DOCUMENT RESUME

ED 427 027 TM 029 391

AUTHOR	Hansche, Linda N.
TITLE	Handbook for the Development of Performance Standards: Meeting the Requirements of Title I.
INSTITUTION	Council of Chief State School Officers, Washington, DC.; Office of Elementary and Secondary Education (ED), Washington, DC.
ISBN	ISBN-1-884037-53-4
PUB DATE	1998-09-00
NOTE	115p.; With contributions by Ronald K. Hambleton, Craig N. Mills, Richard M. Jaeger, and Doris Redfield.
PUB TYPE	Guides - Non-Classroom (055) -- Tests/Questionnaires (160)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Tests; *Compensatory Education; Cutting Scores; Educational Assessment; Educational Practices; Educational Research; *Educationally Disadvantaged; Elementary Secondary Education; Federal Legislation; Low Income Groups; Performance Factors; *Standards; *State Programs
IDENTIFIERS	*Improving Americas Schools Act 1994 Title I; Standard Setting

ABSTRACT

Title I of the Improving America's Schools Act (IASA) of 1994 provides funds for schools with large concentrations of children from low-income families. A fundamental requirement is that children served by Title I funds must be educated according to the same academic standards as all other students. This handbook focuses on methods for developing performance standards in the aligned system of standards and assessments required by IASA Title I. The handbook aims to capture the best of current practice, without relying solely on the published literature, by drawing on the experiences of educators and recent research. The first section (chapters 1-4) defines performance standards in the context of an aligned education system and provides advice for developing a system of performance standards. Chapters introduce the idea of performance standards as a system, provide background about Title I legislation, and define terms related to performance standards. The second section (chapters 5-8) contains several state stories about initiating and developing performance standards and standards-based assessment programs. Chapters focus on Colorado, Maryland, Oregon, and Wyoming. The third section (chapters 9-10) contains the work of nationally recognized researchers in the field of assessment. Chapter 9, "Creating Descriptions of Desired Student Achievement When Setting Performance Standards" by Craig N. Mills and Richard M. Jaeger, describes a method for developing performance standards. Chapter 10, "Setting Performance Standards on Achievement Tests: Meeting the Requirements of Title I" by Ronald K. Hambleton, synthesizes research related to cutting scores. Most chapters contain references. Four appendixes present the instruments. (Contains 16 figures and 4 tables.) (SLD)

Meeting the Requirements of Title I

Handbook for the Development of Performance Standards

Prepared for the
U.S. Department of Education and
The Council of Chief State School Officers

Handbook for the
Development of Performance Standards:

Meeting the Requirements of Title I

Prepared for the U.S. Department of Education and
The Council of Chief State School Officers

by

Linda N. Hansche

With contributions by
Ronald K. Hambleton
Craig N. Mills
Richard M. Jaeger
Doris Redfield

As part of its initiative to assist states in implementing provisions of the reauthorized Elementary and Secondary Education Act, the U.S. Department of Education has worked closely with the Council of Chief State School Officers (CCSSO) to develop technical assistance and documents addressing key aspects of Title I. This report is the result of a collaborative effort between the Department of Education's Office of Elementary and Secondary Education and CCSSO's State Collaborative on Assessment and Student Standards (SCASS), Comprehensive Assessment Systems for IASA Title I/Goals 2000.

CCSSO is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five extrastate jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its standing committees and special task forces, CCSSO responds to a broad range of concerns about education and provides leadership on major education issues.

The views and opinions expressed in this report are not necessarily those of the U.S. Department of Education, CCSSO, or the SCASS.

Council of Chief State School Officers

Gordon M. Ambach, Executive Director

Wayne N. Martin, Director
State Education Assessment Center

Edward R. Roeber, Director
Student Assessment Programs

Phoebe C. Winter, Project Director
Comprehensive Assessment Systems for
IASA Title I/Goals 2000

U.S. Department of Education, Office of Elementary and Secondary Education

Gerald Tirozzi, Assistant Secretary

Mary Jean LeTendre, Director
Compensatory Education Programs

Thomas Fagan, Director
Goals 2000 Programs

ISBN #1-884037-53-4

September 1998

Publication design by Frost Associates, Ltd.
Bethesda, Maryland

Contents

Preface

In 1994 Congress passed a legislative package that was focused on reforming education in the United States and reauthorized the Elementary and Secondary Education Act. The goals of the legislation, the Improving America's Schools Act (IASA), are to underscore the philosophy that schools are responsible for educating all students and to ensure that all students learn to high standards. IASA includes provisions for (1) rewarding schools that demonstrate improvement, (2) providing corrective action for poor performance, (3) increasing parental involvement, (4) increasing high-quality professional development, and (5) supporting instructional improvement through flexible practice.

Title I of IASA provides funds for schools with large concentrations of children from low-income families. A fundamental requirement is that students served by Title I funds must be educated according to the same rigorous academic standards as all other students. The same coordinated system of *content standards* (what students are expected to know and be able to do), *performance standards* (descriptions of "how good is good enough"), and *assessments* that states use to evaluate the education of children in general must also apply to students served by Title I.

The purpose of this handbook is to focus on methods for developing performance standards in the aligned system of standards and assessments required by IASA/Title I. Reasons for the handbook's emphasis on performance standards include (1) the needs of states concerned with successfully implementing legislative requirements and their underlying intent, (2) inexperience with and the relative difficulties of developing and implementing defensible performance standards, and (3) the importance of evaluating the extent to which all students achieve optimal learning and performance.

This handbook is based on best practice and current research. Much of the information related to the procedures and processes for developing content standards and performance standards is still emerging. As educators progress in their investigations into how these new standards are to be created, the insights emerging from firsthand experience are vital and most informative.

One goal of this handbook is to capture the most current practice, without relying solely on the published literature. By the time studies make it into print, their methods are already being improved upon. This handbook attempts to document what is sometimes referred to as "fugitive" literature and practice (i.e., procedures and practices that go on within states or other education agencies that are never formally written about). Information about particular programs and processes may be presented at conferences and professional meetings, and handouts are often provided, but complete information is seldom presented in a manner that can be distributed as a comprehensive or formal piece to guide similar programs along the path toward success.

Contents and Intended Audience

Because Title I is intended to allow states and local agencies maximum flexibility in designing programs appropriate for their students, there are many ideas about what effective programs look like. This handbook does not claim to have the definitive answer to the challenges presented by the development and implementation of comprehensive systems such as those called for by IASA. Variations are not only acceptable but encouraged by the legislation. Arriving at variations that work for a particular jurisdiction requires leadership from the primary audiences for this handbook — Title I coordinators and directors of curriculum and assessment programs. Leaders from other stakeholder groups may also find the information interesting and/or useful.

Each chapter of this handbook is designed to be a stepping stone toward successfully implementing Title I requirements for performance standards. Section I of the handbook defines performance standards in the context of an aligned education system and provides advice for developing a system of performance standards. Chapter 1 introduces the reader to the idea of performance standards as a *system*, setting the context for the remainder of Section I. Chapter 2 provides background about the Title I legislation related to performance standards and essential program requirements. Chapter 3 defines terms in an attempt to foster mutual understanding and consis-

tent communication within this document. (Be aware that the definitions provided are not "standard" yet. In some cases, the terms and their precise meanings are still evolving.) Chapter 4 provides a framework for developing performance standards, one component of the required comprehensive system.

Section II contains several state stories, compiled and edited by Doris Redfield. Much of the content of these stories has been graciously "donated" by individuals who have worked with development and implementation processes in their own states. The stories are about initiating, developing, and implementing standards-based assessment programs, transitioning from existing programs to new ones, or accommodating the alignment of state and/or local requirements within the context of IASA goals.

Section III contains the work of nationally recognized researchers in the field of assessment. The way these researchers use certain terms, which varies somewhat from how the same terms are used in the rest of the document, reflects how psychometricians and other measurement experts/researchers use the terms to communicate precisely with other experts in the field. These variations in vocabulary underscore the emerging nature and development of content standards, performance standards, and performance assessments. The field is fast changing, and agreement about the meaning and use of these terms has evolved, literally, in the months since the papers were written. In Section III, readers should rely on the definitions provided by the authors rather than on the definitions in earlier chapters, which are more general and less technical.

Chapter 9 was written by Craig N. Mills, Educational Testing Service, and Richard M. Jaeger, University of North Carolina, Greensboro. This chapter describes a method for developing performance descriptors used to define levels of performance (e.g., proficient, advanced). Chapter 10 was written by Ronald K. Hambleton, University of Massachusetts, Amherst. He synthesizes findings in the field of standard setting as it pertains to setting "cut scores" (i.e., those scores that determine which test results fit into which levels or categories of performance) and relates this work to Title I requirements.

In its entirety, the handbook covers the multiple aspects of developing performance standards within an aligned system, but each chapter focuses on a particular topic. Because each chapter is intended to stand alone as a resource on its topic, there is duplication in several chapters.

Caveats

There is no one "right" way to implement standards; however, there are some ways that are more right than others. For example, some procedures are *not* appropriate:

- Do not implement performance standards that are based only on cut scores on a standardized norm-referenced test, because it is unlikely that any off-the-shelf test will fully align with the breadth and depth of a state's or local system's content standards.
- Do not rely on a single measure of student learning to determine student proficiency.
- Do not produce overly general content and performance standards in which each progressive level is "more of the same" (i.e., descriptions of quantitative differences only, without consideration of quality, depth, or complexity).

As programs are planned, built, implemented, and reported, they must meet both current needs and long-range needs. They should consider at least two dimensions of alignment:
(1) alignment of student, classroom, school, local, state, and national learning goals; and
(2) alignment of content standards, curricula and instruction, performance standards, and assessments.

It is important to stay focused on the guiding principle underlying this effort: creating, implementing, and accounting for optimal learning opportunities so that *all* students will learn.

Section I:
Performance Standards in Context

Chapter 1. Introduction to Performance Standards

This handbook provides guidelines and options for meeting Improving America's Schools Act (IASA)/Title I requirements, and it also addresses the development of systems of content standards, performance standards, and assessments in general. Because performance standards cannot stand alone, content standards and assessments are addressed insofar as they interface with performance standards. Therefore, the handbook addresses issues of performance standards as part of an aligned system.

Comprehensive assessment systems necessitate the alignment of *content standards* (what students are expected to know and be able to do) and *performance standards* (the descriptions of "how good is good enough"). The vehicles for operationalizing the content and performance standards are familiar — curriculum, instruction, and assessment. In the context of content standards, *curriculum* refers to the subject matter and related skills and processes, and the scope and sequence for delivery. A curriculum is used by teachers to guide instruction and to serve as the teacher's "technical manual."

A curriculum is the "what" and "when" of teaching certain content, and instruction is the "how." Instruction, then, is the mechanism by which teachers deliver the curriculum and engage students in learning.

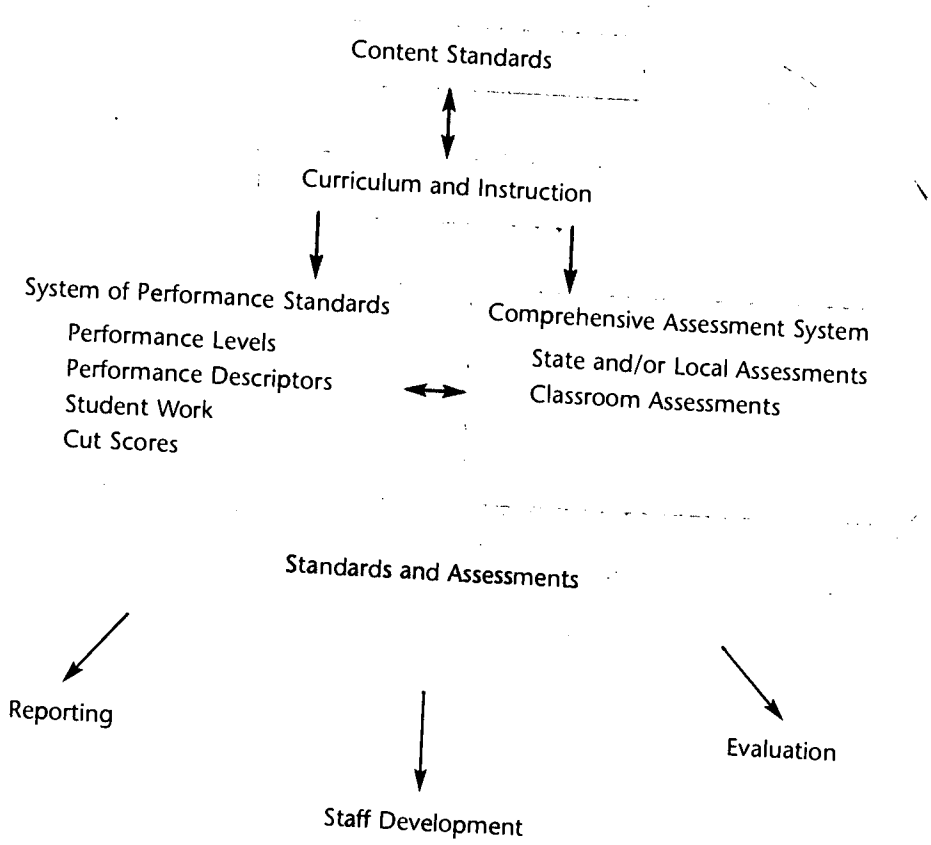
Assessment must be more than a product or series of tests; it must be a process matched directly to the content standards and used to report student progress in terms of the performance standards that are directly tied to the content standards. A comprehensive assessment system of this kind requires that all the parts are linked and work together as a whole.

Figure 1.1 shows the relationships among components of an aligned education system focused on student learning.

Systems of Performance Standards

There is little consistency in the education community in the way terms are used when discussing performance standards. The term *performance standard* itself is a case in point. To test developers and psychometricians, *performance standard* usually refers to the point on a test score scale that separates one level of achievement (e.g., pass) from another (e.g., fail), identified through a technically sound process. To educators involved in the development of curriculum and instruction, *performance standard* often means a description of what a student knows and can do to demonstrate proficiency on a content standard or cluster of content standards. To others, the term *performance standard* indicates examples of student work that illustrate world-class performance.

Figure 1.1: System for student learning



In this handbook, "performance standard" is defined as a *system* that includes performance levels, descriptions of student performance, examples of student work, and cut scores on assessments. A system of performance standards operationalizes and further defines content standards by connecting them to information that describes how well students are doing in learning the knowledge and skills contained in the content standards.

