A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989–2009

Mary Lyn Bourque

March 2009

Paper Commissioned for the 20th Anniversary of the
National Assessment Governing Board
1988–2008

Mary Lyn Bourque is President of Mid Atlantic Psychometric Services of Leesburg, VA. From 1989 to 2000 she was Assistant Director for Psychometrics of the National Assessment Governing Board.

Introduction

A history of the achievement levels for the National Assessment of Educational Progress (NAEP) could be documented in different ways. A chronological history might be the obvious approach, but in the author's view, that could miss the most salient aspects of setting student performance standards on an assessment like NAEP. Further, new initiatives are fostered within a social and political context that is critical to understanding the initiative's development, direction, and destiny. Therefore, rather than following a strict chronology, the author has adopted an issues-based approach, describing first the contextual issues that surrounded NAEP during the beginnings of achievement levels in the late 1980s and early 1990s. The paper will then discuss the critical initial implementation decisions made by the National Assessment Governing Board as it developed the student performance standards for NAEP. Then it will follow new implementation issues into the first decade of the 21st century; finally, it will discuss the broad impact of the achievement levels on American education.

Contextual Issues: Early NAEP[1]

NAEP emerged as a federally supported program in the late 1960s, often pinpointed at October 1969 when the U.S. Office of Education (USOE) awarded a 1-year, $2 million grant to the Education Commission of the States (ECS) to support the planning and initial assessments of NAEP. Prior to that time (during the initial planning phases and dating from spring 1963), support for NAEP came from private sources, including the Carnegie and Ford Foundations, with encouragement by the USOE. Two years later (August 1971), however, the USOE transferred administrative responsibility to the National Center for Education Statistics (NCES), where it remains today (though with some legislative modifications). It took another two years (September 1973) for the USOE to announce that NAEP's funding arrangement would move from a grant to a contract. The fiscal year 1975 budget request by President Nixon included $7 million for NAEP. Eventually, Congress appropriated $3 million.

Interestingly, NAEP emerged as an initiative that Francis Keppel, then U.S. Commissioner of Education, believed would contribute to fulfilling the intent of the legislation that created the USOE 100 years earlier. That legislation required that the Commissioner of Education provide an annual report to Congress on the state of American education. For the first 100 years the annual reports focused on student-teacher ratios, per capita spending on education, number of classrooms and teachers, and teacher salaries, but provided little or no information on the outcomes of education; that is, what students know and can do. Keppel thought that a national assessment of students conducted on a regular basis could fill that void.

The initial design of the NAEP program was developed carefully to protect the rights of the states. It virtually precluded using NAEP as a lever for policy changes in American education: NAEP sampled students by age, not by grade; NAEP reported results at the regional level of aggregation, not by states or districts; and NAEP participation was strictly voluntary. A testimony to the wisdom of the original group of crafters, and to the adage that "change comes slowly," NAEP is still making its presence felt as a partner to the change embodied in the No Child Left Behind (NCLB) legislation (P.L. 107–110).

The early years of NAEP planning occurred during the post-Sputnik era. The U.S.S.R. launched its first satellite orbiting the earth in 1957. This event sent shockwaves throughout the western world, particularly the United States, which viewed itself as a world leader. So while there was an urgent need to know what American students know and can do, there was an equally urgent need to critically evaluate the American system of education and its relationship to our advancement as a nation. This was the era of the National Science Foundation's support for science teachers and foreign language learning as well. In short, it was a time of taking stock in how we fared as a nation, what future generations of American citizens could and would accomplish, and how we could improve education by making it more accountable at all levels.

Internationally, we were (in the 1960s) and still are (in 2009) one of the only developed countries that does not have a centralized structure of elementary and secondary education. Contrary to other institutions, education in the United States has never been "nationalized." Two hundred years after the founding fathers, the United States has no national curriculum, no national content or performance standards, no national testing of individual students, no national accountability, no national expectations for a high school diploma (McNeil, 2008), and no national training program or entry standards for teachers or administrators. In addition to more than 50 publicly funded state education systems and nearly 15,000 local education agencies, there are numerous private schools, charter schools, alternative schools, home schools, and other forms of schooling that satisfy compulsory education laws enacted by states, not by the federal government. The founding fathers awarded full responsibility for the education of the nation's children to the states, and federal intrusion into such matters has always been viewed with suspicion.

The mix of the domestic social and historical contexts of the 1960s and 1970s resulted in a National Assessment that developed slowly.[2] From funding issues to administrative issues, NAEP struggled to develop into a first-class assessment program without a broad-based federal infrastructure to support it. Consensus came at a snail's pace from the various public groups, especially Congress. However, by the mid-1980s the political climate began to change. *A Nation at Risk* (National Commission on Excellence, 1983) was published, the Educational Testing Service (ETS) was awarded the new NAEP contract, and a *New Design for a New Era* (Messick, Beaton, and Lord, 1983) provided a whole new framework for NAEP.

During the first 20 years of NAEP, much was accomplished due to the diligence of its contractors and those at the federal level responsible for its administration. Assessments were developed and administered to samples of K–12 students in a variety of subjects, including science, citizenship, reading, writing, and computer competency; in addition, out-of-school samples and a young adult literacy study (16- to 25-year olds) were implemented. However, despite some progress, the direction of NAEP was too dispersed among contractors, federal agencies, and interested public groups to allow the program to reach its full potential as the nation's leading indicator of academic progress. Someone needed to be "in charge" if NAEP was to move forward.

Contextual Issues: Early Governing Board

By the mid-1980s, states began to realize that better reporting mechanisms were needed to measure student progress. Comparisons of SAT and ACT college admission scores state by state were no longer satisfactory. The National Governors' Association (Alexander, 1986) called for better "report cards" that could more accurately compare states' performances to each other and to the nation.

The Alexander-James Panel, appointed by William Bennett (then Secretary of Education), was a response to the governors' call. The 22-member panel report issued in 1987 (Alexander and James, 1987), along with commentary by the National Academy of Education (Glaser, 1987), ultimately formed the basis of the reauthorization of the Elementary and Secondary Education Act, also known as the Hawkins-Stafford Elementary and Secondary School Improvement Amendments of 1988. The Governing Board was created by this federal legislation (P.L. 100–297, 1988), which was signed by President Reagan in April 1988.

The new legislation covering NAEP was the first attempt to put someone "in charge," at least in the policy sense. In theory, the law gave the Governing Board responsibility for program policy, NCES would oversee program administration, and the NAEP contractors would provide technical expertise. The law set NAEP on what some refer to as a three-legged stool for support: a tripartite, shared responsibility where the distinction between policy and administration was not clear, and policy and technical issues overlapped considerably. In theory it sounded logical. But in fact, some of these distinctions are still being worked out today, some 20 years later.

The Governing Board took its legislative charge seriously. Clearly, its most controversial responsibility was that of setting student performance standards for NAEP, which the Board decided to term achievement levels. P.L. 100–297 makes more than one reference to what could be interpreted as student performance standards. One part of the statute directs the Board to "identify appropriate achievement goals for each age and grade in each subject area to be tested (Sec. 3403, (6)(A)." In another section the law states that "Each learning area assessment shall have goal statements devised through a national consensus approach (Sec. 3403, (6)(E)." Both directives could be interpreted to mean different things and, further, both could be interpreted differently by different groups. In order to better fulfill the intent of the law, the Board leaned heavily on the National Academy of Education (NAE) commentary on the Alexander-James report, which was the basis of the legislation (Alexander and James, 1987; Glaser, 1987). In that document the Academy panel argued that "for each content area NAEP should articulate clear descriptions of performance levels, descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent, and expert. Descriptions of this kind would be extremely useful to educators, parents, legislators, and an informed public (Glaser, 1987)."

During the first 18 months after the Board's formation, it developed a policy statement on NAEP student performance standards. The Board agreed to adopt three achievement levels (Basic, Proficient, and Advanced) for each grade and subject area assessed by NAEP. In arriving at that decision, the Board researched different options, including whether the legislative phrase "identifying appropriate achievement *goals* [emphasis added]" was synonymous with setting "targets" on NAEP, or whether the law's intent was to identify the *content* that groups of

students should know and be able to do if they reached the standard. The Board ultimately rejected the notion of "targets," believing that setting passing rates for the decentralized education system of the United States was beyond the policy competence of a national Board.

The Legislative History of Achievement Levels

Before launching into an issues-based history of the NAEP achievement levels, it might be beneficial to briefly summarize the legislative history of the achievement levels (figure 1). Before 1988, the NAEP legislation was mute on such things as performance standards. This early legislation was content to cover such things as which subject areas would be assessed, which samples would be assessed, where the resources would come from, etc.

