific directions for calculating the rate in these instances. However, **Michigan**, which permits early testing for 10th graders seeking “dual enrollment” at grade 11 to begin concurrently taking courses for college credit, plans to designate the number of students enrolled as the “universe” of students that are required to participate in the high school

assessment. The State's system of assigning a Unique Identification Code for each student allows the matching of the student's enrollment and that student's assessment score. High school results, including achievement and participation, will be report by each 11th grade cohort. At other levels of the State assessment system in **Michigan**, schools are required to administer the tests within a designated "window." The State then designates a single day within the window and uses its Single Record Student Database to determine the actual enrollment on that day and determine whether 95% of the enrolled students participated in the testing on that day.

Including Non-Participants in the Calculation of Participation Rate

Some States proposed counting students who were absent or otherwise not tested as having participated in the assessments, arguing that (1) State law requires testing all students enrolled, and (2) non-participating students were included in AYP determinations as having scored at the State's lowest student academic achievement level (i.e., these students lower the Percent Proficient indicator). There are two possible outcomes to this approach—the first is that the State will always report a 100% participation rate and the second is that "actual" participation rates will not be reported. ED's response to **Delaware** and **Colorado** regarding this matter is illustrative of its position. These States were advised that they needed to account for all enrolled students in the denominator and could not count any student without an achievement test score in the numerator. The State can—but does not have to— automatically assign a 'not proficient' score (in the absence of an actual test score) for enrolled, but non-participating, students when calculating the Percent Proficient, but can not count these students as having participated.

ED has also addressed this in an internal Policy Decision requiring that a State count in the numerator as "participants" only those students who actually complete some portion of the assessment; this is notably different from the information provided to **Delaware** and **Colorado**, which were told that a student must earn an actual score to count as participation. However, in a late approval decision, ED accepted **California's** plan to include all students who sat for the State's assessments as participants including students who fail to respond to enough items to generate a valid score. For accountability purposes, these **California** students will be considered "far below basic."

Varying Approaches to Setting Participation Rate Minimum "N's"

States have proposed a variety of distinct approaches to setting a minimum "n" for Participation Rate accountability determinations:

- First is the use of the same minimum "n" employed in making all other AYP determinations.
- A second is the use of a larger minimum "n" than for other AYP determinations to protect schools and districts from the effects of absences of a few students in small subgroups.
- A third is where a State opts not to remove this requirement because a school or subgroup is less than the minimum "n" but proposes to use a participation rate of "n" minus one. That is, a subgroup of 18 students would have to have 17 students

participating in the assessments to meet the AYP target. **Nevada** will use the latter technique for schools/districts with fewer than 20 students total or in a subgroup; otherwise, the 95% requirement would translate into 100% for these schools. This State argued that there are legitimate extenuating circumstances that can arise preventing a student from participating in or making up missed assessments and it would be unfair not to allow for these circumstances in small schools.

- A fourth is the use of a series of “n’s” for determining the rate. In **Idaho**, the minimum required Participation Rate for all subgroups, schools, and districts is 34. Their plan indicates that “for subgroups less than the minimum “n” the 95% assessment requirement will be applied to the LEA and state levels.” The State further indicates that, “for all districts, schools, and subpopulations the participation requirement will be reduced according to” a table. In the table, for “n’s” between 33 and 13, the number of permitted absences is 2; for “n’s” between 12 and 7, the number of permitted absences is 1; and for “n’s” between 6 and 0, the number of permitted absences is 0.

Alaska will not calculate Participation Rate for subgroups of students of 20 or fewer. For subgroups between 21 and 40, subgroups will have met the participation requirement if all but two students are assessed. For subgroups above 40, the rate will be calculated in the standard manner.

GRADUATION RATE

Taken at face value, the NCLB statute and regulations (including the related comments included in the December 2002 final accountability regulations) appear sufficiently straightforward in terms of definition and calculation. However, as in other central issues, States have posed many related questions for ED, particularly in terms of how to count “non-standard” diplomas, which “class” to count SWDs in who earn “regular” diplomas over a five-year period (or longer) consistent with their IEPs (especially when the diploma is not based on the same academic standards required for other students receiving “regular” diplomas), how to count early and late graduates, and how to calculate graduation rate in the absence of prior year’s data for all subgroups of students.