Performance standards answer the question, How good is good enough? A system of performance standards includes the following components:

- performance levels — labels for each level of achievement
- performance descriptors — narrative descriptions of performance at each level
- exemplars — examples of student work from a representative sample of all students that illustrate the full range of performance at each level
- cut scores — scores on a variety of assessments that separate the different levels of performance

The components of the system are highly interrelated. As a system, performance standards describe student achievement at different levels; these levels are operationalized by assessments used to measure student achievement and exemplars showing the type of student work at each level. Table 1.1 gives definitions and examples of the components of a system of performance standards.

Table 1.1: Performance standards system components

Term	Definition	Examples
Performance Levels	Labels for the levels of student performance in a content area that convey in a general manner the degree of student achievement in the content area Each performance level encompasses a range of student achievement.	Advanced, Proficient, Partially Proficient Exceeding Standard, At Standard, Approaching Standard, Below Standard
Performance Descriptors	Descriptions of what students at each performance level know and can do that are usually referenced to a specific content area Some descriptions are transdisciplinary: they incorporate knowledge and skills that apply to multiple content areas (e.g., reasoning, predicting, using research skills).	NAEP Grade 4 Achievement Level Descriptions for Reading *Basic:* Fourth-grade students performing at the Basic level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for 4th graders, they should be able to make relatively obvious connections between the text and their own experiences. For example, when reading **literary text**, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences. When reading **informational text**, Basic-level 4th graders should be able to tell what the selection is generally about or identify the purpose for reading it; provide details to support their understanding; and connect ideas from the text to their background knowledge and experiences. *Proficient:* Fourth-grade students performing at the Proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to 4th grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear. For example, when reading **literary text**, Proficient-level 4th graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect. When reading **informational text**, Proficient-level students should be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts. *Advanced:* Fourth-grade students performing at the Advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to 4th grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought. For example, when reading **literary text**, Advanced-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with the ideas

Table 1.1: Performance standards system components (continued)

Term	Definition	Examples
		suggested by the text. They should be able to identify literary devices such as figurative language. When reading **informational text**, Advanced-level 4th graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.
Exemplars	Examples of student work that illustrate the range of performance in a content area within each performance level	Collections of student work that include samples related to the full range and depth of content standards. These collections can include products such as responses to assessment tasks, classroom work, and results of projects. The set of work should include examples from students representing the entire population and range of performances. The compilation should provide a concrete illustration of what students within each performance level know and can do in the content area, including examples of student work that represents proficient, but not quite advanced, or that is just above partially proficient, and therefore proficient. It is critical that the set of exemplars illustrate the range of acceptable performance for each level.
Cut Scores	Score points on an assessment that separate one level of performance from another	On the 1992 National Assessment of Educational Progress, Fourth-Grade Mathematics, a scale score of 211 was the minimum score for achieving the basic level, thus separating the below basic and basic levels; a score of 248 was the minimum score needed to achieve the proficient level; and a score of 280 was needed to reach the advanced level. These cut scores were developed through a process involving expert judgment of what student performance should be at each performance level.

Chapter 2. The Challenges of the New Title I

The 1994 Improving America's Schools Act (IASA) legislation, designed to serve as a framework for improving education in the United States, reflects both research and good practice, based upon what educators have learned about providing effective education for the nation's diverse student population. Title I of IASA represents a significant change in the way the federal government provides education support for low-income students.

At the state level, IASA/Title I requires a system of content standards, assessments, and performance standards (or a system for approving local standards and assessments) that will challenge *all* students, including students with limited English proficiency (LEP) and students with disabilities. The legislation clearly states that the same standards and assessment systems must apply to all students, not just students served by Title I. By developing and agreeing on content standards describing what students should know and be able to do, school systems, teachers, students, parents, and community and business stakeholders can better direct their efforts to improve student performance.

More specifically, the new Title I legislation requires an agreed-upon designation of content standards describing the knowledge and skills students are expected to acquire and demonstrate as a function of schooling, assessments aligned to the content standards, and performance standards that describe how well students and schools are doing relative to the agreed-upon content. This handbook provides tools for "do-it-yourself" alignment of content standards, curricula, instruction, assessments, and performance standards, all components of a comprehensive system reflecting good practice.

The goal of Title I has always been to provide special instructional support for low-income children. In the past, it was designed to be a separate program within a school. Students identified to be served by Title I were often pulled out of the regular classroom for special instruction. Achievement of these students was typically measured using standardized tests, with progress determined by improvement in their norm-referenced rankings. Success was measured using the generalized content represented by standardized norm-referenced tests, which might be very different from what students were taught in their Title I or other classes and from the content required of students not served by Title I.

Because the new Title I is grounded in best practice, it has been reconceptualized in several ways. Student progress is no longer to be measured by scores on generic test content. Rather, student progress is to be based on state-developed or state-approved challenging academic content and performance standards, with assessment systems matched to or aligned with those standards. These standards and assessments apply to all students. Although Title I does not directly require *state-level* standards and a statewide assessment program to be created to assess that content, such a plan is indeed a de facto requirement unless states develop a system that allows for the development, review, and approval of *local* standards and assessments. (See the section on Alignment Requirements/Options in Chapter 3 for details.) Title I requires that students served receive instruction on the same content and curriculum required of all students. The content and performance standards, therefore, must be the same for all students. Except for the few children with severe cognitive disabilities, assessments must also target the same content and performance standards for all students.

Because content standards, performance standards, and the assessment system must be the same for all students, the entire student population must be considered when the standards and assessment system are designed. Educators and others with expertise in educating students with special needs, such as students with disabilities, English language learners, and highly mobile students, must be included in the development process. Although standards and the assessment system will be constant for all students, flexibility must be built in to ensure that they are appropriate for all students.

Flexibility in the assessment system might include a measure of English language proficiency determines whether an English language learner should take an assessment in the standard ıner in English, a version translated into his or her native language, or a form in English with

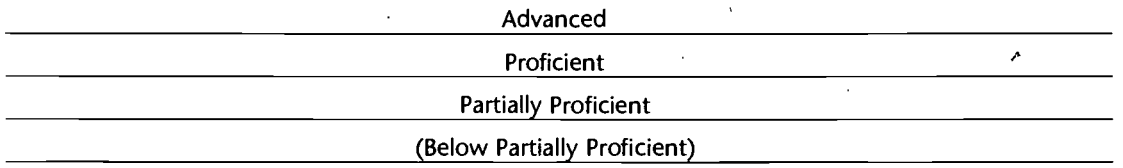
some accommodations, such as a glossary or an oral administration. For students with disabilities, individualized education programs might include specific strategies for delivering content or descriptions of performance that would demonstrate proficiency on the content standards. These descriptions would not be different in rigor or intent from the general performance standards, but they would specify alternative forms of evidence of proficiency for a particular child. The 1997 reauthorization of the Individuals with Disabilities Education Act contains requirements for curriculum, instruction, and assessment for students with disabilities, with an emphasis on applying the same, systemwide rigorous content and performance standards to the education of students with IEPs.

Although the issues involved in creating an education system appropriate for all children are challenging, states such as Kentucky have made progress in generating solutions. Organizations such as the National Center on Educational Outcomes and the Council of Chief State School Officers are also aiding states in accommodating special populations of students as they work to meet the same set of content and performance standards. The U.S. Department of Education is supporting peer consultant networks so that staff in states further along in developing and implementing a system that includes all students can assist less experienced states.

In addition to requiring that each state have a single system of content standards that apply to all students, Title I requires that student success is to be measured using the same system of assessments. The assessment system used must be fully aligned with the content standards. Title I also requires that students be assessed using more than one measure and that results be disaggregated. Disaggregation of assessment results is required so that specific groups of students (e.g., migrant, LEP, or African-American students) are not masked or overlooked and thus ignored (Jaeger & Tucker, 1998).

The legislation further requires states to create performance standards that include at least three levels of performance. The categories of performance, or levels, used in the law are advanced, proficient, and partially proficient, although states are not required to use these specific labels. States may add other levels below the partially proficient level or add levels between the other two mandated levels. Figure 2.1 displays this information graphically.

Figure 2.1: Performance levels



Just as the content standards and assessment system must be the same for all students, the achievement of all Title I schools and students in these schools must be reported based on performance standards. These standards must be the same as those applied to all other students in their system or state. Since the focus is on schoolwide improvement, not just on individual students' gain scores (the focus in the past), improvement cannot be based on only certain categories of students who typically show continuous growth (e.g., only growth of the top-level students). Schools must show adequate yearly progress (AYP) (i.e., improvement in educating all students in all performance categories so that there are fewer students who are below proficient and more students who are proficient or advanced [Carlson, 1996]).

In essence, Title I represents a criterion-referenced point of view. This means that students are no longer required to show gain scores based on a distribution of scores on a norm-referenced test (NRT) but are required to show progress toward certain criteria or standards of performance regardless of how other students score. Rather than any improvement of scores on an NRT defining "success" (notwithstanding the relationship of the NRT to the goals of the program), success

is now defined as improvement in proficiency as defined by the content and performance standards. Under such a standards-based system, all students can be successful. Since success is no longer determined by how well students rank in a group (statistical "norming"), but by what students know and how well they can apply or use knowledge and skills, all students can be successful relative to performance standards.

Figure 2.2 contains excerpts related to standards and assessment from the reauthorization of the Elementary and Secondary Education Act under the IASA Title I, 1994.

Figure 2.2: Title I basic program requirements for standards and assessments

Elementary and Secondary Education Act (ESEA) [Public Law 103-382]

Part A, Subpart 1 — Basic Program Requirements, Section 1111. State Plans

(b) (1) Challenging Standards.— Each State plan shall demonstrate that the State has developed or adopted challenging content standards and challenging student performance standards that will be used by the State, its local educational agencies, and its schools...including at least mathematics and reading or language arts....[S]tandards shall include the same knowledge, skills, and levels of performance expected of all children....

Standards...shall include

Challenging content standards in academic subjects that

- Specify what children are expected to know and be able to do;
- Contain coherent and rigorous content; and
- Encourage the teaching of advanced skills;

Challenging student performance standards that

- Are aligned with the State's content standards;
- Describe two levels of high performance, proficient and advanced, that determine how well children are mastering the material in the state ; and
- Describe a third level of performance, partially proficient, to provide complete information about the progress of the lower performing children toward achieving to the proficient and advanced levels of performance.

(b) (3) Assessments.— Each state plan shall demonstrate that the State has developed or adopted a set of high-quality, yearly student assessments, including assessment in at least mathematics and reading or language arts, that will be used as the primary means of determining the yearly performance of each local educational agency and school served under this part in enabling all children...to meet the State's student performance standards.

Such assessments shall —

- Be the same assessments used to measure the performance of all children...;
- Be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards;
- Be used for purposes for which such assessments are valid and reliable...;
- Measure the proficiency of students in the academic subjects for which a State has adopted challenging student performance standards and be administered at some time during Grades 3 through 5; Grades 6 through 9; and Grades 10 through 12;
- Involve multiple up-to-date measures of student performance, including measures that assess higher order thinking skills and understanding;
- Provide for the participation in such assessments of all students; the reasonable adaptations and accommodations for students with diverse learning needs, necessary to measure the achievement of such students relative to the State content standards; and the inclusion of limited English proficient students who shall be assessed, to the extent practicable, in the language and form most likely to yield reliable information on what students know and can do...in subjects other than English.

Summary

The basic Title I requirements for standards and assessments can be summarized as follows:

- Each state must develop and/or adopt high academic content standards as well as performance standards. These were to be in place by the beginning of the 1997–1998 academic year, but waivers of the time line for developing content and performance standards have been granted to a number of states that are progressing toward meeting this requirement but have not yet met it.
- Each state must determine what constitutes sufficient or adequate yearly progress on the part of schools and school systems toward meeting the state's performance standards.
- Each state must provide and/or develop an assessment system that (1) measures achievement for all students, including those with special learning needs, and (2) provides for accountability of school success as measured by AYP.
- Each state must provide assessments that measure complex skills and challenging subject matter, assess students in at least reading/language arts and mathematics (and more areas, if desired), are administered in at least one grade level in each of the required grade spans (3–5, 6–9, 10–12), and yield reliable and accurate data and data interpretations.

Clear pictures of acceptable standards and assessment plans are still emerging. The challenges to states in designing and creating programs of standards and assessments to meet the new Title I rules and regulations have initiated many long and heated debates, some of which are not yet over. The debates do not focus as much on the dissection of Title I as they do on how to develop and achieve comprehensive and aligned assessment systems that reflect the best of what we know about educating children in the United States. The idea that all students can learn has become paramount, and new ways of educating our children are emerging.

References

Carlson, D. (1996). *Adequate Yearly Progress Provisions of Title I of the Improving America's Schools Act: Issues and Strategies*. Washington, DC: Council of Chief State School Officers.

Jaeger, R. M., & Tucker, C. (1998). *Analyzing, Disaggregating, Reporting, and Interpreting Students' Achievement Test Results: A Guide to Practice for Title I and Beyond*. Washington, DC: Council of Chief State School Officers.

Chapter 3. Standards in an Aligned Assessment System

The expectation that all students can learn challenging content is a concept now guiding education in the United States. All students *can* learn, and each teacher must be able to teach and respond to each student in his or her classroom. It is not enough to teach only those who learn easily; the challenge is to teach those for whom instruction must be varied in design and intensity. Students should be expected to achieve high content standards (what students are expected to know and be able to do) and performance standards (the descriptions of "how good is good enough"), but they must be allowed to achieve those standards in different ways and over different periods of time. The standards must not vary for different groups of students; nor should the expectation that every student can learn to at least the proficient level. Different students will vary in the instructional programs they need, their adaptations to those programs, their need for additional learning support, and the pace at which they learn. Except in extreme cases, variation should not occur in the content standards or performance standards that students are expected to achieve.

Title I reinforces this perspective by building on good instructional practices and targeting three grade spans for assessment; assessments must be conducted at some time during grades 3–5, grades 6—9, and grades 10–12. During each of these three grade ranges, students must be assessed on how well they are meeting their state-approved content standards. How well students meet the content standards is defined by a system of performance standards.

The integration of content standards and performance standards is the cornerstone of the Title I initiative. Defining the content to be learned during each stage of a student's education facilitates the effective use of assessments to measure student progress. Student and school progress is to be measured and reported according to performance standards. For example, a fourth grade student will be expected to have learned all mathematics defined by the mathematics content standards through Grade 4. How well the student has learned that content will be determined by the level of performance he or she demonstrates on assessments that have been created or selected specifically to measure that content. Whether the student is partially proficient, proficient, advanced, or below partially proficient will be determined by an assessment system directly linked or aligned to each content area and its associated performance standards. The assessment system may include multiple-choice tests, constructed-response tasks, writing samples or other performance measures, check lists, and portfolios as well as other forms of assessments. (See Appendix 3.1 for a continuum of these assessment alternatives.)