The first NAEP legislation that spoke to performance standards was P.L. 100–297 of 1988, which established the Governing Board to set policy and required it to identify "appropriate achievement goals for each…grade… (and) subject area" to be assessed. Although the law was ambiguous about exact expectations, the Board understood that it was charged with developing student performance standards. The legislation did not spell out how, but left that to the wisdom of the broadly representative Board.[3]

The 1994 NAEP legislation, Improving America's Schools Act of 1994 (P.L. 103–382), came out of the same congressional session as the Goals 2000 law, Educate America Act (P.L. 103–227). The latter codified the education goals promulgated by the Charlottesville Summit and reflected the early policy definitions adopted by the Board, which, in turn, borrowed from the language used at Charlottesville. The National Education Goal 3 stated in part that, *"All students will leave grades 4, 8, and 12 having demonstrated competency over challenging subject matter… so they may be prepared for responsible citizenship, further learning, and productive employment in our nation's modern economy."*

In the 1994 NAEP legislation, the achievement levels were judged "developmental" for the first time. This came as a result of two outside evaluations, which concluded that the student performance standards on NAEP did not meet technical expectations and should not be used as the primary means for reporting NAEP. (National Academy of Education, 1993; U.S. General Accounting Office, 1993).

The current legislation (P.L. 107–110), adopted in 2001, places the achievement levels within the accountability framework of NCLB, though they still are to be used by NAEP on a "trial basis." States receiving federal education aid must participate in the National Assessment. They must also set standards on their own state assessments using the terms of the NAEP achievement levels (Basic, Proficient, and Advanced), though there is no requirement to have the same academic content or rigor.

Figure 1

Legislative History of Achievement Levels: 1988–2001

P.L. 100–297 (1988)	P.L. 103–382 (1994)	P.L. 107–110 (2001)
(6)(A)(ii) "identifying appropriate achievement goals for each age and grade in each subject tested under the National Assessment; (6)(E) Each learning area assessment shall have goal statements devised through a national consensus approach, providing for active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the general public."	Sec. 411 (e) STUDENT PERFORMANCE LEVELS "PERFORMANCE LEVELS. The National assessment Governing Board, established under section 412, shall develop appropriate student performance levels for each age and grade in each subject area to be tested under the National Assessment. DEVELOPMENT OF LEVELS. Such levels shall be – Devised through a national consensus approach, providing for active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the general public; Used on a developmental basis until the Commissioner determines, as a result of an evaluation…, that such levels are reasonable, valid, and informative to the public; Updated as appropriate. In using such levels on a developmental basis, the Commissioner and the Board shall ensure that reports that use such levels do so in a manner that makes clear the developmental status of such levels. REPORTING – After determining that such levels are reasonable, valid, and informative to the public, as a result of an evaluation …, the Commissioner shall use such levels or other methods or indicators for reporting results of the National Assessment and State Assessments."	Sec. 602 (e) STUDENT ACHIEVEMENT LEVELS "ACHIEVEMENT LEVELS. The National Assessment Governing Board shall develop appropriate student achievement levels for each age and grade in each subject area to be tested under assessments authorized under this section, except the trend assessment… DETERMINATION OF LEVELS – In general such levels shall be determined by identifying the knowledge that can be measured and verified objectively using widely accepted professional assessment standards; and developing achievement levels that are consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge for grade levels to be assessed, or the age of the students, as the case may be. NATIONAL CONSENSUS APPROACH – After the determinations described in subparagraph (A), devising a national consensus approach. TRIAL BASIS – The achievement levels shall be used on a trial basis until the Commissioner determines, as a result of an evaluation…, that such levels are reasonable, valid, and informative to the public. STATUS – The Commissioner and the Board shall ensure that reports using such levels on a trial basis do so in a manner that makes clear the status of such levels. UPDATES – Such levels shall be updated as appropriate by the National Assessment Governing Board in consultation with the Commissioner. REPORTING – After determining that such levels are reasonable, valid, and informative to the public, as a result of an evaluation …, the Commissioner shall use such levels or other methods or indicators for reporting results of the National assessment and State assessments. REVIEW – The National Assessment Governing Board shall provide for a review of any trial student achievement levels under development by representatives of State educational agencies or chief State school officers…"

The differences between these pieces of authorizing legislation are quite nuanced in one sense. At first blush, they all look pretty much alike, although the most current one is far more explicit than the initial one. On the other hand, there are distinctions that affect how the achievement levels are viewed, used, and maintain their currency. As the paper discusses the various issues below, these differences will be highlighted.

Implementation Decisions: The Beginning of Achievement Levels

The Board faced many implementation questions during the early years of developing achievement levels, including the number of levels and what they should be called, a description of the achievement levels, the methodology to be used to develop them, the composition of standard-setting panels, how to report student performance standards, and the transition from anchor levels to achievement levels. The following sections will examine each of these topics.

How Many Levels?

This question was one of the first to come up in developing Board policy. The 1970s and 1980s experienced the growth of the minimal competency movement. The historical result of that movement was one of mediocrity. Indeed, standards were set, usually a single passing score and usually on tests of basic skills, and at a level to ensure most students would "pass." However, the minimum competency paradigm would not work for NAEP in the context of *A Nation at Risk,* the Charlottesville summit, and the push for international competitiveness in a global economy. On the other hand, setting standards too high on a challenging test such as NAEP could result in irrelevance for NAEP and, more importantly, in a lack of involvement in the completely voluntary National Assessment. Furthermore, it would be far more beneficial to the states (at whose behest improved reporting of academic student performance was initiated) to be able to describe the performance of students across the whole distribution, not just at a single "passing score." States wanted to know how *all* their students performed—those at the top of the distribution as well as those in the middle and lower ranges. Indeed, they wanted to know if some of their students met or even exceeded expectations, and if some of their students did not meet any standards or were in the "almost there" category. This called for more than just a "pass."

Although the Board considered reporting a single level of performance, it was convinced to consider more than one level, especially by the late Albert Shanker, then President of the American Federation of Teachers.[4] In fact, NAEP was ideally suited to employing multiple levels since the item pool used at any one grade level was substantial (e.g., the 1990 math assessment used nearly 200 items to measure grade 8 performance), and the use of large item pools was a necessary condition for having multiple standards. In the final analysis, the Board adopted three levels—Basic, Proficient, and Advanced—that were defined by Governing Board policy and applicable to all subject areas and grades.

Why three? The reasons are both practical and technical. The technical reason is that the scale range would probably not support more than nine levels of performance (three grades x three levels) on a single cross-grade scale and still provide clear distinctions between them, as Board policy called for. Although this was an empirical question at the time, NAEP data have

borne this out. The practical reason is somewhat related. Immediately before the adoption of the Board's policy, NAEP had been using the ETS-developed NAEP "anchor levels"—one in the middle of the distribution, two higher, and one lower.[5] Since one of the purposes of adopting the achievement levels was to "improve the form and use of NAEP," three levels at each grade seemed to be the "Goldilocks approach" that would be "just right."

Other labels were considered during the policy development process, including the suggestions found in the NAE commentary (Glaser, 1987) on the Alexander-James report (1987). These included such labels as "novice, journeyman, highly competent, and expert." There was also much discussion about using the label "proficient" precisely because the anchor points were collectively referred to as NAEP proficiency. Similarly, the Board entertained a numerical labeling (e.g., Levels 1, 2, 3, etc.) and the terms "fundamental" and "master" for the Basic and Proficient levels, respectively. However, none of these seemed to meet the legislative charge of "improving the form and use of NAEP," nor did they comport with the descriptions of the standards the Board had in mind. The Board wanted a more descriptive label that corresponded to the content of the levels of expectations envisioned by the policy. In the final analysis, alternatives were eventually set aside in favor of the current identifiers.[6]

Policy Definitions

Who should set the standards on NAEP does not seem to be a salient question today, but in 1989 it truly was an issue. The Board struggled in the early days to find a way to develop a process for standard setting that invoked the Board's role as the policy body for NAEP. The Board believed that, since it was the legal entity responsible for setting standards under the federal statute, it should set the *expectations* for "how good is good enough." By the 1992 mathematics standard-setting initiative the Board had arrived at a three-prong solution: (1) the Board would develop policy definitions (PDs) that articulate the *expectations* for each level in general terms, (2) the PDs would be operationalized into grade- and subject-specific statements of performance levels for use in the standard-setting process (achievement level descriptions, or ALDs), and (3) the Board would be the final arbiter of the recommendations provided on all aspects of the levels, including the ALDs, the cut scores, and the exemplar items.

The first general definitions of performance (i.e., PDs) were developed to ensure that the standards from subject area to subject area, and from grades 4 to 8 to 12, were aligned. In other words, the Board did not want different standards (easier or harder) in reading than in math, or in science than in writing, nor did they want different standards in grades 4, 8, or 12.[7] The first definitions were called policy definitions to distinguish them from the achievement level descriptions developed later. The original policy definitions appear in figure 2.