Inclusion of “Non-standard” Diplomas

A few States initially proposed counting as graduates students who receive diplomas or certificates such as those based on completion of a high school equivalency program or completion of a special education program. ED consistently indicated that inclusion of such diplomas/certificates, including GEDs, is not permissible.

Virginia offers four types of diplomas: standard, advanced studies, modified standard, and special. A student receiving any of these “is able to respond in the positive when asked if she or he has received a high school diploma” and a student with any of these diplomas is eligible to apply for federal tuition grants. The standard, advanced, and modified standard have specific course content requirements. The special diploma is awarded to students with disabilities who have met the requirements of their IEPs and do not meet the requirements for other diplomas. **Virginia** will not include the special

diploma in determining graduation rate, as deleting it from the NCES formula will have no negative impact on the graduation rate. The modified diploma requires 20 standard units of credit, and students pursuing a modified diploma must pass the English/reading and mathematics Standards of Learning tests. Students receiving the Modified diploma, under directive from ED, will not be included in the graduation rate.

Absence of Prior Years' Data

Prior to NCLB, many States had not collected graduation data that could be disaggregated by all the required categories. Absence of the required data means that States need to establish timelines for the collection of these data and a “phase-in” process to use in the interim. Until a State is able to disaggregate the data consistent with NCLB requirements, ED has signaled a willingness to consider the use of a “proxy” indicator for “safe harbor” reviews (absent approval to use the All Students graduation rate). But, typically, a State must add new data to the proxy each year until a full four-year rate is achieved.

Vermont, for example, will combine two years of data and apply a 90% CI for the interim Graduation Rate and Other Academic Indicator. Until the State can disaggregate Graduation Rate, it will use the grade 10 New Standards Reference Exam: Reading Basic Understanding for all subgroups at the high school level. The criteria for not making AYP will be having 15 percent or more students in Below the Standards and Little or No Evidence. Two years of results will be combined and a CI of 0.01 will be used. **Idaho** will use either the current language arts ISAT or student growth assessment (Compass Learning Program) as a proxy for Graduation Rate to disaggregate for “safe harbor” purposes.

Current Year or Prior Year's Graduation Data

Some States have indicated that they will use Prior Year's graduation data in AYP determinations. ED has approved this approach if the primary reason is that the State includes summer school graduates in its calculations; thus, final rates are not available in time to meet parent notification requirements related to public school choice and supplemental educational services.

Graduation in More or Less than Four Years

Oregon proposed to include graduates who take more than four years to graduate in both the numerator and denominator but did not include this in their final accountability workbook. The State recognizes that this does not change the resulting AYP determination but feels that it acknowledges and reinforces the effort not to lose students even if it takes longer for them to graduate.

Maine's System of Learning Results requires high levels of performance for the issuance of a high school diploma. The State anticipates that some students will require five years to complete the diploma requirements and received approval from ED to extend the timeframe for consideration of dropouts (allowing the NCLB accountability criterion here to align with **Maine's** accountability system). State law provides that, “The intent of the system of Learning Results is to provide the time that students need in order to meet the content standards. This may involve more or less than the typical four years

of secondary school.” The State will compare the number of students that entered grade 9 with the number that receive a high school diploma in accordance with State regulations by the fifth year after entering grade 9. Under law there, “Secondary students are eligible for extended years of study to complete the requirements of a diploma if they have not reached the age of 20 at the start of the school year...Extended study for students with disabilities shall be specified in the student’s IEP.” Students who receive a GED or Adult Education Diploma will not be counted as having received a high school diploma.

ED approved **Rhode Island’s** proposal to include in graduation rate calculations students who take less than four years to graduate. **Vermont’s** State Board of Education recently approved a five-year definition for purposes of school accountability. However, the State cannot accurately collect the data until 2005, thus deferring until that time their request for approval of a five-year graduation rate.

ED also approved **Georgia’s** graduation rate definition which states that a “standard number of high school years for students with disabilities will be determined by each student’s IEP team, even if the standard set exceed four years. However, **Georgia** will not include students seeking a Special Education Diploma” in the numerator for calculating graduation rate. SWDs taking more than four years to graduate, consistent with their IEP, and earning a regular diploma will be counted in the numerator for determining the graduation rate in that State. In **Nevada**, students with IEPs will be given up to seven years to earn a standard diploma. The State will not, however, recognize students with “adjusted” diplomas in calculating graduation rate (other than in the denominator). **Michigan** and **Idaho** include a provision for SWDs that allows the IEP team to determine the standard number of years for graduation.