Why Standards?

Standards address the needs of most groups of stakeholders. For students, the expectations outlined in content and performance standards provide a framework for understanding (1) what they need to know and be able to do to meet the requirements for each performance level and (2) what is expected to enable them to move from one level to the next. The standards can promote challenging, equitable, and rewarding learning experiences for all students by describing what is expected in understandable terms. Students who understand what is expected are more likely to feel ownership of their own progress toward meeting the standards.

For teachers, content standards provide a broad framework to help them focus on the curriculum and what is most important for students to learn. Performance standards and assessments provide them with feedback about how well their students are progressing toward meeting the content standards. When teachers use content standards, performance standards, and assessments that are aligned in a single comprehensive system, instruction becomes more powerful than in a learning context where only details of curricula are addressed, assessments are generic, and there are no stated goals or standards against which to measure student progress.

Through collaborative efforts that include parents, business people, and community members, standards communicate shared visions for learning and provide a common language for

talking about the process of learning and teaching. Standards allow these stakeholders to understand student progress in intelligible, stakeholder-friendly terms.

For schools and districts, standards provide a way to understand learning and instruction. A standards-based approach to education provides descriptions of what students should know and be able to do, and how well they know the information or are able to handle the tasks. Performance descriptors refer directly to the terms used in the content standards and are generally narrative rather than numeric. Being "proficient" is more educationally meaningful and understandable than being "ranked at the 87th percentile," for example, because proficiency can be tied directly to content standards via performance descriptors.

For states, standards are a common reference tool for ensuring that the components of the education system are working together. Standards provide a vehicle for making learning comparable across districts and from school to school. Content standards engender vital conversations about what students should be learning and how performance standards can provide the tools for determining how much learning should take place and is taking place.

One way to answer the question, Why standards? is to think about content and performance standards in real terms. Those who travel by airplane probably want to be assured that the pilot is fully proficient in the art of piloting airplanes. Most passengers really don't care how long it took the pilot to reach the proficient level on the prescribed content standards for airplane flying, only that the standards, both content and performance, were achieved. The same applies to a person who repairs cars or re-roofs houses. Proficiency is highly desirable.

Consider an example from education. A student is in school in a particular class or grade level for a certain number of weeks. At the end of that fixed period of time, whatever content has been learned and whatever grade has been assigned are pretty much fixed. Most students move on after that specified amount of time regardless of success or failure and regardless of how prepared they are to succeed in the next class or grade. Time is the fixed, or specified, variable rather than content, and it is the acceptable levels of performance that are allowed to vary, often dramatically, among students.

Content and performance standards can be used to address this problem and can reverse what varies and what shouldn't vary in education today. When content and standards of acceptable performance remain fixed, and time is allowed to vary, students will be prepared for success throughout their educational careers.

Definitions

To create and establish standards, then, it is essential to understand how performance standards relate to content standards and how they both fit into a comprehensive assessment system. The definitions and examples that follow are intended to provide a foundation from which to communicate clearly and efficiently and to serve as a springboard for generating ideas. Although the focus of this handbook is on performance standards, content standards must come first, since they define the foundation for learning and instruction. For that reason, a brief section on content standards is included in the discussion of performance standards.

Content standards. Content standards answer the question, What should students know and be able to do? They are descriptions of the knowledge and skills expected of students at certain times throughout their education, often targeted at a specific grade level or at a cluster of grade levels. Content standards are clear, broad statements of content in academic areas such as mathematics, language arts, science, and social studies. They are few in number and are defined more specifically by curriculum, instruction, and examples of student work. They cite important ideas, concepts, issues, and other relevant information and skills to be taught and learned. They also include the specialized skills — ways of thinking, working, reasoning, investigating, and communicating — within each content domain. Such skills are sometimes referred to as "habits of the mind" or "cross-cutting competencies."

Content standards should

- be specific enough to provide a vision of expectations relative to a curriculum (e.g., the student can apply lessons learned from and make extensions of a text and evaluate texts critically);
- be aligned with performance standards, assessments, principles of learning, curriculum, and instruction;
- be clear and understandable to teachers, students, and parents;
- be assessable in a variety of ways;
- be illustrated by examples of student work; and
- be useful for defining and supporting good instruction

Following is an example of a content standard taken from *Curriculum and Evaluation Standards for School Mathematics* by the National Council of Teachers of Mathematics.

In grades 5–8, the mathematics curriculum should include exploration of statistics in real-world situations so that the students can —

- systematically collect, organize, and describe data;
- construct, read, and interpret tables, charts, and graphs;
- make inferences and convincing arguments that are based on data analysis;
- evaluate arguments based on data analysis;
- develop an appreciation for statistical methods as a powerful means for decision making. (1989, p. 105)

This content standard illustrates the difference between a standard and curriculum. Note that the *standard* says the "curriculum should include exploration of." Remember, standards are broad statements. The standard does not detail the specific content or materials; a curriculum will do both. Indeed, several pieces of curriculum are needed to represent the knowledge and skills in this one standard. An example of curriculum (information specifically written as a "technical manual" for educators) related to the standard might be for the student to explain different ways to use and interpret circle graphs, bar graphs, and histograms. The content standard (the broad view) describes statistics at the general level of understanding, but it is specific enough to determine what to do to teach this information (i.e., which components of the curriculum are relevant to this standard). Figure 3.2 contains a checklist useful in exploring how to formulate content standards.

Appendix 3.2 contains sample content standards for mathematics and language arts. These examples were created by the Metropolitan Atlanta P–16 Community Council for work with its local systems. (P–16 refers to preschool [P] through college or postsecondary [16].) The content standards are intended to be either approved by local systems or used as a guide in developing local standards. Readers may wish to use the standards as guides for developing or refining their own standards.

Figure 3.2: Checklist for content standards*

Content standards are concerned with "big ideas." Standards should contain the major concepts and essential ideas that students must master in order to grasp the content. Being able to understand mathematics by making inferences is different from memorizing formulas.

Content standards are accurate and sound. Standards should reflect the most recent, widely accepted scholarship in the discipline. Because facts and concepts change rapidly today, when new information is constantly being generated, maintaining accuracy and balance among the important concepts requires continual revision. Documents related to content learning should be updated regularly.

Content standards are clear and useful. Standards should be specific enough to drive the curriculum. They should not be written in language so abstract or technical that teachers, parents, and students cannot easily understand them.

Content standards are parsimonious. Standards should reflect the depth of learning. Standards should be few and brief, and short enough to be memorable because they are strong, bold statements, not details of content description (the details are in the curriculum).

Content standards are built by consensus. Standards must be arrived at by most of the constituency who will use them. Conversations about standards are as important as the standards themselves.

Content standards are assessable. Standards should have verbs that indicate an assessable action. Words like "compare," "explain," or "analyze" are useful for assessments. Words like "understand" or "appreciate" are not.

Content standards are for students. Standards should describe to students what they are responsible for knowing. The standards should be clear and understandable to them.

Content standards are developmental. Standards should evidence a clear sense of increased knowledge and sophistication of skills. Standards that simply repeat content and specify "more" at successive levels are not useful. Benchmarks, or target levels for assessment, should indicate developmentally appropriate content knowledge and skills.

Content standards are visionary. Standards should be the goal of student learning. They should not describe "what is" or "this is where we are," but rather, "this is where we want our students to be."

* Adapted from Ruth Mitchell's *Front-End Alignment: Using Standards to Steer Educational Change* (1996, pp. 22–23)

Systems of performance standards.[3.1] As discussed in Chapter 1, this handbook defines performance standards as a system that includes (1) performance levels, labels for levels of achievement; (2) performance descriptors, descriptions of student performance at each level; (3) exemplars, illustrative student work for each level; and (4) cut scores, score points on a variety of assessments that differentiate between performance levels.

A system of performance standards has certain characteristics:

- The system defines several distinct levels of performance.
- Examples of student work from the full population of students are used to articulate each level.
- The system is based on specific content, and components are interpreted relative to content standards.
- Components of the system can be understood by someone who is not an educator.

[3.1] As noted in the preface, these definitions are designed to foster mutual understanding and consistent communication within this document. The definitions provided are not "standard" yet. In some cases, the terms and their precise meanings are still evolving. For additional discussion of performance standards, see Chapter 1.

A common source of confusion is the difference among performance descriptors, scoring rubrics, and cut scores. An example of a performance descriptor for the proficient level in mathematical problem solving follows:

> To be proficient, the student solves mathematical problems in more than one way. The student clearly articulates each step in a process correctly and accurately. The student is able to demonstrate multiple approaches and can make a determination about the effectiveness and feasibility of various possible solutions. The student is able to provide accurate solutions.

The performance descriptor for "proficient" is broad enough to be applied to a variety of tasks and/or assessments and covers multiple content standards. It clearly states in broad terms what the student must be able to do to achieve proficient status in mathematical problem solving.

In contrast, an example of a rubric for scoring a particular problem-solving task follows:

> To be acceptable, a response must identify and explain each step in the series to solve the problem correctly and accurately. At least two methods must be presented; a third is optional. Estimates are not acceptable.

This sample rubric appears to describe a single response and therefore is not generalizable across more than one class of problems. It is also quite specific about what is and is not acceptable and, therefore, appropriate for scoring a piece of student work. However, it lacks the richness needed for use with a variety of products or performances. As such, it is not an acceptable description for a performance level tied to content standards.

Cut scores define a particular point or points on a score scale. These points differentiate between performance levels (e.g., pass/fail or advanced/proficient/partially proficient). A cut score typically does not include information about any other performances beyond those on a specific assessment instrument.

The checklist in Figure 3.3 may serve as a general guideline for developing a system of performance standards.

Figure 3.3: Checklist for a system of performance standards

Performance standards should be understandable and useful for all stakeholders. The system should describe to stakeholders what is expected of students who perform at a given level. If a stakeholder wants to know what it means to be proficient, then the stakeholder should be able to understand the kind of work that is required by reading the descriptor for that level of performance and looking at examples of student work.

Performance standards clearly differentiate among levels. Performance descriptors should be easy to apply to collections of student work. When they apply the descriptors for the performance levels, teachers, parents, and students should clearly see why certain sets of student exemplars or student profiles are assigned to one performance level and not to another.

Performance standards are grounded in student work but not tied to the status quo. The system should reflect the major concepts and accomplishments that are essential for describing each level of performance. Student work that reflects the diverse ways various students demonstrate their achievement should be used to inform the descriptions during various stages of development, illustrating where students should be as a result of the educational process rather than where they are now.

Performance standards are built by consensus. The system of standards must be arrived at by the constituency who will use them. It must be built around agreed-upon statements of a range of achievement with regard to student performances. Not only should teachers and students understand the standards, but the "end users," such as colleges and universities, technical schools, and employers, should also understand what performance standards mean for them.

Performance standards are focused on learning. Performance descriptors should provide a clear sense of increased knowledge and sophistication of skills. Descriptors that simply specify "more advanced" at each successive level are not particularly useful. The "more" should be clearly described or defined to show progression of learning. Cut scores on assessments must be based on this learning, and exemplars of student work should illustrate learning at each level.

Examples of Performance Descriptors

Performance descriptors vary. Each example in this section was selected to show the variety of approaches to development. The first example is from the National Assessment of Educational Progress (NAEP). The NAEP achievement levels reflect one of the earliest efforts to develop performance standards. The NAEP descriptions include both generic policy definitions and content- and grade-specific statements about student performance. The other two examples are from individual states. Maine began its standards development in the early 1990s, and its initial descriptors were general in nature, serving as the foundation for developing performance descriptors for specific content areas. Colorado recently developed detailed performance descriptors tied to content standards from their initial development, rather than creating more general descriptors.

NAEP Achievement Levels and Performance Descriptions

Since 1990, the NAEP results have been reported using achievement levels authorized by Congress. These achievement levels were developed by the National Assessment Governing Board in the major content domains, including reading, writing, mathematics, science, U.S. history, geography and civics. The policy definitions are general and apply across all content domains and grade levels.

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

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.

Advanced: This level signifies superior performance.

For each domain, these general achievement or performance levels have been operationalized with content-related detail to make them more useful and to tie them to the NAEP assessment frameworks. NAEP achievement levels for Grade 4 mathematics (1992–1994) are as follows:

Basic: Fourth-grade students performing at the basic level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content strands.

Fourth graders performing at the basic level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve simple real-world problems in all NAEP content areas. Students at this level should be able to use — though not always accurately — four-function calculators, rulers, and geometric shapes. Their written responses are often minimal and presented without supporting information.

Proficient: Fourth-grade students performing at the proficient level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content strands.

Fourth graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.

Advanced: Fourth-grade students performing at the advanced level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP strands.

Fourth graders performing at the advanced level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should be beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

The NAEP performance level descriptors for Grade 4 Reading are shown in Table 1.1.

State Examples

Although the NAEP achievement levels are grade and content specific, some states have developed performance descriptors that are generalized across grade and content as a starting point for developing more specific descriptors, much like NAEP's policy definitions. Other states have created elaborate and detailed performance descriptors tied to content standards, without reference to general descriptions.

Maine began working with standards in the early 1990s. Its initial performance descriptors, presented below, are general in nature. Colorado recently completed developing its content standards and performance descriptors; the excerpted sections are elaborate and specific. These two examples were selected to illustrate the two ends of the continuum from general to specific. Examples from other states are included in Section II.

Maine

Maine has had integrated performance descriptors in place since 1994, and it has since developed descriptors tied to content standards. Assessments of its content standards include multiple-choice items and performance tasks. Performance levels for students in Maine are defined as novice, basic, advanced, and distinguished. The integrated descriptors, used as the starting point for developing content area descriptors, are shown below.

Novice. Maine students display partial command of essential knowledge and skills. With direction, these students apply their knowledge to complete routine problems and well-defined tasks. The students' communications are rudimentary, and sometimes ineffective.

Basic. Maine students demonstrate a command of essential knowledge and skills with partial success on tasks involving higher-level concepts, including applications of skills. With some direction, these students make connections among ideas and successfully address problems and tasks. Their communications are direct and reasonably effective, but sometimes lack the substance or detail necessary to convey in-depth understanding of concepts.