Figure 2

Original Policy Definitions

(Excerpt from the Governing Board Policy (May 10, 1990))

Proficient. This central level represents solid academic performance for each grade tested—4, 8, and 12. It will reflect a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. For grade 12, the proficient level will encompass a body of subject-matter knowledge and analytical skills, cultural literacy, and insight that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.

Advanced. This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For grade 12, the advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement tests.

Basic. This level, below Proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade tested. For grade 12, this will be higher than minimum competency skills (which normally are taught in elementary and junior high schools) and will cover significant elements of standard high school level work.

The opening page of the Board's policy statement, unanimously approved on May 11, 1990, stated:

> *"The National Assessment Governing Board is not authorized to establish any overarching national goals for education. It does have authority to define levels of achievement that will serve as 'appropriate achievement goals' on National Assessment exams . . . Hence the achievement levels defined by the Board will be used for reporting group data and making it more meaningful* (National Assessment Governing Board, 1990, p. 1)."

That initial policy statement goes on to say:

> *"The proposed achievement levels will add to assessment frameworks and objectives the specific definitions of basic, proficient, and advanced achievement at each grade tested, which are based on the content of the National Assessment exams. These are not broad general goals of education or curriculum, but substantive descriptions of levels of achievement tied firmly to National Assessment questions and objectives* (National Assessment Governing Board, 1990, p. 6)."

These content-specific statements became known as the achievement level descriptions. The title of the early Board policy, "Setting Appropriate Achievement Levels for the National Assessment of Educational Progress" indicates that the Board initiative was never intended to reflect national goals for school subjects. Level setting was conceptualized as follows:

- limited to better reporting the results of the NAEP survey
- limited to three grade levels (4, 8, and 12)
- limited to specific subjects areas selected by the Board
- limited in scope, reflecting a limited assessment framework

The levels would answer the question, "How good is good enough, *on NAEP?*"

Despite policy arguments regarding the limited nature of the NAEP standards, the original policy definitions were ambitious. This was understandable given the educational climate existing at the time. Congress passed the Hawkins-Stafford Act because better information was needed on the performance of American students. States wanted better reporting as well, and it was thought that making the NAEP scale understandable to the public would be a way of reaching that goal. Grade 12 was singled out in all three original definitions because grade 12 is generally viewed as the gatekeeper for any postsecondary choices, including military, employment, or advanced training. Most importantly, the initial Governing Board policy statements were fashioned after a set of common goals for the decentralized system of American education crafted at the Charlottesville Summit, a meeting of the nation's governors during the first Bush Administration (National Governors Association, 1991).

These initial policy definitions were sorely criticized by various groups; they were judged as extending far beyond the capabilities of what a cross-sectional survey such as NAEP can substantiate. Several evaluations examined the initial policy definitions and concluded that the predictive statements (e.g., at the Advanced level "…show readiness for rigorous college courses…") could not be validated using NAEP data (Linn et al., 1991; Stufflebeam, Jaeger, and Scriven, 1991). Consequently, the PDs were revised in 1993. The newer versions were streamlined and had no predictive statements, were fully balanced in that they applied to all grades and subjects equally well, and tapped into the cognitive processes related to the levels.

The full achievement level policy statement has been modified several times over the years (including as late as August 2007) to align it with recent legislative changes.[8] However, the revised PDs in figure 3 have retained their saliency over the past 15 years. They serve to ensure that the standards being set on all NAEP subjects reflect the Board's expectations for students in grades 4, 8, and 12, thereby fulfilling the Board's legislative charge as the policy body responsible for setting the NAEP achievement levels.

It is important to state that, although the Board sets the expectations and adopts the achievement levels as policy at the completion of the process, there is much that happens in between that the Board monitors, directs, and approves with the assistance of many content and technical experts, policymakers, and stakeholders. More will be said about each of these activities in later sections of this paper.

Performance Level Descriptions

Perhaps one of the most important contributions that the National Assessment has made to the standard-setting movement is the pivotal role of performance level descriptions in the standard-setting process.[9] The use of ALDs in standard setting in general was not common in 1990. In fact, there were no ALDs in the 1990 NAEP initiative (Hambleton and Bourque, 1991). Panelists were required to translate PDs directly into cut scores on the NAEP scale without benefit of the intermediary steps of using grade- and subject matter-appropriate descriptions of content. However, starting in 1992, the use of ALDs became standard operating procedure.

The timing of ALD development has varied from cycle to cycle. In one case the ALDs were developed after the fact for reporting purposes (1990), in another case they were developed by panelists during the standard-setting process itself (1992), and in another case they were developed a second time after the Board made the decision to adopt cut scores significantly different from those recommended in the standard-setting process (1996). In subsequent cycles, preliminary ALDs (viewed as *working descriptions*) were developed during the assessment framework development process by a national consensus content panel. These working descriptions were expanded and refined by panelists during the standard-setting process, and the modified and finalized ALDs were then used to report NAEP performance on each assessment along with exemplar items from the assessment.

Figure 3

Current Policy Definitions

(Excerpt from the Minutes of the Achievement Levels Committee
November 24, 1993)

Proficient. This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Basic. This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Advanced. This level signifies superior performance beyond Proficient.

In the early cycles that used this approach (1994 and 1996), the standard-setting panels were responsible to refine the preliminary ALDs during the standard-setting process in order to align the final version of the ALDs with the exact content of the assessment, which involved deleting content from the ALDs that was not measured and adding content to the ALDs that was measured.

However, in 1998, the process for crafting final ALDs was removed from the standard-setting panels altogether (ACT, 1997). Before implementing the 1998 level-setting process, content panels were brought together to examine the public comments from a broad review of the preliminary ALDs. The review, conducted via the Internet and focus groups, allowed the content panels to finalize the ALDs prior to the standard setting. The Board-approved ALDs became inputs to the standard-setting process as givens, just as the frameworks and the assessment items are givens. This approach is still being used in 2009.

When developing the ALDs, content experts have access to a number of assessment documents, including the assessment framework, test and item specifications, and the generic PDs. It is important to note that the ALDs do not derive from the item pool directly but from these more global assessment documents. The ALDs may be checked later for alignment with the item pool. However, because the ALDs are derived from assessment frameworks, they have a durability that serves NAEP well. The frameworks are reviewed only about every 10 years, while the item pools are revised (at least partially) each time a subject is given on a 2- or 4-year cycle. The approach allows the achievement level definitions to be used for reporting NAEP over the entire life of the frameworks.

How did the use of ALDs come about? Taking a closer look at the 1990 and 1992 standard-setting efforts, the Board concluded that an incorrect chronology was being used. It seemed backwards to develop the cut scores first and then, as an afterthought, develop descriptions that aligned with the cut scores.[10] As a result, in 1994 preliminary ALDs were crafted by the consensus panels that develop the assessment frameworks. Who better knew the content and what students should know and be able to do than the panels that recommended the content of what would be assessed? The preliminary ALDs were considered preliminary because they would serve to delimit the domain and the assessment content and could act as guides for the item writers. In addition, it was likely that some content identified in the preliminary ALDs might not be included (for one reason or another) in the final selection of test items on the assessment. Therefore, having preliminary ALDs provided some flexibility later on. Further, the preliminary ALDs served a very important role in the standard-setting process; that is, they communicated to the panelists the Board's expectations for Basic, Proficient, and Advanced for each grade tested and in the particular content measured. Figure 4 displays the initial NAEP ALDs in grade 12 mathematics excerpted from the 2003 NAEP assessment framework.

Figure 4
Grade 12 Mathematics ALDs

(Excerpt from 2003 NAEP Mathematics Framework)

Basic

Twelfth-grade students performing at the *Basic* level should demonstrate procedural and conceptual knowledge in solving problems in the five NAEP content strands.

Twelfth-grade students performing at the *Basic* level should be able to use estimation to verify solutions and determine the reasonableness of results as applied to real world problems. Twelfth graders performing at the *Basic* level should recognize relationships presented in verbal, algebraic, tabular, and graphical forms, and demonstrate knowledge of geometric relationships and corresponding measurement skills.

They should be able to apply statistical reasoning in the organization and display of data and in reading tables and graphs. They should be able to generalize from patterns and examples in the areas of algebra, geometry, and statistics. At this level, they should use correct mathematical language and symbols to communicate mathematical relationships and reasoning processes and use calculators appropriately to solve problems.

Proficient

Twelfth-grade students performing at the *Proficient* level should consistently integrate mathematical concepts and procedures into the solutions of more complex problems in the five NAEP content strands.

Twelfth-grade students performing at the *Proficient* level should demonstrate an understanding of algebraic, statistical, geometric, and spatial reasoning. They should be able to perform algebraic operations involving polynomials, justify geometric relationships, and judge and defend the reasonableness of answers as applied to real-world situations. These students should be able to analyze and interpret data in tabular and graphical form; understand and use elements of the function concept in symbolic, graphical, and tabular form; and make conjectures, defend ideas, and give supporting examples.