Approval of **Georgia’s** graduation rate again underscores the importance of a sound rationale and adequate explanation. It also underscores the considerable range of related practice among the States. In some States, IEPs are written to standards (at the high school level) and if students meet the standards, they receive regular diplomas. In other States, SWDs can only receive regular diplomas if they meet the same standards other students meet. Otherwise, they receive an alternate diploma. This latter practice seems to more closely reflect NCLB regulations.

Inclusion of GEDs

ED consistently informed States that GED recipients may not be included in the numerator when calculating graduation rates.

OTHER ACADEMIC INDICATORS

Typically, States chose to use attendance rate as the other academic indicator at the elementary and middle school levels. A few States chose instead to use results from other assessments, such as writing or science. Others choices included reduction in below basic performance and performance increases in percent proficient. **Nebraska** will use its statewide writing assessment as the other academic indicator (implemented in 2001-02 for grade 4, 2002-03 for grade 8, and annually in grades 4, 8, and 11 beginning in 2003-04), although the State does note that some districts or schools may use science assessments in place of writing. **Georgia** will use attendance for 2002-03 but will allow each school district to select its other academic indicator from a menu beginning in 2003-

04. This choice will be in effect for three years, at the end of which time districts may select a different indicator from the menu or continue with the same one they have been using.

Validity and Reliability

Throughout the NCLB statute and regulations, there are numerous requirements that States are to ensure that their decisions, methodologies, and procedures are valid and reliable (according to Dale Carlson in Marion, et al [p. 21, 2002], there are 59 references to the phrase, “validity and reliability” in Title I of NCLB). In completing their accountability workbooks, States were required to address these requirements and provide substantiating evidence under Critical Elements 9.1 and 9.2. Peer Reviewers were asked¹⁵ to determine, for example, whether States provided sufficient evidence that

- Their use of an index (if using one) contributes to the reliability and validity of their accountability system (Reviewer Question A3);
- They have plans for conducting reliability/validity analyses of their minimum “n” decisions (Reviewer Question A6);
- Policy makers accept the balance between validity and reliability for their minimum “n” decisions (Reviewer Question A6);
- Each of their other academic indicators is reliable and valid for the intended use (Reviewer Question B2);
- Their approach to determining participation rates, if unable to include the total number of students enrolled in the tested grade, does not compromise the validity of the AYP decision (Reviewer Question C3);
- Their proposed methods for calculating AYP were developed to maximize the validity of the system (Reviewer Question E1); and
- Their approach will enable them to determine the reliability (decision consistency) for AYP decisions (Reviewer Question E2).

From interviews with a number of State representatives and reviews of the accountability workbook “approval” letters received by several States, it does not appear that ED consistently asked States to address these requirements across all dimensions of their proposed accountability systems. For example, in **Ohio’s** approval letter, ED requested information about the validity and reliability of the State’s uniform averaging procedures once they have been in operation. However, beyond this instance, ED did not request additional information about how the State plans to examine the technical merit of other specific aspects of its accountability system, including the quality of the AYP decisions and the array of consequences that follow them.

In general, States tended to be vague in their responses to Critical Elements 9.1 and 9.2 (Reviewer Questions E1 and E2) in which they must describe plans for how they will examine the reliability and validity of their accountability systems. Very few States, (e.g., **Minnesota**), submitted a comprehensive plan for these examinations, which is understandable given the lack of precedence in the accountability context. Yet, to date, ED does not appear to have specifically asked any State to address this deficit.

In the long run, it likely will be up to States themselves to develop and implement sound plans for validating their accountability systems; the evidence gathered in the

¹⁵ There were a total of 19 questions Peers were asked to respond to regarding each State’s accountability workbook.

validation process will be critical for determining whether these systems are working as intended. Further, States will need this evidence to help defend their AYP decisions and the imposition of correspondingly serious consequences. It is also worth noting here that some States have expressed frustration with the expectation that they must validate a system that has in many ways been imposed upon them, a “system” that they consider seriously flawed at the outset. However, from this perspective, as what is deemed acceptable continues to evolve and the law’s next reauthorization takes shape, it will be important for States to have clear validity evidence of what works and what does not in order for more informed models to find their way into use.