Advanced. Maine students successfully apply a wealth of knowledge and skills to independently develop new understanding and solutions to problems and tasks. These students are able to make important connections among ideas and communicate effectively what they know and are able to do.

Distinguished. Maine students demonstrate in-depth understanding of information and concepts. The students grasp "big ideas" and readily see connections among ideas beyond the obvious. These students are insightful, can communicate complex ideas effectively (and often creatively) and can solve challenging problems using innovative, efficient strategies.

Colorado

In March 1997, Colorado developed performance descriptors for individual content standards in reading and mathematics. The excerpts that follow illustrate performance descriptors for each content standard at the partially proficient, proficient, and advanced levels for each targeted grade range.

READING

Content Standard 5: Students read to locate, select, and make use of relevant information from a variety of media, reference, and technological sources.

Performance Descriptors for Grades K–4

Partially Proficient. Students inconsistently find and make use of information using organization features of a variety of printed texts and electronic media. Students take notes, outline, and identify main ideas in resource materials, but there may be inaccuracies, limited understanding, omission of important facts and details, or direct copying.

Proficient. Students are able to find and make use of information, using organizational features of a variety of printed texts and electronic media for a specific topic or purpose. Students accurately take notes, outline, and identify main ideas in resource materials and give credit by listing sources.

Advanced. Students can easily and without assistance find information, using organizational features of a variety of printed texts and electronic media. Students sort, record, and synthesize information from a wide variety of sources and give credit by listing sources.

Performance Descriptors for Grades 5–8

Partially Proficient. Students locate and select relevant, but sometimes insufficient, information, using organizational features of printed text and electronic media. Students' use of technology is dependent upon direct teacher assistance. The end product is inaccurately or incompletely documented.

Proficient. Students locate and select relevant information, using organizational features of printed text (for example, prefaces, afterwords, appendices) and electronic media (for example, microfiche headings and numberings, headings for accessing nested information in hypertext media, electronic media, CD-ROM, laser disc). With minimal assistance, students use available technology to research and produce a quality end product that is accurately documented and contains a bibliography.

Advanced. Students locate, select, synthesize, and evaluate relevant information effectively, using the more complex organizational features in a wide variety of printed texts and electronic media. Students are able to efficiently and independently use available technology to research and produce an end product that is effectively presented and that shows unusual depth, is accurately documented and contains a substantial bibliography.

Performance Descriptors for Grades 9–12

Partially Proficient. Students use a narrow range of research strategies to find, record, and evaluate information from a limited number of print and electronic media. Students produce an end product which is insufficiently researched, omits important information, or is not carefully documented.

Proficient. Students independently use research strategies to find, record, and evaluate information from a number of different print sources (journals, research studies, technical documents, footnotes, endnotes, bibliographies) and electronic media (bulletin boards, keyword searches, e-mail) to produce a carefully documented, quality product.

Advanced. Students independently use extensive research strategies to find, record, and evaluate information from a variety of print sources (journals, research studies, technical documents) and electronic media (bulletin boards, keyword searches, e-mail) to produce a carefully documented product demonstrating depth and insight.

MATHEMATICS

Content Standard 6: Students link concepts and procedures as they develop and use computational techniques including estimation, mental arithmetic, paper-and-pencil, calculators, and computers in problem-solving situations, and communicate the reasoning used in solving these problems.

Performance Descriptors for Grades K–4

Partially Proficient. Students demonstrate limited conceptual meanings for the four basic arithmetic operations of addition, subtraction, multiplication, and division. With some errors, students can add and subtract commonly used fractions and decimals using physical models. Students demonstrate some understanding of and limited proficiency with basic addition, subtraction, multiplication, and division facts without the use of a calculator. Students demonstrate limited ability to construct, use, and explain procedures to compute and estimate with whole numbers. With limited success, students select and use appropriate methods for computing with whole numbers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods.

Proficient. Students demonstrate conceptual meanings for the four basic arithmetic operations of addition, subtraction, multiplication, and division. Students add and subtract commonly used fractions and decimals using physical models. Students demonstrate understanding of and proficiency with basic addition, subtraction, multiplication, and division facts without the use of a calculator. Students construct, use, and explain procedures to compute and estimate with whole numbers. Students select and use appropriate methods for computing with whole numbers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods.

Advanced. Students demonstrate a comprehensive understanding of the conceptual meanings for the four basic arithmetic operations of addition, subtraction, multiplication, and division. Students add and subtract commonly used fractions and decimals using physical models and justify the procedure used. Students demonstrate thorough understanding of and proficiency with basic addition, subtraction, multiplication, and division facts without the use of a calculator. Students construct, use, and thoroughly explain procedures to compute and estimate with whole numbers. Students show insight in selecting and using appropriate methods or combinations of methods for computing with whole numbers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods.

Performance Descriptors for Grades 5–8

Partially Proficient. Students use models inaccurately when explaining how ratios, proportions, and percents can be used to solve real-world problems. With some procedural errors, students construct, use, and give an incomplete explanation of procedures when computing and estimating with whole numbers, fractions, decimals, and integers. Students develop, apply, and explain a limited number of estimation strategies in problem-solving situations, or provide an incomplete explanation of why an estimate may be acceptable in place of an accurate answer. With limited success, students select and use methods for computing with commonly used fractions and decimals, percents, and integers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and make an attempt to determine whether the results are reasonable.

Proficient. Students use models when explaining how ratios, proportions, and percents can be used to solve real-world problems. Students construct, use, and explain procedures when

computing and estimating with whole numbers, fractions, decimals, and integers. Students develop, apply, and explain a variety of different estimation strategies in problem-solving situations, and explain why an estimate may be acceptable in place of an exact answer. Students select and use appropriate methods for computing with commonly used fractions and decimals, percents, and integers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and determine whether the results are reasonable.

Advanced. Students use models when explaining how ratios, proportions, and percents can be used to solve a variety of real-world problems. Students construct, use, and explain multiple procedures when computing and estimating with whole numbers, fractions, decimals, and integers. Students develop, apply, and explain a wide variety of different estimation strategies in problem-solving situations, defend the estimation strategy chosen, and explain why an estimation may be acceptable in place of an exact answer. Students select and justify the use of appropriate methods for computing with commonly used fractions and decimals, percents, and integers in problem-solving situations from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and determine whether the results are reasonable.

Performance Descriptors for Grades 9–12

Partially Proficient. With some procedural errors, students use ratios, proportions, and percents in problem-solving situations. Students use a limited range of methods for computing with real numbers, selecting from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and they make an attempt to determine whether the results are reasonable. Students describe some limitations of estimation, and incompletely assess the amount of error resulting from estimations.

Proficient. In problem-solving situations, students use ratios, proportions, and percents. Students select and use appropriate methods for computing with real numbers, selecting from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and determine whether the results are reasonable. Students describe the limitations of estimation, and assess the amount of error resulting from estimation within acceptable limits.

Advanced. In problem-solving situations, students use ratios, proportions, and percents. Students select and use appropriate methods for computing with real numbers, selecting from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods, and provide insightful arguments that the results are reasonable. Students thoroughly describe the limitations of estimation, and assess the amount of error resulting from estimation within acceptable limits.

What Is Alignment?

Alignment is a match between two or more things. *Webster's New World College Dictionary* defines *align* as "to bring into a straight line; to bring parts or components into a proper coordination; to bring into agreement, close cooperation." In an aligned system of standards and assessments, all components are coordinated so that the system works toward a single goal: educating students to reach high academic standards.

Content standards, performance standards, and assessments must be aligned so that what is taught is tested and what is tested is taught. No surprises, no questions, no controversy, and no confusion. Although the primary use of a system of performance standards may appear to be its connection with the tests or assessments as results are reported, the system remains rooted in content. Performance descriptors must be carefully written to reflect the content for which students are to be held accountable and must be used to guide the development and selection of test items. Without performance descriptors drafted prior to the development of an assessment sys-

tem, tests may be created or selected that do not align with the content and performance standards in terms of depth, or complexity, or fully reflect the breadth, or coverage, of the standards.

So, how does one achieve alignment? Begin with content, what students should know and be able to do. **Content standards** should be written for each content domain (e.g., reading, mathematics). Content standards provide the overview of content so that parents, students, and other stakeholders can understand what it means to educate our country's children. One way to develop content standards is to begin with considering what students should know and be able to do at specific points in their education. For example, by the end of fourth grade, students should be able to use reading for learning, as a research tool, and as a means of communication. In mathematics, students should understand how to use numbers and number theory to solve problems and use computation and estimation to communicate solutions to real-world problems. These broad end-of-grade descriptions can be further refined by indicating just what fourth-grade students should know. These "benchmarks," or points of reference, become the content standards.

A **curriculum** is developed for each set of the broad content standards. Teachers use a curriculum to determine what to teach their students to help them meet the content standards. Think of a curriculum as a bridge, or conduit, between the broader vision of what is important in lay terms and what teachers should teach in their classrooms. The curriculum is simply an elaborated or "technical" version of the content standards. Content standards and curricula are related tools; they do not contain different content to be learned, and they are not in conflict. The sets of content standards are the models, and the curricula are the blueprints for building those models. If they are created in this way, they automatically align.

Systems of **performance standards** and **assessments** must be created or selected and matched with the content. In an aligned system, all content standards must be accounted for in some manner. A common misconception is that when prebuilt or off-the-shelf assessments appear fairly well matched to content standards and curricula, there will also be alignment. Many of the test items or tasks, usually around 80 to 90 percent, or even 100 percent, may match what is specified in a set of content standards. What is missing is the rest of the matching process, which, it turns out, is critical. That is, how much of the content standards and curriculum is actually assessed using the prebuilt system? What the test assesses may reflect only 50 to 60 percent (or even less) of the content standards and curriculum the test is selected to measure. What about the parts of the content standards/curriculum that are not assessed at all with the prebuilt system or even a custom-built system?

The critical questions are (1) Is the relative emphasis of content covered by the test the same as the intended emphasis in the content standards? and (2) Is the range of difficulty of the test consistent with the range of performance levels, as defined by the performance descriptors? The answers to these questions will help determine which tests and other assessment techniques comprise the entire comprehensive assessment system. If one part of the system, such as an on-demand test, measures only some of the content standards, another component of the system, such as classroom assessment embedded in instruction, might measure other content standards.

Real alignment means that the assessment system matches the depth (difficulty and complexity) and breadth (content and coverage) of the content nearly one-to-one. Spending time developing visionary content standards, elaborate curricula, and carefully crafted performance descriptors and then using an assessment system that assesses only a portion of the content is unfair. It's unfair to teachers and students and their parents. It's unfair to future employers. In a fully aligned system where content and assessments match completely, the question of whether a teacher should teach the curriculum or teach to the content of the assessments becomes irrelevant. Since the system is aligned, it shouldn't make any difference. If teachers do ask whether they should teach to the test or teach to the curriculum, then the assessments and content standards may not be completely aligned.

Educators are very good at matching curriculum with series of books and other materials that are used in classrooms. They do it all the time. The matching process is the same for aligning content standards, curriculum, and assessments. Each curriculum statement must be tagged to one or more content standards. Each content standard or curriculum statement must be tagged to an assessment activity or activities. Some activities will be part of a formal assessment instrument

or test; other content will be assessed in informal ways. In either case, the match between content and assessment must be complete (i.e., all of the content standards and curriculum must be accounted for and matched to assessments). The driving force is not that all the test items are accounted for, but the reverse. *All content must be accounted for in creating an aligned assess*ment system.

The discussion of aligning content and assessments illustrates how these two components must work together. The system of performance standards operates at several levels and is the "how good is good enough" part of a comprehensive aligned system. During the process of developing content standards and designing an assessment system, performance descriptors must be created. To become operational, content standards need performance descriptors and examples of student work to give them life.

Performance descriptors articulate the various levels of learning. Words such as *Basic, Proficient,* or *Accomplished* commonly label performance levels. The terms themselves are not important except as hooks on which to hang descriptions and illustrations of student performance.

A system of performance standards can be used to determine AYP (i.e., whether schools and local districts are making progress in educating students to reach the standards). The goal is to constantly and consistently decrease the number of students at lower performance levels and to increase the number of students at higher performance levels.

Remember, performance standards apply to all students, not just part of the school population. They include students living in poverty, LEP students, migrant students, and special education students. *Everyone* is assessed based on the same content and performance standards.

Alignment Requirements/Options

Title I has three major options for states to follow in creating an aligned system of standards and assessments:

- establish a single statewide system of standards and assessments;
- establish a statewide system with provisions for the addition of local standards and/or assessments, given state approval;
- establish state models for or criteria for approval of local system standards and/or assessments, to ensure high quality and comparability statewide.

An explanation of each option follows.

Option 1. States may establish content and performance standards and assessments that apply uniformly to all schools and local school districts or systems in the state. This option implies that every student is expected to meet the same content and performance standards, no matter where he or she attends school within the state.

Option 2. States may elect to combine state content standards, performance standards, and assessments with local district standards and/or assessments. This option maximizes a district's opportunity to be flexible and/or tailor standards for its own use. However, care must be taken in establishing and documenting an approval process. Under this option, the state must establish content and performance standards and assessments. Local districts must adopt those standards, but they are allowed to develop additional standards that meet or exceed the state standards. In such a program, content and performance standards are aligned across the state by the "core" standards, with each district having the option to enhance the state's established content and performance standards. The core standards are in no way less rigorous or excellent than content and performance standards established under Option 1 above. By allowing local districts some authority to expand the standards and assessments, the issue of local control and ownership can be addressed without jeopardizing the requirement of rigorous content and performance standards for all students. Under such a plan, the state must carefully articulate the linking process

for approval of district programs and, according to the law, must facilitate alignment between the content standards, performance standards, and assessments for all local districts.

Option 3. States may elect to allow local systems to establish their own content and performance standards and assessments instead of adopting a uniform system across the state. In this case, the state must establish model standards and/or criteria for review of local standards to ensure comparability, rigor and conformity with the state model or criteria. The state must also ensure that all school districts set standards that meet the state's criteria and/or model standards in terms of challenging content and rigor. As an example, districts with high concentrations of poor or low-achieving students must establish standards and assessments that are no less demanding than a district with more affluent or higher-achieving students. The standards must be sufficiently high in both cases to raise performance and enhance learning of all students.

Under Options 2 and 3, states must be extremely careful about the various sets of standards. The state must ensure that all sets of content standards meet the range of desired content coverage, the depth of content coverage, and the degree of emphasis on topics or areas within the content. For performance standards, the degree of rigor for each performance level must be ensured across the accepted and approved content standards.