Advanced

Twelfth-grade students performing at the *Advanced* level should consistently demonstrate the integration of procedural and conceptual knowledge and the synthesis of ideas in the five NAEP content strands.

Twelfth-grade students performing at the *Advanced* level should understand the function concept and be able to compare and apply the numeric, algebraic, and graphical properties of functions. They should apply their knowledge of algebra, geometry, and statistics to solve problems in more advanced areas of continuous and discrete mathematics.

They should be able to formulate generalizations and create models through probing examples and counter-examples. They should be able to communicate their mathematical reasoning through the clear, concise, and correct use of mathematical symbolism and logical thinking.

The use of ALDs (sometimes referred to as performance level descriptors (PLDs) in other settings) in standard setting has become de rigueur for most agencies today; it was almost unheard of before the National Assessment. Today, ALDs/PLDs are used with virtually all standard-setting methods (Hambleton and Pitoniak, 2006). Hambleton and others argue that PLDs are critical to interpreting the score results and providing evidence of procedural validity of the process (Hambleton, 2001; Mills and Jaeger, 1998). Furthermore, most states now use ALDs to develop standards on their own testing programs. The ALDs represent a range of performance across the score range. These descriptions help the panelists conceptualize and internalize the policy definitions before engaging in the tasks associated with the particular standard-setting method used. Without them, standard-setting panels are left to their own devices (and perhaps creativity) to recommend appropriate cut scores. Furthermore, because the panels are so diverse it is unlikely they would have the "common understanding" of the expectations of student performance that the policy board has in mind.

Finally, the ALDs are used to report assessment results. They provide a level of interpretation for the knowledge and skills that students within the levels know and can do. The ALDs, along with the exemplar items and the percentages of examinees at or above the cut scores, provide a reasonably broad picture of what the nation's students in grades 4, 8, and 12 know and can do.

Another nuance of the ALDs worth mentioning here is the use of borderline descriptions in the level-setting process. As mentioned above, the ALDs are inputs to the process and, as such, they outline the content that students should know for those whose performance is in the designated score range. For example, the Proficient achievement level description outlines the content expectations for students whose performance is in the Proficient range; i.e., from the Proficient cut score up to the Advanced cut score. However, in training panelists to assess item content and recommend cut scores, it is necessary for them to think about the borderline performance of examinees; that is, what do students need to know to move from the Basic level up to the Proficient level? Thus, borderline descriptions are a subset of the ALDs and include the content necessary to move from a lower level up to a higher level.

In NAEP, the borderline descriptions are developed by the panels during the standard-setting process after extensive training, during which they have developed an understanding of the assessment, framework, item pools, policy definitions, and most especially the ALDs. The borderline descriptions are usually developed in grade-level groups and are working documents than can be refined during the process.[11]

At the end of two decades of level setting, the Board seems to have mastered the process for developing the descriptions of achievement. The Board sets the policy definitions, which are given to consensus panels or independent content panels to operationalize in terms of specific grade levels and content areas. The draft statements are widely vetted with a variety of NAEP audiences, including other content specialists, stakeholders, NAEP users, and policymakers. The results are incorporated into a revised version of the ALDs that the Governing Board approves before the standard setting panels use them to develop their recommendations on cut scores and exemplar items. Finally, the whole package (i.e., recommended cut scores, descriptions, and exemplars) is brought to the policy Board for review before a final decision is made.

Methodology for Developing Standards

In the early 1990s, as the Governing Board was initiating the NAEP standard-setting process, the choice of methods was limited. Berk (1986) and others (Cizek and Bunch, 2007) have described some of these early procedures, including the Nedelsky (1954) method, the Ebel (1972) procedure, and the Angoff method (1971). While all three of these early methodologies were primarily designed for use with multiple-choice items, the Angoff method was the most researched in the literature of the 1970s and 1980s. So, at the time of NAGB's initial decision to choose a methodology for NAEP standard setting, the recommendation from experts was to use Angoff. The method is a "judgment" method, meaning that panels examined items and item content and made a "judgment" about the probability of examinees answering the item correctly.[12] A distinct advantage was that the Angoff method did not require any empirical data such as examinee performance on the items.[13] Further, it was fairly straightforward to train panelists to complete the required tasks in the Angoff method; it was easily explained to standard-setting panelists (educators and noneducators alike) and could be adapted to accommodate multiple levels and multiple item formats.

By the 1992 standard setting in mathematics, the Board had contracted with ACT in Iowa City to implement the Board's policy. ACT was responsible for developing ALDs; convening national samples of grade 4, 8, and 12 panels; implementing pilot and research studies; conducting the standard-setting meetings; providing recommendations to the Board; and producing all process and technical reports. Reckase (2000) and Loomis and Bourque (2001a) both provide comprehensive descriptions of the research conducted by ACT on the different methodologies explored during this period (1992 to 2000).

Between the 1992 and 1998 cycles, ACT developed standards on seven NAEP subjects, including mathematics, reading, science, writing, civics, U.S. history, and geography. All seven used a modified Angoff procedure to develop the achievement levels.[14] The method used by NAEP was eventually modified to the extent that it took on a new descriptor and eventually was called the ACT/NAEP method. It was during this period that the ACT technical staff, with the advice of their Technical Advisory Committee on Standard Setting (TACSS), expanded and refined the standard-setting process in a number of ways.[15,16]

First, training of panelists was standardized. That is, all grade-level panelists (4, 8, and 12) were trained in large group sessions by the coordinator of the session. All groups were exposed to the same training principles. Large group training sessions were complemented by grade-level group sessions moderated by a facilitator who was trained by the coordinator. The grade-level groups reinforced the large group training sessions and offered opportunities for panelists to ask questions and explore the tasks at hand. Standardization of training is desirable in standard setting just as it is in test administration. Results in either case should not be a function of who was involved. Minimizing the unintended effects of different facilitators is a laudable and necessary goal if replication across grades, content areas, etc. is to achieve common standards.

Second, panelists were trained and encouraged to internalize every aspect of the NAEP assessments (including the NAEP assessment framework, PDs, ALDs, and item formats being used in a particular assessment) before moving on to the standard-setting task. Customized

"practice exercises" were developed to assist in this process (e.g., panelists take a student-length version of the NAEP assessment or, in areas such as reading (1992), panelists review student-constructed response papers). It was critical for participants to understand the content aspects of the NAEP assessments and the expectations for the standards embodied in the PDs and ALDs. The better part of two days (of a five-day meeting) was spent at each standard-setting event to ensure that panelists had indeed internalized the salient content aspects of NAEP before proceeding.

Third, systematic feedback to panelists during the standard-setting process was designed to make the panelists better informed "judges" as they developed the recommended standards. The standard-setting process was viewed not as providing simply a professional *opinion* about the standards, but rather one's professional *judgment* about the appropriate standards. It is believed that the better informed judge is the better standard setter. In 1992, the notion of feedback was somewhat novel in most standard setting, and was carefully crafted and cultivated as ACT improved on each NAEP cycle. Feedback was staged and provided judiciously so as to not influence unduly the judgments being made during the rating process. Such feedback examined inter- and intra-rater location data, rater consistency feedback, empirical data such as p-values, whole booklet feedback, and other useful "reality checks" for panelists to ground themselves.[17] Panelists were always free to use (or not to use) the feedback data as they saw fit in making their judgments and recommendations.

Fourth, the process (not the panelists) was monitored from beginning to end using a series of self-report questionnaires completed by every panelist. The evaluations were used to improve future processes and to determine whether the panelists felt confident in the work they had completed. These process evaluations also served as evidence of procedural validity, asking a variety of questions about their understanding of the assessment content, the training provided, the methodology being used, the feedback provided, and their level of confidence in and satisfaction with their overall recommendations to the Board.

Standardized and comprehensive training, extensive use of feedback, and formal process evaluations were all modifications added to the original Angoff method. In addition, three cut scores were recommended and the probability of a correct response to an item was expanded from "0% *or* 100%," as the original footnote suggested (Angoff, 1971), to 0 percent *up to* 100 percent and any probability in between.

Reflecting on the Legislation

The first four implementation issues described above—number of levels, policy definitions, performance level descriptors, and methodology—are not directly addressed by the legislation. The Board was initially, and still is, free to view these issues as operational and/or program policy issues, and has great latitude to formulate its decisions. However, that being said, the decisions should be made with advice from experts in the field, and full consideration of the professional guidelines for standard setting issued by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The current set, issued as part of the *Standards for Educational and Psychological Testing* in 1999, provides guidance for (1) evidence of validity, e.g., the qualifications of experts

used in the test construction process (including formulating content, standards or scoring); (2) evidence of reliability, e.g., conditional standard errors of cut scores and repeated measures reliability; and (3) test interpretation, e.g., documentation of rationale and procedures for establishing cut scores. Each of these areas provides specific guidance for the Board and its contractors, and attempts have been made over the years to collect this kind of information during the standard-setting process and to align the standard-setting process with such guidelines.