To assist States in gathering and examining validity and reliability evidence, CCSSO is currently developing a paper through two of its State Collaborative on Assessment and Student Standards (SCASS) groups, the Accountability Systems and Reporting SCASS (ASR SCASS) and the Comprehensive Assessment Systems SCASS (CAS SCASS). This paper will provide a framework for how to conceptualize the validity issues related to accountability systems, gather and examine relevant evidence, and evaluate how the systems are (or are not) working. States will need to engage in rigorous validation processes, such as those that will be addressed in the ASR/CAS SCASS paper, in order to defend their accountability systems to their stakeholders and in courts of law as well as to offer evidence of the impact of NCLB within their States.

AYP Consequences and Reporting

ED has consistently maintained that States must be able to make AYP determinations prior to the beginning of each school year in order for eligible parents and students to be notified of their opportunity for public school choice and supplemental educational services. For States with late spring testing windows, meeting these requirements can be extremely challenging. In some instances, States have proposed making school/district identification for improvement determinations on the basis of “preliminary data.” To the extent that these preliminary determinations can be made reliably, ED has approved them while sometimes encouraging the State(s) to make changes in their testing windows as well as scoring and reporting procedures/timelines.

Forcing States to change testing windows and tighten up timelines related to scoring and reporting can also have a significant impact on the type and quality of assessments States use. For example, States opting for testing windows near the end of the school year may decide against the use of, or sharply reduce the use of, open-ended or constructed response items because of the additional scoring time involved. Further, it is not known whether ED considered, in developing its position, the optimal time for State assessments to be administered, permitting parents the opportunity to visit prospective choice schools while they are in session, or beginning enrollment in choice schools at mid-year or the beginning of the following year.

This issue is driven largely by when a State’s testing window occurs. However, it also can become an issue for States with year-around schooling or those in which districts are free to determine their own starting and ending dates without regard to related parameters. Such practices definitely affect consistency in the application of decision rules across schools and districts. The lack of uniformity may serve to place a State in jeopardy of challenge by schools and districts identified for improvement.

As its solution to the tight timeline, **Massachusetts** will make AYP decisions based on preliminary data. Beginning with 2002-03 data, this preliminary AYP determination would be delivered to schools and districts before the end of August and schools and districts that were preliminarily identified would be required to provide public school choice and supplemental educational services according to the requirements of sec. 1116. Schools and districts that were preliminarily identified and in the final determination made AYP would continue to provide choice throughout the remainder of the school year.

Another issue related to State report cards is the law's provision calling for the first report to be issued not later than the beginning of the 2002-03 school year [sec. 1111(h)]. The law and regulations are silent as to whether subsequent report cards must be issued prior to the beginning of each school year. This did not appear to have been an issue for ED during the Peer Reviews.

The next section offers a set of conclusions about the accountability system designs proposed by the States and the "approval" decisions released to date by ED.

Part III: Conclusions

This section identifies areas that, so far, (1) have been pretty much “non-negotiable” in terms of variation from the NCLB law, regulation, or ED policy, (2) have produced some unanticipated approvals, and (3) have resulted in ED approvals that could have either short-term or long-term impacts on AYP determinations.

Non-Negotiable Issues

The following are areas where some States have sought to “push the envelope” with respect to NCLB requirements and which ED has almost consistently ruled against:

- **Annual accountability determinations for ALL schools including new schools, schools created by consolidations, and alternative schools**—In promulgating final accountability regulations in December 2002, ED noted in comments regarding §200.20, “In those instances in which schools and districts are too small to include any subgroups, the school and district will need to make a decision about AYP at least on the basis of all students who were enrolled for a full academic year. The Department of Education will issue nonregulatory guidance to provide examples of methodologies for handling this issue.” (*Federal Register*, December 2, 2002, p. 71744). ED’s Peer Reviewer Guidance and Report (March 6, 2003) include related questions under Critical Elements C4 and D3.

In its final round of “approvals,” ED modified its position on this issue with respect to new schools (not those created as a result of reorganization or re-districting). For new schools, an AYP determination does not have to be made until the end of the second school year following the school’s opening. However, the school district must provide a public report of the school’s progress based on its first year of operation.