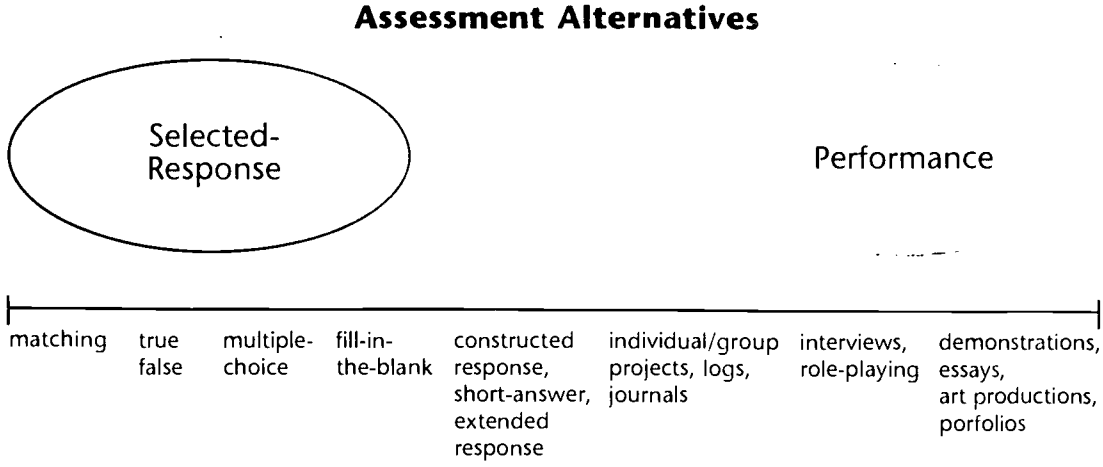
The following chapter describes ways to connect content standards, performance standards, and assessments as part of a comprehensive system. Regardless of which of the above options is selected for the Title I plan, a comprehensive assessment system must be clearly articulated and put into place. All system components must be "cut from the same cloth" and boldly and seamlessly pieced together.

References

Mitchell, R. (1996). *Front-End Alignment: Using Standards to Steer Educational Change.* Washington, DC: The Education Trust.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Appendix 3.1. Continuum of Assessment Alternatives



Appendix 3.2. Metropolitan Atlanta P–16 Community Council Language Arts and Mathematics Standards

Language Arts Academic Standards, May 1997

Introduction

Draft academic standards are being developed by persons from member organizations which are affiliated with the Metropolitan Atlanta P–16 Community Council. Conversations among K–12 and post-secondary faculty within core academic disciplines help to minimize gaps in expectations between high school graduation and post-secondary admission. Benchmarks at levels 3, 5, 8, 12, and 14 are intended to suggest the appropriate level of accomplishment for students at that juncture. The setting of academic standards is the prerogative of local school boards for K–12 or the governing bodies of the particular post-secondary institutions. Our desire is to provide individual systems a set of draft standards which they may wish to affirm.

The voluntary model standards are intended to reflect a continuous learning process. They are based on content which is cumulative at each benchmark within each discipline. They also were written based on the newly revised state-mandated Quality Core Curriculum. Progress toward each benchmark may vary in terms of time according to the intellectual development of each student. The benchmarks are meant to insure an acceptable level of learning so that student success is enhanced. By preparing properly at each preceding benchmark, students will be responsible for work for which they are developmentally and academically prepared.

Cross-cutting competencies are also included in the voluntary academic standards. Cross-cutting competencies refer to the knowledge and skills required during any learning process, regardless of academic content. Cross-cutting competencies also provide a foundation for preparation for various career paths, workplace know-how, and solid job performance. Examples include conducting research, evaluating data, allocating resources, using technology, and communicating effectively across diverse cultures. These competencies are woven within each discipline.

Our goal is to support the various stakeholders in the Metro Atlanta P–16 community and to provide an education based on high academic standards in which all students can be successful. As part of this effort, performance standards are also being developed. The performance standards will be framed as model assessments which can be used to describe and determine student progress matched to the voluntary academic standards as well as all learning across the curriculum.

The Metropolitan Atlanta P–16 Community Council is housed at Georgia State University. For additional information and updates on the Council, please consult our web site at http://www.gsu.edu/MetroP-16, or call 1-800-822-8515.

Language Arts Vision Statement

The vision guiding these standards is rooted in the fact that all students are capable of learning. They must be provided with the resources to develop the language skills they need to be lifelong learners and to participate fully as informed, productive members of society. Literacy growth begins before children enter school as they experience and experiment with literacy activities — reading and writing, and associating spoken words with their graphic representations. These standards encourage the continued development of the emerging literacy abilities that children bring to school as well as furthering the development of the necessary tools essential for all learning — comprehending the written and spoken word. The standards provide ample room for innovation and creativity essential to teaching and learning. The language arts — reading, writing, listening, and speaking — form the basis for all other learning. They are not separate or distinct; they are, by nature of the process they represent, interrelated and must be considered as a whole as students move through their educational career from preschool through college.

Language Arts Academic Standards

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 1: Students read a wide range of print and nonprint materials to build an understanding of text, of themselves, and of the cultures of the world in order to acquire and create new information, to respond to the needs of society and the workplace, and for personal fulfillment.				
Read with accuracy, fluency, and comprehension using materials of appropriate difficulty (picture books, simple narratives, directions) Use basic technology to explore text	Read with accuracy, fluency, and comprehension using materials of appropriate difficulty (longer works of fiction and nonfiction) Use basic technology to explore text	Read with accuracy, fluency, and comprehension using materials of appropriate difficulty (various themes and genres, reference works) Use basic technology to extract meaning from texts	Analyze and synthesize texts of appropriate difficulty (various themes and genres, technical and research documents) Use basic technology for a variety of purposes	Recognize and respond appropriately to rhetorical situations Use technologies for a variety of academic and professional purposes Understand the purposes of genres of professional discourse (memoranda, proposals, reports, agenda, minutes)
Standard 2: Students employ a variety of writing strategies and different writing elements to communicate with different audiences for various purposes.				
Write as a way of exploring ideas Plan and write in each of the four discourses (telling, describing, convincing, explaining) Use different technologies for communicating with and learning about others and for creating original works Participate in prewriting, drafting, revising, editing, and publishing	Write as a way of creating and clarifying ideas Plan and write for the four discourses (telling, describing, convincing, explaining) by demonstrating competency in the five domains (content and organization, style, sentence formation, grammar/usage, mechanics) Use different technologies for communicating with and learning about others and for creating original works Participate in prewriting, drafting, revising, editing, and publishing	Write as a way of discovering and clarifying ideas Plan and write multi-paragraph compositions (narrative, persuasive, descriptive, expository) by demonstrating competency in the five domains (content and organization, style, sentence formation, grammar/usage, mechanics) Use different technologies for communicating with and learning about others and for creating original works Participate in prewriting, drafting, revising, editing, and publishing	Write as a way of analyzing ideas Plan and write multi-paragraph compositions (narrative, persuasive, descriptive, expository) by demonstrating competency in the five domains (content and organization, style, sentence formation, grammar/usage, mechanics) Write problem/solution papers based on multiple perspectives and research Use different technologies for communicating with and learning about others and for creating original works Participate in prewriting, drafting, revising, editing, and publishing	Write as a way of evaluating and synthesizing ideas Use technical writing to provide information for journals, research studies, technical manuals and technical documents Formulate and defend theses (extended documents) Distinguish the difference between arguable and unarguable issues and ideas Use different technologies for communicating with and learning about others and for creating original works Participate in prewriting, drafting, revising, editing, and publishing

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 3: Students employ a variety of strategies (e.g. focusing, managing barriers, paraphrasing, formulating appropriate responses) to become effective listeners.				
Practice listening skills to follow directions, answer questions, and gather basic information Listen actively to gather information and respond appropriately	Practice listening skills to follow directions , answer questions, and gather basic information Listen actively to gather information and respond appropriately	Listen actively and critically and respond to varied communications to determine meaning Listen actively by paraphrasing to demonstrate understanding of other points of view Empathize with other people(s) so as to know what they said and understand what they meant	Listen critically and respond to informal and formal situations and develop points of view for diverse issues, audiences, and ideas Comprehend, interpret, and evaluate literal and implied meaning in a variety of listening situations, debates, dramatic presentations and readings from literature and poetry Analyze and question spoken language in order to become an informed citizen capable of participating in a democratic society Empathize with other people(s) so as to know what they said and understand what they meant	Comprehend, interpret and evaluate literal and implied meaning in a variety of listening situations, including lectures, speeches, debates, dramatic presentations, literature Analyze and question spoken language in order to contribute to a democratic society Empathize with other people(s) so as to know what they said and understand what they meant
Standard 4: Students use spoken language to communicate effectively and appropriately in a variety of situations and for a variety of audiences for different purposes.				
Demonstrate oral language skills of pace, volume, pronunciation, and appropriate word choice Participate in oral presentations, both group and individual, such as drama activities, choral reading, readers' theater, storytelling	Demonstrate oral language skills of pace, volume, emphasis, pronunciation, and appropriate word choice Participate in oral presentations using notes, visual aids, and technology, such as drama activities, choral reading, readers' theater, storytelling, debating	Demonstrate verbal and non-verbal language skills of pace, volume, emphasis, articulation, appropriate word choice, eye contact, gestures Participate in oral presentations using notes, visual aids and technology, such as drama activities, choral reading, readers' theater, storytelling, debating	Demonstrate verbal and non-verbal language skills of pace, volume, emphasis, articulation, appropriate word choice, eye contact, gestures Deliver extemporaneous and planned presentations and speeches	Become a competent speaker in a variety of settings including personal, political and social situations in classrooms and workplaces using a variety of technologies
Standard 5: Students use conventions of standard English for oral and written communication.				
Begin to apply language conventions to subject/verb agreement, punctuation, capitalization, sentence types, and spelling	Begin to apply language conventions of subject/verb agreement, modifiers, punctuation, abbreviation, capitalization, sentence types, spelling, pronoun agreement, possessives, prefixes and suffixes Explore the use of technology (editing software) to write, revise, and edit	Apply skills and use appropriate terminology for language conventions of subject/verb agreement, modifiers, punctuation, abbreviation, capitalization, sentence types, spelling, pronoun agreement, possessives, affixes and roots, essay formats Explore the use of technology (editing software) to write, revise, and edit	Evaluate one's own and peers' written works using language conventions of pronoun usage, modifying phrases and clauses, parallel structure, internal capitalization, secondary quotations, research documentation, manuscript forms specified in various style manuals Explore the use of technology (editing software) to write, revise, and edit	Recognize the metacognitive power of language conventions (the difference between knowing and doing)

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 6: Students apply critical reading and thinking skills to spoken, written, and visual language.				
Identify relationships Make predictions Draw conclusions Determine main ideas Identify prior knowledge	Draw conclusions Make judgments Interpret graphs and charts Determine relationships Determine main ideas with supporting details Identify prior knowledge	Determine relevancy Make predictions and judgments Determine cause and effect Identify propaganda techniques Determine main ideas with supporting details Make inferences Identify prior knowledge	Compare, defend, criticize, judge and appraise Synthesize through formulation of a theory Prepare an alternative Use persuasive/propaganda techniques Manipulate analogous relationships Identify prior knowledge	Compare, defend, criticize, judge and appraise Synthesize through formulation of a theory Prepare an alternative Use persuasive/propaganda techniques Manipulate analogous relationships Identify prior knowledge
Standard 7: Students conduct research by locating, selecting and making use of relevant information from a wide variety of sources.				
Begin to use basic reference materials	Use reference materials, e.g., encyclopedias, periodicals Find information in more than one source Identify and limit topics Locate evidence to support topics	Use reference materials, e.g., encyclopedias, periodicals, Internet Find information in more than one source Identify and limit topics Locate evidence to support topics Gather information (outline, paraphrase) and write a research report	Write a research paper Formulate a thesis Document sources Provide evidence in a coherent form Synthesize ideas Write precisely Use various manuscript styles (e.g., MLA, APA, Turabian)	Conduct research for a variety of purposes and audiences Synthesize data Write precisely Use various manuscript styles (e.g., MLA, APA, Turabian)

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 8: Students use study techniques and test taking strategies to comprehend, retain, and organize information.				
Use a variety of informational sources for study purposes (pictures, books, dictionaries, technology) Begin to use study techniques	Use a variety of informational sources for study purposes (e.g., thesaurus, almanac, encyclopedias, technology) Begin to use study techniques (SQ3R [survey, question, read, recite, review], PQRST [preview, question, ready, study, test], KWL charts, anticipation guides) Use test preparation and test taking strategies essential for a variety of tests Develop a personal time management system Identify individual learning styles Develop outlines for comprehension and study purposes	Use a variety of informational sources for study purposes (e.g., thesaurus, almanac, encyclopedias, technology) Begin to use study techniques (SQ3R, PQRST, KWL charts, anticipation guides) Use test preparation and test taking strategies essential for a variety of tests Develop a personal time management system (study schedules, priority lists) Identify individual learning styles (modalities, metacognition) and study accordingly Develop outlines for comprehension and study purposes	Use a variety of informational sources for study purposes (e.g., thesaurus, almanac, encyclopedias, technology) Use study techniques (SQ3R, PQRST, anticipation guides) to guide and enhance learning Use test preparation and test taking strategies essential for a variety of tests (post-secondary entrance exams, career option exams) Develop a personal time management system (study schedules, priority lists) Identify individual learning styles (modalities, metacognition) Develop outlines for comprehension and study purposes	Use test preparation and test taking strategies essential for a variety of tests (professional exams, GRE, LSAT, comprehensives) Maintain a time management system and study schedule
Standard 9: Demonstrate knowledge of literary genres representative of diverse cultures, time periods, and ideas.				
Read, respond to and discuss literature as a way to explore the similarities and differences in story structure Recognize a variety of literature, including picture books, folk tales, etc. Use literary elements such as setting, plot, character, problem and solution	Read, respond to and discuss novels, poetry, short stories, folk tales, legends, non-fiction, plays, and content area and technical material Distinguish the elements that characterize and define literary form and structure Use literary terminology such as setting, character, conflict, plot, resolution, theme, foreshadowing and figurative language	Read, respond to and discuss novels, poetry, short stories, folk tales, legends, non-fiction, plays, and content area and technical material Distinguish the elements that characterize and define literary form and structure Use literary terminology such as setting, character, conflict, plot, resolution, theme, foreshadowing and figurative language	Read a literary text analytically, understanding relationship between form and content Use literary terminology accurately, such as mood, diction, style, point of view Identify recurrent themes in literature	Differentiate among genres such as allegory, epic, masque Demonstrate knowledge of various literary periods and an understanding of how literature reflects the habits and practices of the people who lived at the time a specific text was written Demonstrate an understanding of how cultures affect interpretive practices

Mathematics Academic Standards, May 1997

Introduction

Draft academic standards are being developed by persons from member organizations that are affiliated with the Metropolitan Atlanta P–16 Community Council. Conversations among K–12 and post-secondary faculty within core academic disciplines help to minimize gaps in expectations between high school graduation and post-secondary admission. Benchmarks at levels 3, 5, 8, 12, and 14 are intended to suggest the appropriate level of accomplishment for students at that juncture. The setting of academic standards is the prerogative of local school boards for K–12 or the governing bodies of the particular post-secondary institutions. Our desire is to provide individual systems a set of draft standards which they may wish to affirm.