Composition of the Standard-Setting Panels

The Governing Board's first policy statement said that, "the panels be composed of individuals with expertise in the education of students of the ages and grades under consideration …[and] with knowledge of the typical subject area achievement…" (National Assessment Governing Board, 1990, p. 18). The approach was good; this was the acceptable standard in the field at the time. The problem was how do you find these panelists? The policy statement went on to say that, "Major national organizations will be contacted to recommend from among their members individuals who might serve on the panels…" (National Assessment Governing Board, 1990, p. 18). Two additional criteria included (1) continuity with the framework panels and (2) states participating in the new 1990 state-by-state assessments must be represented, especially at the eighth grade level.

Unfortunately, the 1990 effort resulted in panels that were too one-sided. Most participants were educators: classroom teachers, school administrators, and representatives of national education organizations. There were no business participants and no experts in the noneducation fields of mathematics. Because the education establishment was primarily represented, a tone was set at the meeting that was not conducive to setting standards. In reviewing that meeting, the Board came to the conclusion that the panels needed to be more diversified. To paraphrase one Board member at the time, setting standards on NAEP is just too important a task to involve only educators; we must involve others who also have a stake in the future of our country: employers, business and industry, and other stakeholders. With that in mind, the Board's 1992 policy on panel composition requqired a distribution of 55 percent teacher-educators and 45 percent nonteachers, split between 15 percent nonteacher educators and 30 percent general public who are noneducators (American College Testing, 1992). In 1995, the Board policy was slightly modified to state that about two-thirds of the panels should be teachers and other educators, and one-third should be from the public/noneducator sector. These members are drawn from a national sampling frame and are broadly representative of the NAEP regions, types of communities, various ethnicities, and genders. In 1994, the notion of having broadly representative standard-setting panels was codified in the Goals 2000 Educate America law. The 1995 policy on panel composition is still being used in 2009 and is consistent with the current legislation.

Reporting Student Achievement Levels

One of the initial reasons for moving in the direction of a standards-based approach to NAEP was to provide better and more understandable information to users, including Congress, governors, state test users, policymakers at all levels, and the American public. Therefore, how

the NAEP data were reported became a paramount concern to the Governing Board. Two issues had to be dealt with: (1) how to transition from *anchor* levels to *achievement* levels and (2) how to report what students *do* know when the standards describe what students *should* know.

NAEP assessment results in the late 1980s were reported using anchor levels. Scale anchoring was developed by ETS in the early 1980s in an effort to improve the understandability of NAEP data. Previously, NAEP results were reported using fairly simplistic approaches such as item cluster reporting and sometimes domain reporting. But in general, these approaches were unsatisfactory. ETS' *New Design for a New Era* (Messick, Beaton, and Lord, 1983) spelled out a new way of providing insight into the NAEP results called anchor levels. As conceptualized by ETS, the anchoring process involved:

- selecting a range of points on the score scale;
- constructing "item maps based on examinee performance;"
- identifying items that "anchor" around the selected points;
- describing the collection of anchored items at each of the selected points.

The five points selected for NAEP were the mean and one and two standard deviations above and below the mean. The cross-grade scale for each subject ranged from 0 to 500 (anchored at grade 8) with a mean of 250 and a standard deviation of 50, yielding the five points: 150, 200, 250, 300, and 350. Items were identified for each anchor point, and the descriptions of the points were crafted by content experts. For each subject there was one set of anchor levels that spanned the three grades at which NAEP tests. Anchor items were identified based on several criteria, including the percentage of students answering the items correctly at the selected points (80 percent), the percentage answering correctly at the next lower point (30 percent fewer), the percentage answering *incorrectly* at the next lower point (at least 50 percent), and the sample size on which the percentages are based (at least 100). The first criterion is one that can have a major impact on which items are identified as anchor items. Research has shown that using higher percentages (e.g., 80 percent) can yield anchor items and descriptions that are considerably easier in content, while using lower percentages (e.g., 65 or 50 percent) results in more difficult descriptions (Kolstadt et al., 1998).[18] Easier items on the item maps result in easier descriptions; harder items on the item maps result in harder descriptions.

Clearly this new statistical approach was an improvement over the earlier methods, but it still remained difficult for policymakers to make decisions based on such data. Even if one knew that the mean mathematics performance of grade 8 students was 242, no one knew whether 242 was "good enough." What was known was that the mean performance was slightly below the overall mean on the assessment, and the scale value could be described using the content of the anchor items (what students know and can do). However, even though the anchor levels described what students *did* know and *can* do, do they reflect what students *should know and be able to do?* In other words, was the performance of 242 good enough?

The Board was authorized to set performance levels and in 1990 it did so on the new mathematics framework. The Board preserved the older frameworks and trend lines (called long-term trends) and launched the new assessments with a new way of reporting their results—

achievement levels. The Board took on the legislative charge of answering the question, "How good is good enough?"

Transitioning from anchor levels to achievement levels has not been an easy journey. Initially, both anchor levels and achievement levels appeared together in the NAEP reports (NCES, 1994; American College Testing, 1995). The nuanced distinctions between the two were difficult for NAEP users and policymakers to grasp. One was statistically based; the other standards based—easier said than understood. One described what students know and can do (the current condition of education); the other was based on what students *should know* and *be able to do* (the goals of education). The anchor levels were statements of the status quo; the achievements levels were *expectations,* desired outcomes for performance goals on NAEP. This issue of *can* and *should* became a hotly debated topic in the mid-1990s. The author's sense is that it has never been resolved completely. However, the most cogent argument toward resolution was framed by Reckase, who argued that "can" and "should" represent a tautology. When speaking of the achievement levels as goals to be reached, the terminology "should" is used; when reporting on the percentage of examinees whose performance on NAEP is within the achievement level categories, the terminology "can" is used.

This is where the legislative terminology came into play. The 1994 legislation used the term "developmental" to describe the levels during the transition from anchor levels to achievement levels.[19] The 2001 legislation changed that language (perhaps moving it a bit forward by using the terminology "trial"). However, the decision regarding when the levels are no longer a "trial" still resides with the Commissioner, not the Board. The NAEP state trial assessments were dubbed "trial" for three cycles (1990, 1992, and 1994). By 1994, they were simply the NAEP state assessments. The same caution will probably hold for the Trial Urban District Assessments and that label could be removed in the foreseeable future. However, an evaluation of the levels by the NAE (1997) concluded that, "…the current achievement levels be abandoned by the end of the century…" A study by the National Academy of Sciences (NAS) (Pellegrino, Jones, and Mitchell, 1998) concurred. The NAS study, however, presented no new research on the issue, but depended on older studies to reach its conclusion. It seems unlikely that the label "trial" as it applies to achievement levels will be removed any time soon.

Contextual Issues at the Start of the 21st Century

As the Board moved into the first decade of the 21st century, the NAEP frameworks in most subjects were approaching 10 years in use. This was especially true for the oldest in the set, mathematics and reading. It had always been the Board's intention and policy to renew and revitalize the frameworks about every 10 years, and as movement took place in the curriculum field. That is not to say that new fads became the impetus for change. However, in the decade of the 1990s there were new developments in learning theory, and some of those developments impacted how and what was being taught in the schools. National professional organizations were rethinking their earlier curricula scope and sequence (National Council for Teachers of Mathematics, 2000); states were completing new, or revising older, state content standards in an effort to meet NCLB requirements; and new demands were being articulated for postsecondary education and training. Therefore, taking a serious look at the NAEP frameworks by curriculum experts and others who comprise the NAEP national consensus approach seemed to be in order.

This brought to the foreground new issues to be solved for this new era of the achievement levels.

Implementation Decisions: The New Century of Achievement Levels

Early in the current decade, several implementation issues needed to be resolved if the achievement levels were going to maintain their currency, and especially if they were to be helpful to states as they meet the requirements of NCLB and look ahead to its anticipated reauthorization. Should there be new frameworks in the major subjects—reading and math—covered by NCLB? If the answer to that question is "yes," then that would call for, at the very least, a review of the original achievement levels to see if they still would apply to a new framework. Second, if the original achievement levels cannot apply, then new ones could be adopted by the Board. However, that could disrupt trend lines. Can the original and new achievement levels be linked? Should there be changes in the Governing Board achievement levels policy based on what has been learned in the 1990s? For example, should the Governing Board continue to use the original ACT/NAEP methodology for developing the levels or should a different method be employed? We will explore each of these issues in turn.

New Levels for New Frameworks?

The 2000 and 2003 NAEP mathematics assessments continued the trend started in 1990 with the original mathematics framework developed under the Board's guidance. Figure 4 (earlier in this paper) displays the ALDs for the original achievement levels in grade 12. However, in 2005 the Board developed and adopted a new framework for the grade 12 mathematics assessment. According to Board documents, this change was warranted because the grade 12 mathematics curriculum had become more challenging over the last decade. This movement was based partly on the international studies that reported less than stellar performance for U.S. students in both mathematics and science. In addition, curricula changes at the state and local levels had the effect of sequencing more difficult content at lower grade levels, consequently making grade 12 content more demanding and thus out of step with the grade 12 original NAEP framework.