- **AYP determinations that consider which subgroup(s) fail to meet the AMO are not acceptable**—ED’s policy is that there are two ways to be identified for improvement based on not meeting the AMOs: (1) not meeting the AMO in either reading or language arts and mathematics for two consecutive years or (2) not meeting the AMO in the same subject for two consecutive years. Since both the NCLB Act and the related regulations are silent on this issue, ED appears to have chosen to interpret the law’s intent in this manner. Obviously, the former way is more stringent than the latter and, to date, only Louisiana has opted for this approach.
- **Starting Points and Trajectories must be based on all students statewide** and cannot be established by subgroups, schools, or districts. ED has been fairly consistent in this area with the exception of a few late “approvals” permitting the use of school-based trajectories in “safe harbor” reviews.
- **Index systems must treat reading or language arts and mathematics separately and cannot allocate additional points for the advanced**

proficient level. ED has consistently indicated to States that while the use of indexing in making AYP determinations is permissible, States cannot establish index weights that might serve to mask low student performance.

- **Apportioning student membership across subgroups in making AYP determinations** when individual students belong to two or more subgroups. Delaware proposed a plan to apportion these results across subgroups to mitigate the over-representation of the lowest performing students (e.g., a student who appears in a race and ethnicity subgroup, an LEP subgroup, and an economically disadvantaged subgroup). In this example, the student would be apportioned at .33 for each subgroup. ED rejected this proposal outright.
- **Participation Rate based on enrollment at time of testing**—ED has been unwavering in its requirement that Participation Rate be calculated on the basis of all students enrolled at the time of testing and not on the basis of students enrolled for a full academic year.
- **Delaying the first Intermediate Goal beyond 2004-05**—ED has not approved any plan in which the first Intermediate Goal occurs after the 2004-05 school year.
- **Delaying or extending the 2013-14 goal**—ED has not approved any plan in which the goal of all students proficient in reading or language arts and mathematics occurs after the 2013-14 school year.
- **Using Confidence Intervals in indicators that are a “count”**—Throughout the Peer Review process, ED did not approve the use of confidence intervals for determining Participation Rate and Other Academic Indicators such as attendance. However, ED did approve the use of CIs in recent negotiations with at least two States for “count” indicators.
- **“Parent opt-out Laws”**—Some States have statutes permitting parents to have their children “opt-out” of State assessments. ED’s position is that this is a State issue but it does not permit States to exclude those students from Participation Rate determinations or other accountability requirements under NCLB.
- **Exclusion of GEDs from graduation rate**—ED has not approved any plan that includes students who earn GEDs in the definition of a graduate.

Unanticipated Approvals

As the reviews of State accountability workbook plans progressed from January through April of 2003, what could be approved by ED seemed to change. As a result, some issues that were called “non-negotiable” in early reviews were approved during the latter stages of the Peer Reviews and the ensuing negotiations between ED and various States. As noted earlier in this paper, recent approvals by ED have even included proposals from some States that appear to conflict with NCLB provisions. These include

- **Dual Accountability Systems.** Early in the review process, it seemed that States would be allowed to implement secondary accountability systems, including extra rewards and sanctions, only if the NCLB AYP outcomes served as a sort of ceiling for the second system. That is, a school could not achieve a high performance level on the second system if it was identified for improvement

under AYP. Now, it appears that several States have won approval for systems that recognize schools regardless of their AYP outcomes.

- **Use of Larger Minimum “n’s” for Subgroups**—ED approved the use of larger minimum “n’s” for several States. Some States successfully argued for the use of higher minimum “n’s” with only the SWDs subgroup.
- **Including Scores for “Exited” SWDs** in that subgroup’s AYP determination. Although several States have now received approval for including exited LEP students in the LEP subgroup for a couple of years, this option seemed off the table for SWDs. However, Georgia plans to include exited SWDs in the SWD AYP group as long as they are being monitored in some way.
- **Including Scores for “Exited” LEP Students** in that subgroup’s AYP determination.
The fine distinction ED is making here is that the students in question must have not technically “exited” the program as long as they are still receiving services. Establishing “exit” criteria and monitoring to ensure that a student can read, write, and understand the English language permits the student to stay LEP for classifying and reporting purposes (and AYP determinations). This will also increase the subgroup size and may have other implications that States should carefully analyze such as the need for continued participation in annual assessments under Title III.
- **Out-of-Level Testing**—As noted earlier in this paper, States involved in early Peer Reviews were advised by ED that, consistent with NCLB regulations, out-of-level testing was not permissible. In more recent reviews, that position has changed. Several States have received approval to use out-of-level testing provided that student results are reported as less than proficient on State academic content and student achievement standards. However, how results can be reported and applied for AYP determinations was reversed in a June 27, 2003 letter from Secretary Paige to Chief State School Officers.
- **Using Number Tested Rather than Number Enrolled** to calculate Percent Proficient. Several States have now been allowed to use the number of students tested rather than the number of students enrolled when calculating Percent Proficient. Many States did not directly indicate what they would use in the denominator.
- **Allowing States to Use Non-Uniform Averaging across Jurisdictions**—Several States won approval for plans that allow decisions about the number of years of data used for AYP to be made at the unit level. Thus, even within a district, one school may use one year of data and another might use three years of data.
- **Exemptions from Testing for Some Students**—Delaware successfully argued for approval of its policy to exempt students from State assessments, in extreme cases and rare situations, where an unexpected medical or psychological condition prohibits inclusion provided that the school district requests and receives approval from the State.
- **Use of Statistical Tests in “Safe Harbor” Reviews**—Until late in the series of approvals, ED rejected proposals to use a priori criteria for statistical significance for safe harbor. However, Louisiana, Maryland, and Nevada will all use a pre-specified confidence interval when invoking safe harbor.