The voluntary model standards are intended to reflect a continuous learning process. They are based on content which is cumulative at each benchmark within each discipline. They also were written based on the newly revised state-mandated Quality Core Curriculum. Progress toward each benchmark may vary in terms of time according to the intellectual development of each student. The benchmarks are meant to insure an acceptable level of learning so that student success is enhanced. By preparing properly at each preceding benchmark, students will be responsible for work for which they are developmentally and academically prepared.

Cross-cutting competencies are also included in the voluntary academic standards. Cross-cutting competencies refer to the knowledge and skills required during any learning process, regardless of academic content. Cross-cutting competencies also provide a foundation for preparation for various career paths, workplace know-how, and solid job performance. Examples include conducting research, evaluating data, allocating resources, using technology, and communicating effectively across diverse cultures. These competencies are woven within each discipline.

Our goal is to support the various stakeholders in the Metro Atlanta P–16 community and to provide an education based on high academic standards in which all students can be successful. As part of this effort, performance standards are also being developed. The performance standards will be framed as model assessments which can be used to describe and determine student progress matched to the voluntary academic standards as well as all learning across the curriculum.

The Metropolitan Atlanta P–16 Community Council is housed at Georgia State University. For additional information and updates on the Council, please consult our web site at http://www.gsu.edu/MetroP-16, or call 1-800-822-8515.

Mathematics Vision Statement

- Use mathematics to explore, describe and model real world situations.
- Use a variety of tools, technologies, and techniques to solve problems, to reason critically, to think logically, and to conduct mathematical research.
- Use problem solving approaches to investigate and understand mathematics.
- Learn to explore meaningful mathematical situations that require self-confidence and persistence.
- Communicate mathematics effectively through speaking, listening, reading, writing, and spatial visualization.

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 1: Develop number sense, use number relationships in problem solving situations and be able to communicate the reasoning involved.				
Construct number meanings through real-world experiences and the use of physical materials Understand our numeration system by relating counting, grouping and place value concepts Interpret the multiple uses of numbers Develop meaning for the four basic operations Relate the mathematical language and symbolism of operations to problem situations and informal language Identify fractions using physical models, both as parts of a whole and parts of a set Develop concepts of fractions and decimals with standard symbols Apply fractions and decimals to problem situations, including money	Construct number meanings through real-world experience and the use of physical materials Understand our numeration system by relating counting, grouping and place value concepts Investigate whether numbers are odd or even, prime or composite Interpret the multiple uses of numbers encountered in the real world Develop meaning for the basic operations by modeling and discussing a rich variety of problem situations Compare numbers to each other in terms of greater than, less than, or equal and explore different representations of the same number Relate the mathematical language and symbolism of operations to problem situations and informal language Develop concepts of fractions, mixed numbers and decimals Apply understanding of whole number operations to fractions and decimals Apply fractions and decimals to problem situations, including money	Understand, represent, and use numbers in a variety of equivalent forms (integer, fraction, decimal, percent, exponential and scientific notation) in real-world and mathematical problem situations Develop the real number system Understand and apply ratios, proportions, and percents to a wide variety of situations Investigate relationships among fractions, decimals, and percents Understand and appreciate the need for numbers beyond the whole numbers Develop, represent, and use order relations for whole numbers, fractions and decimals (rational numbers) and integers Apply the basic operations to integers Understand how the basic arithmetic operations are related to one another Develop and apply number theory concepts (e.g., primes, factors, and multiples) in real-world and mathematical problem situations	Develop and use fractional exponents, negative exponents, radicals and complex numbers Apply operations for the real number systems to a variety of mathematical situations	Perform arithmetic operations with real and complex numbers Estimate, and judge reasonableness of numerical results Use percentages, orders of magnitude, ratios, and proportions to express relationships between quantities

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 2: Apply geometric concepts, properties and relations and communicate the reasoning used in the application.				
Describe, sort, and classify shapes Investigate the results by subdividing, combining and changing shapes Construct two- and three-dimensional shapes with physical models Identify and draw two- and three-dimensional shapes Develop spatial sense (include near, between, etc.) Recognize geometric relationships	Describe, model, draw, and classify shapes Investigate and predict the results of combining, subdividing and changing shapes Identify, describe, and draw lines, line segments, lines of symmetry, rays, angles and parallel and perpendicular lines Identify and draw three-dimensional shapes Relate geometric ideas to number and measurement ideas Determine when figures are congruent and similar Recognize geometric relationships in the world	Identify, describe, compare, and classify geometric figures Visualize and represent geometric figures with special attention to developing spatial sense Explore transformations of geometric figures Determine when figures are congruent and similar Represent and solve problems using geometric models Understand and apply geometric properties and relationships Recognize geometric relationships in the world Investigate properties of triangles and develop connections among right triangle ratios	Interpret and draw two- and three-dimensional objects Represent problem situations with geometric models and apply properties of figures Classify figures in terms of congruence and similarity and apply these relationships Deduce properties of, and relationships between, figures from given assumptions and using transformations Compare, contrast and translate across synthetic and coordinate geometry Deduce properties of figures using transformations	Synthesize geometric concepts into algebraic, functional, and problem solving activities
Standard 3: Use probability and statistical models to analyze data and make inferences about real-world situations.				
Formulate and solve problems that involve collecting, organizing and analyzing data Explore concept of fairness, uncertainty and chance	Collect, organize, and describe data Construct, read, and interpret displays of data, including picture, bar, circle, and line graphs Formulate and solve problems that involve collecting, organizing and analyzing data Determine probability of an event Explore concepts of fairness, uncertainty and chance	Collect, organize and describe data Construct, read and interpret tables, charts and graphs Apply measures of central tendency Make inferences and convincing arguments and evaluate arguments that are based on data analysis Recognize statistical methods and probability models as powerful decision making tools Devise and conduct experiments or simulations to determine probabilities Construct a sample space to determine the theoretical probabilities Make predictions that are based on experimental or theoretical probabilities	Construct and draw inferences from charts, tables and graphs Use curve fitting to predict from data Understand and apply measures of central tendency, variability, and correlation Use experimental or theoretical probability, as appropriate, to represent and solve problems involving uncertainty	Use and analyze probability models Use descriptive and inferential statistics

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 4: Understand algebraic concepts and use algebraic methods to explore, model, and describe situations that can be represented symbolically.				
Relate mathematical symbols to mathematical ideas Relate subtraction to addition Recognize, extend, and create patterns Explore the use of variables and open sentences to express relations	Extend the number system to fractions and decimals Relate algebraic ideas to geometric representation Relate multiplication to addition, arrays, and Cartesian products Relate division to subtraction and multiplication Explore the use of variables and open sentences to express relations Use models to represent mathematical ideas Represent situations and number patterns with tables, graphs and verbal rules	Understand the concepts of variable, expression and equation Formalize situations and number patterns with tables, graphs, verbal rule, and equations and explore the interrelationships of these representations Analyze tables and graphs to identify properties and relationships Solve linear equations using concrete, informal and formal methods Investigate inequalities and nonlinear equations informally Apply algebraic methods to solve a variety of mathematical problems	Represent situations that involve variable quantities with expressions, equations, inequalities and matrices Use tables and graphs as tools to interpret expressions, equations and inequalities Operate on expressions and matrices, solve equations and inequalities, and use matrices to solve linear systems Recognize the power of mathematical abstraction and symbolism	Represent mathematical situations symbolically Use a combination of algebraic, graphical, and numerical methods to solve problems Use a variety of algebraic structures including polynomials, rational expressions, absolute value, exponents, logarithms, matrices, and the algebra of functions Explore the deductive nature of mathematics–recognize the role of definitions, axioms, theorems, and proofs.
Standard 5: Use discrete mathematical algorithms and combinatorial concepts to solve problems.				
Represent data in tables and graphs	Represent data in tables and graphs	Represent data in tables and graphs Exhibit relationships graphically	Use Venn diagrams and truth tables in problem solving involving set theory and logic Use inductive and deductive reasoning Use counting principles Use proof by induction	Synthesize the use of algorithmic and combinatorial techniques into problem solving situations

P–3 Benchmarks	4–5 Benchmarks	6–8 Benchmarks	9–12 Benchmarks	13–16 Benchmarks
Standard 6: Express functions using various representations and incorporate the concept of function in broad areas of mathematics				
Relate numbers to points on a line Understand betweenness, closeness, rounding and approximating	Understand closeness, rounding and approximating Recognize and describe mathematical relations and functions Explore concepts of operational inverses	Expand the number system Use relations and functions to solve problems Recognize and describe relations and functions	Determine the maximum and minimum points of a graph and interpret the results in problem situations Investigate limiting processes by examining infinite sequences and series and areas under curves Apply functions to problem-solving situations Solving trigonometric equations Develop and use functional operations (inverse and composition) Apply functional operations to problem solving situations including trigonometric, exponential and logarithmic functions Understand the connection between trigonometric and circular functions Apply general graphing techniques to trigonometric functions	Interpret functional relationships between two or more variables Translate functional information from one form to another and effectively use these representations Investigate properties of functions

Chapter 4. Development and Alignment of a System of Performance Standards

Developing performance standards is challenging work but not without rewards. Because the process involves members of the education community and the greater community as well, conversations about developing standards can have as much value as the products themselves.

Creating an effective system of performance standards requires developing adequate, accurate descriptions that can be applied to students' classroom work as well as other, more formal assessments used for decision making by administrators and policy makers. Because performance standards describe student learning, they must (1) be general enough to apply to the range of student achievement within a single category (e.g., achievement demonstrated by students who are just proficient and by those who are close to the advanced level), and (2) be specific enough to clearly distinguish among the levels of performance.

This chapter is focused on the development of performance standards aligned with content standards and assessments at the state and district levels, because these are the levels on which Improving America's Schools Act (IASA)/Title I assessment and accountability requirements are focused.

There is no best way to create an aligned system. However, until assessments are in place and student work and data are available for validating those assessments, neither the content nor the performance standards should be considered "final." It is only as the three facets — content standards, performance standards, and assessments — interlock that a fully aligned system can be achieved (see Figure 1.1). They must form the strong connections between the remaining links of curriculum, instruction, and program evaluation. Table 4.1 describes these linkages.

Although presented in a linear fashion, the development of content standards, performance standards, and assessments is anything but linear. The development process and products are iterative in nature; all three must be tightly interwoven and aligned with each other.

Ensuring Alignment for All Students

Just as educators are responsible for reaching all students, and not just those who learn easily, it is important that educators take responsibility for ensuring that content standards, performance standards, and assessment systems are aligned for all students. This responsibility includes making sure that performance descriptors are inclusive enough to fully embrace the varied ways students demonstrate what they know and can do. An aligned system must be appropriate for students who are English language learners, those who have disabilities, and those who have learning, processing, and/or response styles that might be different from most other students' styles.

Alignment for *all* students, then, means that when formal and informal assessments are matched to standards, the critical question becomes whether the tests and other evaluation methods selected are measuring the *same content* (depth, breadth, and relative emphasis) for all groups making up the diverse population of learners. The extent to which the matches are different for some students needs to be identified, and the assessment system must be supplemented to make sure that all content is accounted for, with the same breadth, depth, and relative emphasis, in evaluating what students know and can do.

Ensuring this alignment requires active involvement at all stages of the process by educators familiar with each type of student. It also requires an evaluation of the development, scoring, and reporting elements of the formal and informal assessments as well as an evaluation of the performance standards. Several states are working with the Council of Chief State School Officers on how to supplement systems and accommodate diverse learners.

Table 4.1: Integrated links in an aligned comprehensive assessment system

Links	Questions	Integration
Content standards	What does a student need to know to be successful at each grade level? What does a student need to know to be successful at each performance level (e.g., basic, proficient)?	Statements of learning goals at predetermined levels throughout a student's educational career Success tied to clear, predetermined expectations of performance Content appropriate for the complete population of students
Curricula and pedagogy	What should be taught to meet content standards? What are the implications for instructional content and teaching techniques?	Fully articulated descriptions of specific content to be taught to support the content standards
Performance Standards	What does good performance look like? How good is good enough?	Performance descriptors at multiple levels (at least three) for each content domain, appropriate to describe performances of all students Profiles/exemplars of student work at each performance level based on assessments measuring content-specific standards
Assessments	What should be assessed? How should it be assessed? How are assessment results tied to the content and performance standards?	System based on multiple measures to determine how well content standards are implemented and how much students are learning System appropriate for full range of learners, with any variations in assessment instruments evaluated in terms of link to content and performance standards System designed to accommodate the diverse student population in full
Instruction	How are content and performance standards reflected in teaching and learning? How are assessment results used to inform instruction and define instructional strategies? How is instruction advanced through the use of assessment results?	Development of strategies for improving instruction based on assessment results Professional development for all educators in assessment literacy Determination of critical content teachers must teach so that instruction can be accomplished successfully
Reporting	What is important to know about student, school, system, and state performance? How well do reports align with content and performance standards?	Documents delineating progress toward achieving content and performance standards that are useful for informing instructional decisions
Evaluation	How well is the student, the school, the system, or the state doing? How much improvement is needed and should be expected? Where should resources and efforts be targeted to improve student learning?	Feedback that informs program development and improvement Determination of adequate yearly progress Identification of "best practice" and student success Feedback that targets improvement in the areas of curriculum and assessment, aligned with content and performance standards

Development of Systems of Performance Standards

In order to develop a system of performance standards, a process incorporating multiple activities is necessary. The process itself is iterative in nature: developers must work back and forth from one activity to another to achieve a final product that can withstand the rigors of constant and close investigation from any and all stakeholders. Although the process is presented in linear fashion, the activities must be integrated, and some may take place concurrently. Though the general components of the process are standard, each development effort will be unique.