Whether or not the old and new assessments were aligned enough to be placed on the same reporting scale, and whether the new assessment could be reported using the original achievement levels is an empirical question. Even though the statistical methodology of NAEP (item response theory) can accommodate minor shifts in item content, difficulty, and format, NAEP always undertakes an extensive empirical effort to undergird major decisions when important changes are made. In this case, the evidence showed that the two math assessments could not be closely aligned because of new content, changes in administration and block design, and different rules on calculator usage (National Assessment Governing Board, 2007).[20] As a result, the Board developed new achievement levels for the 2005 grade 12 assessment in math. Figure 5 displays the ALDs for the new grade 12 assessment. The reader can compare the differences in the grade 12 ALDs and judge the degree of difference in content.

Figure 5
Grade 12 Mathematics ALDs

(Excerpt from 2005 Nation's Report Card: 12th Grade Mathematics)

Basic Twelfth-grade students performing at the *Basic* level should be able solve mathematical problems that require the direct application of concepts and procedures in familiar situations. For example, they should be able to perform computations with real numbers and estimate results of numerical calculations. These students should also be able to estimate, calculate, and compare measures and identify and compare properties of two- and three-dimensional figures, and solve simple problems using two-dimensional coordinate geometry. At this level, students should be able to identify the source of bias in a sample and make inferences from sample results; calculate, interpret, and use measures of central tendency; and compute simple probabilities. They should understand the use of variables, expressions, and equations to represent unknown quantities. They should be able to solve problems involving linear relations using tables, graphs, or symbols; and solve linear equations involving one variable.

Proficient Students in the twelfth grade performing at the *Proficient* level should be able to select strategies to solve problems and integrate concepts and procedures. These students should be able to interpret an argument, justify a mathematical process, and make comparisons dealing with a wide variety of mathematical tasks. They should also be able to perform calculations involving similar figures including right triangle trigonometry. They should understand and apply properties of geometric figures and relationships between figures in two and three dimensions. Students at this level should select and use appropriate units of measure as they apply formulas to solve problems. Students performing at this level should be able to use measures of central tendency and variability of distributions to make decisions and predictions, calculate combinations and permutations to solve problems, and understand the use of normal distribution to describe real-world situations. Students performing at the *Proficient* level should be able to identify, manipulate, graph, and apply linear, quadratic, exponential, and inverse proportionality ($y = k/x$) functions; solve routine and one-routine problems involving functions expressed in algebraic, verbal, tabular, and graphical forms; and solve quadratic and rational equations in one variable and solve systems of linear equations.

Advanced Twelfth-grade students performing at the *Advanced* level should demonstrate in-depth knowledge of mathematical concepts and procedures represented in the framework. They can integrate knowledge to solve complex problems and justify and explain their thinking. These students should be able to analyze, make and justify mathematical arguments, and communicate their ideas clearly. *Advanced* level students should be able to describe the intersections of geometric figures in two and three dimensions, and use vectors to represent velocity and direction. They should also be able to describe the impact of linear transformations and outliers of measures on central tendency and variability, analyze predictions based on multiple data sets, and apply probability and statistical reasoning in more complex problems. Students performing at the *Advanced* level should be able to solve or interpret systems of inequalities, formulate a model for a complex situations (e.g., exponential growth and decay), and make inferences or predictions using the mathematical model.

Developing new achievement levels in grade12 mathematics was not an easy decision, but it exemplifies the complexities ahead of the Board as they review, renew, and/or revise the current frameworks in nearly a dozen different NAEP subject areas. In 1989, the Board maintained the initial trend lines (from approximately 1971 to 1989) by preserving the Long-Term Trend assessment, still administered and reported separately from the main NAEP.[21] If that approach is used again, there could be two long-term trend lines, 1971 to 1989 and 1990 to 2005. The difference this time is that so much is riding on the trend lines from 1990, because, the NAEP state assessments began that year and the NCLB accountability requirements started in 2001. Admittedly, NCLB accountability applies to three subjects (reading/language arts, mathematics, and science). However, currently, the 2005 mathematics results in grade 12 are reported on a single-grade scale, with no links to earlier mathematics assessments. This approach in other subjects and other grades may not be entirely suitable to moving the accountability embedded in NCLB forward. More recently, a second new mathematics framework was developed for use in the NAEP 2009 cycle along with a new framework in reading at all three grades.[22] A proliferation of frameworks without a clear plan to develop a unified assessment program with meaningful achievement levels will likely not serve the National Assessment well.

To Link or Not to Link?

Equating the results of one assessment to another, or one grade to another, has been statistically possible for a number of years. However, the procedure is not without its difficulties and critics. Equating is as much an art as it is a science. Depending on the decisions made in the linking process, results can vary substantially. At the August 2007 Board meeting, the Board heard a presentation on NAEP trend line issues that included a discussion of linking, achievement levels, and scaling in the context of the 2009 reading assessment that will employ a new NAEP reading framework (National Assessment Governing Board, 2007).

Representatives from the Education Information Advisory Committee of the Council of Chief State School Officers, the NAEP Validity Studies Panel, and the NAEP Design and Analysis Committee (an advisory committee to the current NAEP contractor, ETS) presented findings from studies and discussions that each group had conducted on the issue.

The recommendation resulting from the panel presentation was that new achievement levels should be set based on the new 2009 reading framework. The question regarding whether or not to link the old and the new levels is still open and under study by NCES and others. Whatever is decided for the 2009 reading assessment must be carefully and cautiously considered, since it paves the way for dealing with other subject areas and sets a precedent for the future of the NAEP achievement levels.

Original Methodology or New Methodology?

By 2003, the new NCLB law was beginning to take hold. Many states had received approval by the U.S. Department of Education to move forward with their own assessments and to develop accountability standards aligned to their assessments. As required under the law, states began to develop performance standards using NAEP's standards as a model. The law requires two higher levels (e.g., Proficient and Advanced) and at least one lower level in order to

report on the performance of the full distribution. Those labels are not required and, indeed, many states use quite different labels to describe student performance, (e.g., Levels I, II, III, IV, or other descriptors such as Low, Intermediate, High). Also, states use variations on the number of levels (from three to six levels) (Perie, 2008). A review of state performance standards shows that 12 states use a 5-level system, 29 use a 4-level system, 10 use a 3-level system, and 1 uses a 6-level system. All states with three or four levels have positioned the required "Proficient" at the second highest level. Of the 13 states that use 5 or 6 levels, 9 have positioned the required "Proficient" at the third highest level, i.e., three levels down from the top level, thus having the likely effect of depressing the definition of Proficient.

Table 1 summarizes the number of performance levels by state, and where Proficient is positioned in the distribution of performance. Further, the definition of Proficient can vary from state to state, and is not required to reflect the NAEP definition. Both of these aspects, the positioning and definition of Proficient, affect the relationship of NCLB and the achievement levels. This is no small matter and its resolution would go a long way to resolving the disparity between NAEP results for the states and the states' performance on their approved NCLB assessments.

Table 1
State Performance Standards under NCLB*

	No. of Levels			
Position of Proficient	3 levels N = 10	4 levels N = 29	5 levels N = 12	6 levels N = 1
2nd highest level	CO GA IN IA MD NJ PR TN TX VA	AL AK AZ AR DC HI ID IL KY ME MA MI MS MT NB NE NH NM NY NC ND OK PA SC SD UT WA WI WY	CA MO OR VT	RI
3rd highest level			CT DE FL KS LA MN OH WV	

*Adapted from Perie (2008).
Notes:
Example of 3 levels – Does not meet standard, meets standard, exceeds standard.
Example of 4 levels – Far below Proficient, below Proficient, Proficient. Advanced.
Example of 5 levels – Levels 1, 2, 3, 4, 5, where Level 3 = Proficient.

In 2003, the Board awarded a contract to ACT to develop the achievement levels on the grade 12 mathematics assessment. At the November 2004 meeting, the Executive Director's report contained a summary of the Board's ongoing work, including the new contract. ACT had already completed a number of pilot studies to look at the viability of moving to a new methodology for NAEP. A Committee on Standards, Design, and Methodology (COSDAM) meeting prior to the November meeting fully discussed the pros and cons of such a move, and recommended the change from Angoff to the MAP Mark method of setting achievement levels (Schultz and Mitzel, n.d.). This method is a variant on the Bookmark method, used by many states, in which items are arrayed according to difficulty and judges pick a "passing" score. (Nellhaus, 2000).[23]

According to the Board meeting transcript, the following factors influenced the Board's decision:

1. *"The new NAGB framework for 12th grade mathematics is sufficiently different from the previous framework to require a new trend line;*
2. *The MAP Mark and item rating methods are likely to produce valid outcomes;*
3. *The MAP Mark approach is based on the bookmark method that is widely used by states in setting achievement levels; and*
4. *MAP Mark is less complex and easier to explain and defend than the modified Angoff method"* (National Assessment Governing Board, 2004).