Approvals Not Likely to Have Long-Term Impacts on AYP Determinations

Briefly highlighted below are areas where ED has approved portions of State accountability plans that are likely to have some initial—but not long-term—impact on the number of schools and districts that may be identified for improvement.

- Extending time in which LEP students and SWDs are included in subgroup AYP determinations.
- Uniform averaging—permitting schools and districts to select from several options instead of requiring a single model statewide.
- Using out-of-level testing (referred to as “Instructional Level Assessments” in Secretary Paige’s June 27, 2003 letter) results in AYP determinations if this is permitted only for 2002-03 results.

Approvals that May Have Long-Term Impacts on AYP Determinations

ED has approved several aspects of State accountability models that may indeed have long-term impact on the number of schools and districts identified for improvement. These include

- Modifying student academic achievement standards;
- Application of certain statistical tests to data analyses including SEMs and other confidence intervals;
- Use of most recent scores when students are permitted to take assessments multiple times;
- Use of higher minimum “n’s” for some subgroups;
- Not rolling up data over multiple years to make subgroup AYP determinations; and
- Requiring schools and districts to make “progress” on the other academic indicators rather than meet or exceed a specific target.

Appendix A – References/Resources

U. S. Department of Education

- Public Law 107-110, The No Child Left Behind Act of 2001.
- Final Regulations, Title I—Improving the Academic Achievement of the Disadvantaged, Standards and Assessments, July 5, 2002, *Federal Register*, 34 CFR Part 200, pp. 45038-45047.
- Final Regulations, Title I—Improving the Academic Achievement of the Disadvantaged, Accountability including Adequate Yearly Progress, December 2, 2002, *Federal Register*, 34 CFR Part 200, pp. 71711-71771.
- Peer Review Guidance for Title IA Accountability Provisions of the No Child Left Behind Act of 2001. Washington, DC: U. S. Department of Education (March 6, 2003).
- Peer Review Report for Title IA Accountability Provisions of the No Child Left Behind Act of 2001. Washington, DC: U. S. Department of Education (March 6, 2003).
- Final Policy Decisions (2003). Washington, DC: U. S. Department of Education.
 - ◆ Participation in a State’s Academic Assessment
 - ◆ Use of the Other Indicator for “Safe Harbor” Purposes
 - ◆ Apportionment of Students Across Subgroups
 - ◆ Use of the Uniform Averaging Procedure
 - ◆ AYP Determinations Based on Subject Triggers
 - ◆ Proficiency Trajectories
 - ◆ Integration of State Index Accountability Systems with NCLB

Others

- ASR-CAS Joint Study Group on Adequate Yearly Progress. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.
- Gong, Brian & Marion, Scott (2003). *Some notes about confidence intervals, reliability, and NCLB*. Dover, NH: The Center for Assessment. Paper available at www.nciea.org
- Marion, S., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

Appendix B

The following are the ten principles for NCLB accountability systems specified in Secretary Paige's letter to Chief State School Officers, dated July 24, 2002:

1. **A single statewide accountability system applied to *all* public schools and LEAs.**¹⁶
 - “All schools and LEAs” includes Title I and non-Title I schools and LEAs.
 - Student assessments are administered and the accountability system is applied in the same manner for all schools, regardless of receipt of Title I funds.¹⁷
2. ***All* public school students are included in the State accountability system.**¹⁸
 - A student attending the same school for a “full academic year” must be included when determining if a school has made AYP.
 - A student that attends more than one school in a district during the school year is only included in determining if a district has made AYP.
 - All student results are included in the school level report card.
3. **A State's definition of AYP is based on expectations for growth in student achievement that is continuous and substantial, such that all students are proficient in reading and math no later than 2013-2014.**¹⁹
 - Accountability systems must establish proficiency goals statewide, based on assessment data from the 2001-02 school year that progressively increase to reflect 100 percent proficiency for all students by 2013-14.
 - These goals must increase at steady and consistent increments during the 12-year timeline, although not necessarily annually throughout the 12 years (i.e., States cannot establish goals that will require the most substantial progress toward the end of the 12-year timeline).
 - Increases in proficiency rates must occur for a school to make AYP. Progress in student achievement from the “below basic” to the “basic level” is not in and of itself sufficient to meet AYP requirements. However, States and LEAs are strongly encouraged to develop systems to recognize very low-performing schools that are making such improvement.
4. **A State makes annual decisions about the achievement of all public schools and LEAs.**²⁰
 - States may calculate AYP for a school using up to three consecutive years of data.
 - If a State chooses to average data over two or three years, it must still determine whether a school or district made AYP on an annual basis.
5. **All public schools and LEAs are held accountable for the achievement of individual subgroups.**²¹
 - Accountability decisions must be based on the achievement of each subgroup in the law, as well as overall achievement.

¹⁶ Sections 1111(b)(2)(A) and 1111(b)(2)(C)(i).

¹⁷ Requirements for school improvement, corrective action, and restructuring under Section 1116 only apply to schools receiving Title I funds.

¹⁸ Sections 1111(b)(2)(A), 1111(b)(3)(C)(xi), 1111(b)(3)(C)(xi), and 1111(b)(3)(C)(xiii).

¹⁹ Sections 1111(b)(2)(C)(iii), 1111(b)(2)(F), and 1111(b)(2)(H).

²⁰ Section 1111(b)(2)(J).

²¹ Sections 1111(b)(2)(C)(v), 1111(b)(2)(C)(v), and 1111(b)(2)(C)(v)(II).

- States must set separate, measurable annual objectives for each of these subgroups that ensure they meet the deadline to reach proficiency within 12 years.
 - Subgroups for accountability are major ethnic/racial groups, economically disadvantaged students, limited English proficient (LEP) students, and students with disabilities. The goals for each subgroup may be the same as long as each subgroup reaches 100 percent proficiency in 12 years.
- 6. A State's definition of AYP is based primarily on the State's academic assessments.**²²
- Decisions about school and LEA progress must be primarily determined by achievement on academic assessments.
- 7. A State's definition of AYP includes graduation rates for high schools and an additional indicator selected by the State for middle and elementary schools (such as attendance rates)**²³
- Other academic indicators may be included in addition to these required indicators.
 - These indicators may only have the effect of indicating a school did not make AYP. In other words, a State may use these indicators to identify a school for improvement, but they may not be used to prevent a school from being identified for improvement.
- 8. AYP is based on separate reading/language arts and math achievement objectives.**²⁴
- Each subgroup of students enrolled in schools and LEAs must meet annual objectives in reading and math for the school or LEA to make AYP.
- 9. A State's accountability system is statistically valid and reliable.**²⁵
- In determining AYP, a State is not required to use disaggregated data when the number of students in a subgroup is (a) too small to yield statistically reliable information or (b) the results would reveal personally identifiable information.
 - Each State determines a minimum size of a group, below which the results would not be statistically reliable for use in determining AYP. States make a reasonable determination of that number based on the technical specifications of their assessments.
- 10. In order for a school to make AYP, a State ensures that it assessed at least 95% of students in each subgroup enrolled.**²⁶
- Schools must report all student results by subgroup. The number of students in a subgroup must be of sufficient size to produce statistically reliable results for the 95% requirement to affect AYP. In other words, if the number of students in a subgroup is too small to produce statistically reliable results, the State need not, on the basis of the 95% requirement, identify the school as not making AYP, even if fewer than 95% of the students in that subgroup take the State's assessment.

²² Section 1111(b)(2)(C)(iv).

²³ Section 1111(b)(2)(c)(vi).

²⁴ Section 1111(b)(2)(G)(i)

²⁵ Section 1111(b)(2)(C)(ii)

²⁶ Section 1111(b)(2)(I)(ii).