The activities that follow are based on the assumption that strong content standards already exist and that content-based assessments will be administered for reporting the performance levels of individual students and/or schools. There are no answers provided to the questions that follow, because every situation and state is different, and users will have to provide answers that fit in their particular contexts. The questions are posed as reminders about the important issues and concerns that will most likely need consideration as content standards, performance standards, and assessments evolve. The information that follows each question represents only a partial selection of possible options.

Although "finalization" is an attractive concept, meaningful, useful standards and assessments are never really final. The nature of learning and teaching is such that, as changes occur in what students need to know and be able to do, the process of "maintaining" standards must include modifying them. However, once adopted, standards need to be consistent to allow sufficient time for appropriate evaluation and decision making.

Create a process by determining the foundation for developing performance standards.

- How will performance standards be used in your state/district?
  - to exemplify content standards
  - to inform the public about expectations

- Which Title I–appropriate option will be pursued?
  - state only
  - state/local
  - local approved by state

- Who will be involved in development?
  - committees of educators
  - members of professional organizations
  - community forums
  - policy makers

- How will participants be involved?
  - writers
  - consultants
  - reviewers
  - final authorities

- Will the same participants be involved in various stages of development?
  - planning
  - developing
  - reviewing and revising
  - steering and oversight
  - implementation procedures

- What resources are needed?
  - personnel
  - time
  - money
  - materials

- Who will review and approve the standards?
  - educators
  - parents
  - business leaders
  - policy makers

- Who will adopt the standards?
  - state board of education
  - local boards of education
  - legislative bodies

- How will the standards be disseminated to the public?
  - print media
  - audiovisual media
  - news media
  - Internet/electronic delivery
  - school/system publication

Draft performance levels and performance descriptions,[4.1] incorporating student work.

- How many performance levels will be developed?
  - number of performance levels
  - labels for performance levels

- What is the starting point?
  - adopt state and national standards
  - develop standards "from scratch"
  - modify existing standards

- Who will be involved in this development?
  - educators
  - parents
  - content experts
  - experts on students with diverse learning needs
  - legislators
  - business partners

- What processes will be involved in writing the descriptors?
  - collect copies of the best examples available
  - converse with experts
  - review content standards
  - review assessments
  - review existing performance descriptors from other sources
  - describe levels of expectation (not status quo, but what is expected at each performance level)

[4.1]Refer to Chapter 9, contributed by Craig Mills and Richard Jaeger, for information that should be helpful for this work.

- How will initial sets of student exemplars be selected?
  - work with teachers who work with diverse student populations
  - collect samples of student work, including work from English language learners and students with disabilities, that cover the full range of performance at each level
  - reference existing works in the field of standards development (e.g., National Council of Teachers of Mathematics, other states)
  - select work representative of local districts and schools
  - create and administer sample or draft assessments from which to select exemplars

- Who needs to review the draft descriptors?
  - teachers
  - administrators
  - parents
  - students

Administer assessments based on content standards and draft performance descriptors by developing or adopting an assessment system.

- What type of assessment system will be developed/adopted?
  - plan a system that will cover all of the content described in the content standards
  - plan a system that will cover the full range of both expected and desired performances

- What types of assessments will be used?
  - determine content needs in terms of depth and breadth
  - select assessment methods that match content requirements:
    - selected response
    - performance based
    - formal
    - informal
    - combination

- What existing tools can be used to assess student performance in relation to the content standards?
  - inventory current assessments
  - identify areas of content not covered or not adequately covered

- How will additional assessments be created?
  - original
  - off-the-shelf
  - custom built

- How will you know if students are meeting the content standards?
  - examine assessment scores
  - check alignment of content standards
  - check alignment of performance descriptors
  - use assessments that elicit performance at all levels
  - develop descriptors that adequately capture performance all levels
  - determine opportunity to learn and perform

Revise performance descriptors and exemplars based on assessment results.

- Who will have opportunity for revision input?
  - teachers
  - administrators
  - students
  - parents
  - community
  - steering committees
  - members of other committees

- What will revision decisions be based on?
  - assessment results
  - content/assessment match
  - whether descriptors are inclusive of all students
  - where students are versus where they should be
  - descriptors that are clearly distinguishable from each other
  - clear descriptions for each level of performance
  - clear descriptions for "borderline" students

- How will student exemplars be used?
  - as anchors, or exemplars, for each performance level
  - for each content domain
  - for each performance standard
  - to represent multiple ways to meet a standard
  - to provide stable reference points for consistency at each level of performance

Set cut scores on assessments by selecting a process and finalize procedures.[4.2]

- Which processes and/or methods will be used to define the levels of performance described in the standards?
  - determine an appropriate method for setting cut scores
  - select a process for setting cut scores
  - convene panels to recommend cut scores
  - determine the final sets of performance descriptors to use during the process

- Who will be panel members?
  - teachers
  - administrators
  - content experts
  - experts in special education, or second language or bilingual education
  - parents
  - members of the business community
  - legislators

- Who will approve the process for setting cut scores?
  - state department of education personnel
  - assessment experts
  - local boards of education
  - state board of education

[4.2]See Chapter 10, contributed by Ronald Hambleton, for specific information on how to accomplish this work.

- Who will adopt the panel results?
  - state board of education
  - local boards of education
  - legislative panels

Report results by publishing standards and assessment results.

- How will content and performance standards be shared within schools and across the district and state?
  - as impact on instruction
  - as action plans for teachers and administrators for program improvement
  - as promotional guides

- How will results be communicated to students and parents?
  - reports on what is working to foster increased student success
  - action plans created for community leaders and parents, focusing on what it will take to improve student learning

- How will report cards and other records reflect the standards?
  - create formats for reporting results reflecting standards information
  - create alternative formats for different audiences (parents, school boards)
  - verify accuracy of visually presented information

- How can assessment results be used to review and refine both the content and performance standards?
  - examples of student work based on actual assessments to inform revisions
  - assessment results that match content standards and allow reporting at all levels of performance

- How can results be used to inform instruction and to provide a variety of strategies for improving student learning?
  - maximize the alignment between teaching, learning, and instruction
  - determine classroom, school, district, or state level needs to improve student performance
  - evaluate improvement plans
  - improve understanding of the standards and assessments

Remember that this list is only partial, both in the specific questions asked and in the issues considered under each question. As each program is planned, designed, and implemented, unique needs will arise. The information is provided to serve only as a guide in development efforts.

Of concern to all educators is the need to ensure that performance standards and assessments are accessible to all students. Deserving special emphasis, accessibility is the responsibility of state and local education agencies. Although test developers are involved to the extent that they have contracted to design a test for states or districts, or even if off-the-shelf tests are used as part of an assessment system, the ultimate responsibility for accommodating all students lies with the state or district. It is critical that the system of standards and the assessments are both valid and reliable for all students. State and district educators need to evaluate the content and performance standards and the assessment system. The goal must be to maximize accessibility, so that the same content and performance standards are aligned to the assessment components, and the same performance standards are being applied to all students.

To accomplish this work, content and performance standards must be clearly understood by

everyone in the education community; they must be rigorous, applicable to all students, and used in the design of curriculum and assessments. In developing this handbook, it became painfully clear that there is little reported research or even informal reports available that explain how to write performance descriptors. Although it seems relatively clear who should be involved in developing performance descriptors, it remains unclear exactly how the descriptors for novice or partially proficient levels, for instance, get written down once the proper committees with all the appropriate people are convened. A growing number of states and local education agencies (LEAs) have worked through this process, but it doesn't seem to be formally documented. Section II of this handbook contains the best available documentation to date.

A Closing Encouragement and a Continuing Challenge

Although states are at various stages of effort, ranging from rudimentary awareness of need to several years of experience, they are clearly intent upon rising to a critical national challenge — the challenge of placing standards, assessment, and accountability in the service of learning in ways that are comprehensive, thoughtful, effective, and defensible. Doing so places enormous demands on human resources; demands collaboration among policy makers, educators, and the citizens they serve; and begs for the identification or development of prototypes, models, and strategies that provide optimal learning opportunities for all students while meeting the needs of particular populations and jurisdictions.

Drawing upon the experience of best existing practice and empirical evidence of effectiveness is and will continue to be critical to achieving the goal — a goal that must maintain integrity beyond rhetoric and serve as an omnipresent beacon to the nation's position as a world leader. Thanks go to those who are leading the way, to those who are cutting new paths, to those who are willing to share their stories of successes as well as their stories of lessons learned, and to all of us who are working to meet the challenge of serving our children and our nation through responsible accountability practices.

Section II:
State Stories — A Reality Check

Compiled and edited by Doris Redfield

In recent years, states have taken various approaches to meet ever-increasing demands for higher standards of student performance in academic content and skill areas. Current efforts are particularly focused on comprehensive systems of assessment that consider *content standards* (statements of what students should know and be able to do), *performance standards* (a system that describes and illustrates what kinds of performances represent learning at various levels), *curriculum* and *instruction* designed to effectively deliver the knowledge and skills necessary for student learning and performance relative to content standards, *assessments* designed to determine the extent to which students have mastered the content and skills represented by the standards, and *accountability* indices that show how well schools, school districts, states, and other entities are demonstrating desirable levels of student achievement.

Factors influencing the current focus include recent federal legislation such as the Improving America's Schools Act (IASA), Title I, and Goals 2000 as well as state reform initiatives, grassroots movements, and the pressures of a changing world economy. States and other jurisdictions, such as local schools and communities, along with their governing bodies, are searching for efficient and effective ways to identify, develop, implement, and evaluate systems that can account for student learning in meaningful and useful ways relative to particular sets of standards.

At least two things are becoming increasingly clear:

- The ultimate goal for all concerned is optimal levels of achievement for all students, including students with special needs.
- There is no universally acceptable or desirable method to achieve this goal.

Nonetheless, some states have braved uncharted territory and taken approaches that may serve as models for others in the midst of seeking their way through murky change processes. Approaches have been varied, as the following examples will indicate:

- standards targeted at every grade level (i.e., K, 1, 2, 3, etc.) such as Virginia's Standards of Learning versus those targeted at categories of levels (e.g., primary, elementary, secondary) such as Kentucky's Valued Student Outcomes;
- standards that are content specific versus standards that cut across academic content areas;
- standards linked to high-stakes accountability (e.g., personnel evaluation, school take-overs, high school graduation) versus those applied on a voluntary basis and designed for program improvement decisions only such as North Dakota's;
- highly specific standards that can be used to operationalize actual test items versus intentionally general frameworks for use in guiding curriculum development.

In the end, each enterprise must decide upon an approach that best meets its needs in ways that are feasible and defensible. The following stories offer documentation of various approaches to the challenge. These stories will be updated, and other stories will be added as they evolve.

Special thanks go to those who worked hard to provide and organize the information related in the stories that follow:

- Don Watson, Colorado Department of Education
- Jessie Pollack, Maryland Department of Education
- Mike Dalton and Rosemary Fitton, Oregon Department of Education
- Alan Sheinker and Jan Sheinker, Wyoming Department of Education

Chapter 5. The Colorado Standards-Based Assessment System

In 1993, the Colorado legislature passed HB 93-1313, which required the implementation of standards-based education. Meeting the demands of the legislation meant developing standards in reading, writing, mathematics, science, history, and geography. Assessments within these content areas were to be established at particular grade levels and standards were, therefore, benchmarked to those levels (grades 4, 8, and 11).

Content Standards

In the fall of 1995, the State Board of Education adopted the required academic content standards. These standards were developed by a council, named the Standards Assessment Development and Implementation (SADI) Council, in conjunction with staff from the Colorado Department of Education (CDE). The SADI and CDE staff constituted the group officially sanctioned by the legislature. SADI members were appointed by the governor and approved by the state senate. The SADI process involved convening task forces to develop standards in each of the designated content areas.

Performance Levels and Descriptions

In the spring of 1996, and in compliance with HB 93-1313, the CDE convened a group of community representatives (e.g., business leaders) and educators, including teachers, school administrators, and university faculty with expertise in the designated content areas. This group was charged with drafting performance level descriptions that were related to the content standards upon which students would be tested.

The State Board of Education approved the draft performance levels in the spring of 1996. The approved standards were operational from the spring of 1996 to the fall of 1997, even though some were still in draft form. The goal was to develop performance standards (i.e., cut scores) to accompany the approved content standards and draft performance level descriptions following administration of the new, content standards–based assessments.

Assessments

A contract to develop assessments in fourth-grade reading, writing, and geography was awarded in November 1996 to CTB/McGraw-Hill. Development began in December 1996. Although teachers did not actually write any of the test items, they did impact item development (e.g., by reviewing items for bias).

In February 1997, the Colorado State Legislature could not agree on whether the new assessment program should be funded as a separate line item or taken out of flow-through moneys to local school districts. Members of the Legislature questioned the $1.75 million bid to develop the new system, because they believed that it would be less costly. Hence, the program was put on hold.

After approximately two weeks, the legislature decided to proceed by funding reading and writing at the fourth-grade level at a cost of $1.6 million. The savings of $150,000 was realized by omitting the administration and scoring of the geography assessment. The decision was influenced by the fact that the two other assessments were well under development.

Then, HB 97-1249 reconfigured the entire assessment program. The previous model was based upon using the first two years of assessments at any particular grade level to establish baselines. The first year of program implementation would have begun with fourth-grade reading, writing, and geography as described previously. Baseline establishment at Grade 4 was to continue the program's second year with the addition of mathematics, science, and history. In the third year, localities would include every student in the assessment process and the state would

use a sampling procedure to allow for district level comparisons statewide. The intent was to compare state standards and assessments with local standards and assessments to assure that local programs matched or exceeded the state program. In the spring of 1997, assessments in reading and writing were administered to every fourth-grade student. The tests were scored by a commercial vendor (CTB/McGraw-Hill) rather than locally.

Finally, the model specified that similar procedures should be initiated in the third year of the program in Grade 8 and in the fifth year of the program in Grade 11. HB 97-1249 overturned this model. Instead, this legislation, in conformance with the governor's desire, determined to assess every student in each designated content area as follows:

- Spring 1998: - reading literacy in Grade 3 only, with continuation of the fourth-grade reading and writing assessments set in place by the previous model
- Spring 1999 - add Grade 5 mathematics
- Spring 2000 - add Grade 8 mathematics
- Spring 2001 - add Grade 8 science

In February 1998, a new accreditation bill will be presented during the state legislative session and is anticipated to pass. This bill proposes leaving the spring of 1998 plan intact (i.e., reading literacy assessments in Grade 3; reading and writing assessments in Grade 4).