A review of other publicly available Governing Board documents provided no other operational or policy reasons for the shift from Angoff to MAP Mark. According to a staff member, the Board reviewed evidence for both the ACT/Governing Board method and the MAP Mark method and judged them to be very similar. The Board's committee responsible for the achievement levels (COSDAM) was interested in the fact that the newer method could be implemented in a shorter time period (four days rather than five).[24] This new method was subsequently used for the grade 12 Economics assessment in 2006. Table 2 displays the NAEP achievement levels results in the first year the levels were set for all NAEP assessments since 1992.

Table 2
Governing Board Achievement Levels Cut Scores by Content and National Performance at or Above Levels in Initial Year Reported

Content/ Initial Year	Cut Score and Percent at or Above Basic	Proficient	Advanced
	Reported on NAEP 0 to 500 Cross-Grade Scale		
Math/1992			
Grade 4	214/59%	249/18%	282/2%
Grade 8	262/58	299/21	333/3
Grade 12	288/64	336/15	367/2
Reading/1992			
Grade 4	208/62	238/29	268/6
Grade 8	243/69	281/29	323/3
Grade 12	265/80	302/40	346/4
Geography/1994			
Grade 4	187/70	240/22	276/3
Grade 8	242/71	282/28	315/4
Grade 12	270/70	305/27	339/2
U.S. History/1994			
Grade 4	195/64	243/17	276/2
Grade 8	252/61	294/14	327/1
Grade 12	294/43	325/11	355/1
	Reported on NAEP 0 to 300 Within-Grade Scale		
Science /1996			
Grade 4	138/67	170/29	204/3
Grade 8	143/61	170/29	207/3
Grade 12	145/57	178/21	210/3
Civics/1998			
Grade 4	136/69	177/23	215/2
Grade 8	134/70	178/22	213/2
Grade 12	139/65	174/26	204/4
Writing/1998			
Grade 4	115/84	176/23	225/1
Grade 8	114/84	173/27	224/1
Grade 12	122/78	178/22	230/1
	Reported on NAEP 0 to 300 Single-Grade Scale		
Math/2005			
Grade 12	141/61	176/23	216/2
Economics/2006			
Grade 12	123/79	160/42	208/3

Sources: Loomis and Bourque (2001b); Grigg, Donahue, and Dion (2007); Mead and Sandene (2007).

Impact of Achievement Levels Over the Last 20 Years

Where are we after 20 years of standards on NAEP? One thing is certain, the sky did not fall (as predicted by some naysayers). *The Nation's Report Card,* while not on the *New York Times* bestseller list, certainly makes news headlines when the results are released. Education shows evidence of being somewhat better off than it was in 1989. The dialog has picked up speed about quality education and what we as a nation should expect from the system. Not that achievement levels can take all the credit, but they may have helped. The author believes that the Governing Board's achievement levels can stand tall and take credit for a number of advances in the standards movement. The achievement levels:

- improve the form and use of NAEP;
- serve policy decisionmaking efforts at the local, state, and federal levels;
- serve as a model for state assessments under NCLB;
- improve the standard-setting enterprise.

Clearly, the levels have improved the form and use of NAEP. Although there will always be some concerns about the "clarity" of the levels, for most users they are far better than what existed before, namely, average scale scores with descriptions at the mean, the 84th percentile, etc. The levels have tried to answer the question, "How good is good enough?" We may not agree with the answer to the question any more than we agree with how clean our air should be or how "green" our automobiles should be. Nevertheless, we need the answers to those kinds of questions in order to measure our progress toward national and world goals. The same is true in education: we need the answer to "How good is good enough?" to measure our progress as a nation toward an educated citizenry.

The levels also serve a very important policy function at all levels of education. The importance of NAEP results that policymakers may use cannot be overstated. The move in 1990 to add the Trial State Assessments to the mix, and the more recent move to add the Trial Urban District Assessments as well, is a testimony to NAEP's value in general and the achievement levels in particular. The levels have influenced legislation at the federal and state levels, they have been used to provide snapshots of academic performance by news outlets such as *Education Week,* and they are used by private foundations and think tanks whose mission is to keep the public informed on the condition of U.S. education. See, for example, the recent American Institutes for Research report on linking international mathematics performance with NAEP mathematics performance in urban school districts (Phillips and Dossey, 2008).

The levels were also invoked as a model for states in the NCLB legislation. States are required to move toward a standards-based approach in reporting the state's progress in reaching the accountability goals of NCLB. They must set levels of performance similar to the achievement levels in that they span the distribution of performance. All states now have those levels, along with assessments and aligned content standards.

Finally, the Governing Board achievement levels have improved the standard-setting enterprise enormously. In 1989 there was a paucity of methods that could be used to set standards; no one had heard of PDs or ALDs; use of feedback during the standard-setting process

was slim to none; using broadly based panels was never done; and, in many instances, the process to set standards was done "behind closed doors." All that has changed; in no small measure, that change is due to the work of the Governing Board.

More methods are available now than ever before—more than one-third of the states use PDs and all use ALDs; virtually all methods use feedback in the process; and, in many instances, the composition of panels has been expanded and the process is more open. In addition, the level of discourse in the professional literature has increased considerably.

What Does the Future Hold?

The National Assessment Governing Board is to be congratulated on this 20th anniversary for its work on the NAEP performance standards. Over the last 20 years, the Board has developed, defended, and disseminated the achievement levels. But—five Administrations and 10 congressional sessions later—the work is not done. Many major issues are still not resolved. And even though resources may be limited, it is important to keep in mind that progress can be measured only if the yardstick used for measuring is valid and reliable, and reflects the current best practice in measurement technology. Here is the author's list of what should be done:

1. Work to remove "trial" from the next NAEP reauthorization.
2. Resolve the issues of trend line and single-grade scale.
3. Explore the possibility of linking new and old assessments, especially in the short-term trend.
4. Resolve the discrepancies in performance between NAEP and state assessment results, or explain them much better.
5. Explore frameworks and achievement levels in the context of 21st century skills, and build new frameworks in accordance with a deliberate plan that preserves the integrity of the NAEP program for the foreseeable future.
6. Mount a robust research agenda on new standard-setting methodologies and publish the results in the professional literature. Items on the agenda should include:

    a. Evidence of bias or lack thereof in the overall standard-setting results.
    b. Impact of various criteria for selecting anchor items and ordering them on the scale.
    c. Impact of item formats and methodology and their interactions on the levels.
    d. Validity data, validity data, and more validity data

Endnotes

1. Much of the information in this section is taken from Jones and Olkin (2004).

2. This thumbnail sketch of the early NAEP decades does not allow extensive coverage of the pressing domestic issues during this period. Suffice it to say that events such as the landmark 1954 Supreme Court decision on desegregating public schools, the development of the 1965

Great Society programs of President Johnson, passage of the 1965 Title I legislation, and the 1966 Coleman report on educational inequality also influenced the crafting of a national measure of educational progress.

3. The 1990 NAEP cycle was the first developed under the Board's policy. This assessment cycle reflected some significant changes, including developing assessment frameworks through a national consensus process, moving from age-based sampling and reporting to grade-based sampling and reporting, preserving the first 20 years of NAEP assessments as the Long-Term Trend, and reporting NAEP performance in terms of the achievement levels rather than using statistical indicators of the national distribution (e.g., means and standard deviations). See Phillips et al. (1993) for more specific information on this topic.

4. A. Shanker, personal communication (1989).

5. For a more detailed discussion of the distinctions between anchor level and achievement levels, see Bourque (2007).

6. The three achievement levels have been used in all NAEP reports since 1992. However, there is also a "Below Basic" level that the Board does not view as a NAEP standard; it is included only to complete the reporting of the full distribution of student performance.

7. One must distinguish here between comparability of standards and comparability of results. The former is a policy requirement for NAEP. However, comparability of results cannot be guaranteed since achievement is the result of many factors, including curriculum emphasis, time allocations in schools, grade structure of state curricula, the percent of special needs participation in NAEP, and other issues beyond the control of NAEP.

8. A Foreword to the Achievement Levels Policy and Implementation Guidelines was added and adopted by the Board in August 2007; it explains and updates policy changes between 1990 and 2007.

9. This paper will refer to these as achievement level descriptions (ALDs), but in the standard setting literature they are commonly referred to as performance level descriptions (PLDs).

10. The 1992 standard setting did employ operationalized versions of the policy definitions. However, they were developed during the standard-setting meeting by three distinct grade-level groups and, as such, varied in sharpness of the language, degree of specificity, and format. These working versions were subsequently validated by an independent group to sharpen the language and to provide durability to the descriptions.

11. It should be noted that the use of borderline descriptions discussed here applies only to the original method of setting standards adopted by the Board in 1990 (the Angoff method). The method adopted more recently for 2005 grade 12 mathematics and 2006 economics does not employ borderline descriptions.

12. This approach is in contrast to other "norm-referenced" approaches in which targets or quotas are set without much regard for the content of the standards. See Hambleton and Pitoniak (2006), Plake (2005), and Goodman and Hambleton (2005) for further information.

13. Although data are not a requirement, the Angoff method (as well as most other methods) is almost always used with some data—even field test data—to provide feedback to panelists (a reality check) during the process.

14. The Angoff method as originally suggested by Angoff has almost always been accommodated (a.k.a. modified Angoff) to particular circumstances.

15. ACT technical staff included Robert Brennan and Mark Reckase in addition to other ACT technical staff who served on an internal Technical Advisory Team.

16. Initial members on TACSS included William Brown, South Carolina Department of Public Instruction; Robert Forsyth, University of Iowa; Ronald Hambleton, University of Massachusetts; Eugene Johnson, ETS; Michael Kane, University of Wisconsin; Brenda Lloyd, University of Virginia; and William Mehrens, Michigan State University.

17. Consequences data were not allowed to impact the process in any meaningful way until the 1998 cycle. Initially, the Board reserved the use of those data for itself; however, by 1998, the Board was convinced that the use of consequences data could improve the panelists' judgments, and allowed its use in the final round of the process.

18. The purpose of this technical paper was not focused on scale anchoring or standard setting per se, but its results lend support to the statements here and their applicability to both approaches for reporting.

19. Full disclosure would admit that there is another interpretation. NCES, which is responsible for the administration of NAEP (but not policy) argues that as a statistical agency it should report only NAEP data, without any accompanying judgments about "how good is good enough." And so "developmental" also has connotations of "use with caution."

20. NCES did not mount any specific bridge studies to gather empirical data on the impact of the changes.

21. The Long-Term Trend assessment uses smaller age-based sample sizes, includes only a national assessment but no state assessments, and preserves the same exclusion rules that had been initially used in NAEP before 1990.

22. S. Loomis, personal communication (2009).

23. A comprehensive review of the literature pertaining to the Bookmark method can be found in Karantonis and Sireci (2006).

24. S. Loomis, personal communication (2009).

References

ACT (1997). *Developing achievement levels on the 1998 NAEP in civics and writing.* Iowa City, IA: Author.

Alexander, L. (1986). *Time for Results*. (Available from the National Governors Association, Hall of the States, 444 North Capitol Street, Suite 267, Washington, DC 20001).

Alexander, L. and James, H.T. (Eds.) (1987). *The nation's report card: Improving assessment of student achievement. Report of the study group, with a review of the report by a committee of the National Academy of Education.* Cambridge, MA: National Academy of Education.

American College Testing (1992). *Setting achievement levels on the 1992 National Assessment of Educational Progress in reading, mathematics, and writing: Design document.* Iowa City: Author.

American College Testing (1995). *NAEP reading revisited: An evaluation of the 1992 achievement level descriptions.* Washington, DC: National Assessment Governing Board.

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

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (pp. 508–600). Washington, DC: American Council on Education.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research* 56: 137–172.

Bourque, M.L. (2007). Review and commentary on the HUMRRO report. Paper prepared for the California State Board of Education. Available at http://www.cde.ca.gov/BE/ag/ag/yr07/documents/may07item09mlbourque.doc.

Cizek, G.J. and Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications.

Ebel, R.L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Glaser, R. (1987). Commentary by the National Academy of Education (p. 58), In *The Nation's report card: Improving assessment of student achievement. Report of the study group, with a review of the report by a committee of the National Academy of Education,* edited by L. Alexander and H.T. James. Cambridge, MA: National Academy of Education.

Goodman, D. and Hambleton, R.K. (2005). Some misconceptions about large-scale educational assessments. In R. Phelps (Ed.), *Defending standardized testing* (pp. 91–110). Mahwah, NJ: Erlbaum.

Grigg, W., Donahue, P., and Dion, G. (2007). The nation's report card: 12th grade reading and mathematics 2005. (NCES 2007–468). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Hambleton, R.K. (2001). Setting performance standards in educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 90–116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.L. and Bourque, M.L. (1991). *The LEVELS of mathematics achievement,* Vol. III. Washington, DC: National Assessment Governing Board.

Hambleton, R.K. and Pitoniak, M.J. (2006). Setting Performance Standards. In R.L. Brennan (Ed.), *Educational measurement,* 4th ed. Westport , CT: Praeger Publishers.

Jones, L.V. and Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Karantonis, A. and Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice* 25(1): 4–12.

Kolstadt, A., Cohen, J., Baldis, S., Chan, T., DeFur, E., and Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

Linn, R.L., Koretz, D.M., Baker, E.L., and Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics.* Los Angeles, CA: Center for the Study of Evaluation, University of California.

Loomis, S.C. and Bourque, M.L. (2001a). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In C.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–218). Mahwah, NJ: Lawrence Erlbaum.

Loomis, S.C. and Bourque, M.L. (Eds.) (2001b). *National Assessment of Educational Progress achievement levels: 1992–1998.* Washington, DC: National Assessment Governing Board.

McNeil, M. (August 13, 2008). Exit Scramble. *Education Week.* Retrieved October 8, 2008, from http://www.Edweek.org.

Mead, N. and Sandene, B. (2007). The nation's report card: economics 2006. (NCES 2007–475). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Messick, S., Beaton, A.E., and Lord, F.M. (1983). *A new design for a new era.* Princeton, NJ: Educational Testing Service.

Mills, C.N. and Jaeger, R.J. (1998). Creating descriptions of desired students achievement when setting performance standards. In L. Hansche (Ed.) *Handbook for development of performance standards* (pp. 75–85). Washington, DC: U.S. Department of Education and the Council of Chief State School Officers.

National Academy of Education. (Glaser, R., Linn, R., and Bohrnstedt, Eds). (1993). *Setting performance standards for student achievement.* Panel on the Evaluation of the NAEP Trial State Assessment. Washington, DC: National Academy of Education.

National Academy of Education. (1997). *Assessment in transition: Monitoring the nation's educational progress.* Mountain View, CA: Author.

National Assessment Governing Board. (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board. (November 2004). *Transcript of National Assessment Governing Board Meeting* (pp. 5–6). Washington, DC: Author.

National Assessment Governing Board. (August 2007). *Transcript of National Assessment Governing Board Meeting* (pp. 8–14). Washington, DC: Author.

National Center for Education Statistics (1994). *NAEP's 1992 reading report card for the nation and the states.* Washington, DC: Author.

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: Author.

National Council for Teachers of Mathematics (2000). *Principles and Standards for School Mathematics.* Washington, DC: Author

National Governors Association (1991). *Educating America: State strategies for achieving the national education goals.* Washington, DC: Author.

Nedelsky, L. (1954). Absolute grading standards for objectives tests. *Educational and Psychological Measurement* 14: 3–10

Nellhaus, J.M. (2000). States with NAEP-Like Performance Standards. In M.L. Bourque and S. Byrd (Eds). *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements* (pp. 99–130). Washington, DC: National Assessment Governing Board.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (Eds.), (1998). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Committee on the Evaluation of the National Assessment of Educational Progress, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Perie, M. (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice* 27(4): 15–29.

Phillips, G.W. and Dossey, J.A. *Counting on the future: International benchmarks in mathematics for American school districts.* Available at www.air.org.

Phillips, G.W., Mullis, I.V.S., Bourque, M.L., Williams, P., Hambleton, R.K., Owen, E.H., and Barton, P.E. (1993). Interpreting NAEP scales. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Plake, B.S. (2005). Doesn't everybody know that 70% is passing? In R. Phelps (Ed.), *Defending standardized testing* (pp. 91–110). Mahwah, NJ: Erlbaum.

P.L. 100–297 (1988). National assessment of educational progress improvement act. (Article No. USC1221). Washington, DC.

P.L. 103–382 (1994). Improving America's School Act of 1994. Sec. 411.

P.L. 107–110. No Child Left Behind Act of 2001. Sec. 602. Retrieved February 1, 2003, from http://www.nagb.org/about/plaw.html.

Reckase, M.D. (2000). *The evolution of the NAEP achievement level setting process: A summary of the research and development efforts conducted by ACT.* ACT: Iowa City, IA.

Shultz, E.M. and Mitzel, H.C. (no date). *The mapmark standard setting method.* Unpublished manuscript.

Stufflebeam, D.L., Jaeger, R.M., and Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural 1990–81 effort to set achievement levels on the National Assessment of Educational Progress.* Prepared for the National Assessment Governing Board, August 23.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations.* GAO/PEMD–993–12. Washington, DC: Author.