Prior legislation called for all testing to occur in the spring. However, the proposed legislation would move the fifth-grade mathematics assessment to the fall of 1999, with the addition of seventh-grade reading and writing in the spring of 1999. In the spring of 2000, eighth-grade mathematics and science assessments would be added, with Grade 10 reading, writing, and mathematics to be added in the spring of 2001.

There is a current push to administer the fifth-grade mathematics assessment at the same time as the other assessments proposed for fifth-grade administration so that instruction is not interrupted for standardized testing more than once during the school year. There does not appear to be any significant opposition to this proposal on the part of legislative leaders, and the governor has expressed support.

Program Description

Descriptions of proposed grade levels and content areas to be assessed are chronicled above. In terms of format, approximately 60 percent of the assessment items are multiple-choice and approximately 40 percent call for constructed responses or performances. In other words, not all of the assessments are machine scorable.

For purposes of staff development, consideration is being given to the idea of holding sessions in which teachers score non–multiple-choice items using student papers from previous test administrations.

Performance Standards (Cut Scores) Setting Process

In September 1997, Colorado convened a group of 40 teachers, 20 with reading expertise and 20 with writing expertise. These teachers were nominated by the SADI councils and local superintendents. In cases in which councils were unable to nominate a suitable number of candidates, local school administrators were asked to provide recommendations. Those selected were demographically representative and consisted of teachers who were widely recognized for their expertise in the designated content areas at Grade 4.

The selected teachers were asked to participate in a standard-setting process developed by CTB/McGraw-Hill called "bookmarking." Briefly, bookmarking involves ranking items from least to most difficult after analyzing student responses using an item response theory process.

To set cut scores using the results of the statistical analysis, teachers are given a "book" with test items ranked from easiest to most difficult based on student performance. Teachers then mark the book where they believe the partially proficient, proficient, and advanced sections

begin, based on the descriptions of the performance levels. CTB/McGraw-Hill then works with the teachers, through three rounds, to resolve any discrepancies between the teachers' conclusions and results of the statistical analyses. After the last round, the median ranking for each level is selected (partially proficient, proficient, and advanced).

Colorado's currently designated proficiency levels of partially proficient, proficient, and advanced were established and described in draft form by the State Board of Education. Once the convened group of teachers previously described had completed the bookmarking process, their recommendations were clarified by CDE and CTB staff and submitted to the SADI Council. The council made some minor adjustments and then submitted the proposed performance level definitions to the board.

The performance standards, as submitted by SADI, were approved by the board. There are now four performance levels: unsatisfactory, partially proficient, proficient, and advanced.

Approach to Title I/IDEA Requirements

Accommodations are at two levels:

- those requiring documentation (i.e., terms and procedures requiring explanation for lay understanding — such as extended time) and
- those not requiring particular explanation (e.g., communication device).

Accommodations in Colorado are not limited to students with individualized education programs (IEPs). However, in order to qualify for an accommodation, the student must have had a relevant instructional accommodation for a three-month period prior to testing. In effect, most such students have IEPs or 504 plans, and there are a few additional Title I students.

A caveat is that students may be given an extra 10 minutes per testing session without documentation. The assessments are scheduled for a total of six 50-minute sessions.

A task force is currently working on alternate assessment issues. Although accommodations allow for adaptations of existing assessment procedures, alternate assessments require different mechanisms for students who cannot be accommodated in ways that will yield valid results (e.g., severely physically handicapped).

The thinking relative to the assessment of limited English proficient (LEP) students was that only Spanish language versions of the new assessments would be developed since approximately 90 percent of Colorado's LEP students are speakers of Spanish. The number of such students, statewide, is about 2,000.

Given the time line (December to mid-April), development of the Spanish language assessments were initiated in geography, since that test had already been developed and would require only translation. However, this plan was discarded, and a CTB "shelf-ready" test, the SUPERA, was used in fourth-grade language arts (reading, writing) in the spring of 1997.

It is now planned that Spanish language versions of the fourth-grade assessments in reading and the third-grade assessments in reading and writing will be developed this year (1998) with hopes to administer them in the spring. However, there is current discussion under way to abandon the effort; it has been argued that such students should be able to demonstrate their level of learning using English rather than Spanish.

In mid-February 1998, the commissioner of education, with state board approval, issued a memorandum to local school districts indicating that the Spanish language assessments would be temporarily discontinued for financial reasons related to the costs of scoring. Upon receipt of the memorandum, a significant number of local school districts expressed their disappointment. In response to this outcry, administering the tests in Spanish was made a local option, since the Spanish versions had already been shipped to the schools. It is not yet known how many localities will choose to exercise the option, but a number of them provide instruction in Spanish, particularly at the primary level.

Conclusion

Policy, and the acceptance of its changing landscape, are critical to the ability of state and local educators to develop and implement successfully standards-based assessment programs that truly serve teaching and learning. An important lesson learned from the rapid-fire changes in Colorado's recent policy making is that if elegant, precise assessment models do not meet public needs and expectations, they will not be successful. It is important to combine the best of both interests to benefit student learning and performance opportunities.

Chapter 6. Establishing Proficiency Levels, Proficiency Level Descriptions, and State Standards for the Maryland School Performance Assessment Program

Background: Schools for Success and the Maryland School Performance Program

Schools for Success is a comprehensive school reform effort focusing on improving schools through rigorous academic and performance standards. This initiative was shaped by the report of the Governor's Commission on School Performance (August 1989). The Commission's work dramatically altered the state's perspective on accountability. The responsibility for student success shifted from individual students to schools and school systems.

The cornerstone of Schools for Success is the Maryland School Performance Program, a comprehensive system involving three levels of commitment beyond the state: local school system, school, and student. The aim of the program is to address school improvement through system and school accountability. The program honors individual student achievement and accountability through its commitment to the belief that all children can learn, and it holds that such learning is contingent upon school and school system accountability. That is, all children have the right to attend schools in which they can progress and learn equally rigorous content.

Maryland School Performance Assessment Program: An Overview

The Maryland School Performance Assessment Program (MSPAP) is the assessment component of Maryland's Schools for Success reform initiative known as the Maryland School Performance Program. Assessment results are used for accountability at the school and school system levels. Therefore, an understanding of the key features of the design, administration, and scoring of MSPAP is necessary to understanding the MSPAP standard-setting process.

MSPAP is currently administered to all public school students in grades 3, 5, and 8. An annual edition consists of three forms composed of 10 to 12 tasks each. The tasks are composed of activities that relate to a theme that addresses a complex real-world problem. Each activity measures one or more challenging education outcomes that Maryland expects students to have achieved by the year 2000. Content covered includes reading, writing, language usage, science, mathematics, and social studies. Some tasks integrate content across two or more disciplines; some items measure outcomes in two content areas. Total engaged testing time per form is nine hours distributed across five days of testing. Tasks may be administered in one sitting or across several days.

Typical features of MSPAP tasks are preassessment activities, teacher demonstrations, group work that may involve data collection, and the extensive use of tools, materials, and manipulatives. Students work alone, in small groups, or as a total testing group. Responses based on group work are not scored. Responses that are scored are constructed independently by the students, who often have the option to write, graph, or draw an answer. The scoring tools used are rules, rubrics, and keys. Scores are aggregated across the three forms of an annual edition to yield school and district results. Results are in the form of outcome scores, outcome scale scores, proficiency levels, and performance standards.

MSPAP also contains survey questions that measure the tested students' perceptions of their interest/enjoyment, competence, utility of learning, and opportunity to learn these cognitively assessed disciplines. The test and survey questions are packaged into three test books per form: an examiner's manual, a student response book, and a student resource materials book.

Title I Inclusion

56

The Maryland State Department of Education developed *One System for All Children* cember 1995) to demonstrate the alignment of the Title I Assessment and School

Improvement Requirements with the Maryland School Performance Program. Within that document, adequate yearly progress for both schools and school systems for Title I is defined as "at standard or showing substantial and sustained progress in attaining the state's student performance standards" (page 3). Maryland defines *substantial progress* as a positive change in the School Performance Index that is at the 95 percent confidence level. The School Performance Index measures the average distance a school is from meeting standards across multiple databases.

Procedures for Establishing State Proficiency Levels and Proficiency Level Descriptions

Student achievement in each content area and grade level included in the MSPAP is described according to five proficiency levels, with Level 1 representing the highest level of proficiency and Level 5 representing the lowest. Cut scores on the MSPAP were developed for each of the five levels, as were descriptions of student performance at each level.

To develop the cut scores on each content area/grade level, teams of subject area experts matched assessment tasks and activities to the five proficiency levels. The teams then developed descriptions of each level by analyzing the tasks and activities assigned to the level. The result was a description of the knowledge, skills, and processes from the Maryland Learning Outcomes that were required by the tasks and activities at each proficiency level for each subject area and grade level assessed.

A modification of the judgmental behavioral anchoring method used by the National Assessment of Educational Progress was developed by the project team to establish and describe the MSPAP proficiency levels. The method involved two stages: (1) the identification of the cut points that define the proficiency levels for third, fifth, and eighth graders in mathematics, reading, science, social studies, writing and language usage; and (2) the description of performance in terms of content, behaviors, and skills displayed by students whose achievement is within these levels. Steps within these stages included (1) acquiring the materials, (2) training the facilitators, (3) developing the initial definitions, (4) selecting and convening Advisory Committee I, (5) refining the procedures following a debriefing session with Maryland State Department of Education (MSDE) staff, and (6) selecting and convening Advisory Committee II.

1. *Collection of materials*. The 1992 process required the collection and preparation of numerous materials for six content areas within three grade levels. These included the MSPAP test materials (student response book, student resource book, examiner's manual); scoring guides; the Maryland Learning Outcomes; the 1991 proficiency levels and their descriptions for mathematics, reading, and language usage/writing; the 1992 item score level locations and item score level information; 1992 item frequency distributions; and scale scores for each student taking the 1992 test.

2. *Training of facilitators*. Six facilitators were trained in two training sessions. The first, lasting half a day, preceded the convening of Advisory Committee I; the second, lasting one day, preceded the convening of Advisory Committee II.

3. *Development of initial definitions*. The procedures required definitions of the "desired" proficiency levels to guide the process. For reading, writing/language usage, and mathematics, the definitions were available from 1991. Advisory Committee I considered these definitions and the item/activity scale, and on the basis of professional judgment, classified each item/activity to a proficiency level. For science and social studies these definitions had to be developed by Advisory Committee I. The general revised procedures are shown in Figure 6.1 at the end of this chapter.

4. *Advisory Committee I*. MSDE identified and selected the members of Advisory Committee I. All members were required to have extensive training and experience in their respective content areas at their respective grade levels. Advisory Committee I was charged to work as a group to establish the proficiency levels for each content area and grade. In addition, each member pro-

vided a rating indicating the degree of confidence associated with his or her proficiency level recommendation. Half a day was allocated for the review of general issues regarding MSPAP and specific procedural issues. However, because of a keen interest in discussing the issues surrounding the process, the group completed the charge for only one content area, resulting in the need for a review and refinement of the procedures.

5. *Refinement of procedures.* Because Advisory Committee I, operating as a full group, failed to complete the charge for all of the six content areas within the allotted time, MSDE refined the procedures for Advisory Committee II. Subcommittees for each content area and grade were assigned.

6. *Advisory Committee II.* The criteria for membership on Advisory Committee II were essentially identical to those for membership in Advisory Committee I. Advisory Committee II was composed of five separate subcommittees in reading, writing/language usage, mathematics, science, and social studies. These subcommittees worked three full days. For reading and writing/language usage, the subcommittees were charged with establishing the descriptions for the proficiency levels previously defined by Advisory Committee I. For the other content areas, the subcommittees were charged with establishing the proficiency levels (day 1) and then defining the descriptions (days 2 and 3). The descriptions were in the language of the Maryland Learning Outcomes. If an outcome was not assessed at a given proficiency level, the description of that proficiency level did not include that outcome.

Details of the procedures are found in Atash (January 1994) and *Establishing Proficiency Levels and Descriptions for the 1992 Maryland School Performance Assessment Program (MSPAP)* (Westat, June 1993).

State Performance Standards

Maryland developed state performance standards (Thorn, Moody, McTighe, Kelly, & Peiffer, 1990) that define student achievement in five proficiency levels. The process of defining the performance standards involved three phases: consensus by educators, review and/or adjustment by a broader community of stakeholders, and public hearing. In phase I, educators using Delphi techniques reached consensus on performance levels that represent satisfactory and excellent achievements. Excellent is proficiency levels 2 and above. Satisfactory is proficiency level 3. In phase II, community leaders from business and government reviewed the recommendations and made adjustments to address the concerns of the constituents. In phase III, recommendations were submitted by the State Superintendent of Schools to the State Board of Education, which held public hearings prior to finalizing the performance standards.

A sample of the performance standards for Grade 5 reading and mathematics is shown in Figures 6.2 and 6.3 at the end of this chapter.

Promises, Problems, and Challenges of Maryland's Experience

The process of establishing proficiency levels and proficiency level descriptions was an ambitious undertaking in a limited time frame. It required extensive training and preparation of numerous materials. Many of the committee members had difficulty using the materials and following the written step-by-step procedures. Participants needed to discuss emerging general and specific issues related to establishing proficiency levels and descriptions at length. These discussions led to a review and refinement of the process in progress.

The use of proficiency levels and descriptions for describing student performance requires that the public, parents, teachers, and students be educated to interpret them appropriately.

Figure 6.1: Maryland procedures for establishing state proficiency levels and proficiency level descriptions

1. **Independently, for each outcome:**
   a. examine all activities for each proficiency level;
   b. write down the knowledge, skills, and processes in Maryland Learning Outcomes terms that each activity requires from students;
   c. on the form provided write down the characteristics (i.e., knowledge, skills, and processes in Maryland Learning Outcomes terms) in common among the activities for each proficiency level;
   d. also write down the characteristics (i.e., knowledge, skills and processes) which differentiate the activities across proficiency levels;
   e. write down any other observations you note with regard to the activities for each proficiency level.

2. **After you have completed Step 1, as a group:**
   a. summarize on a large sheet of paper your observations and comments for each proficiency level;
   b. refine the descriptions for each proficiency level using the language of the Maryland Learning Outcomes;
   c. discuss your committee's findings for all proficiency levels and on a separate sheet of paper summarize relevant comments, observations, and notes which might be useful in interpreting the descriptions of the proficiency levels.

3. **After you have completed steps 1 and 2 for Grade 3, repeat these steps for Grade 5 and Grade 8.**

4. **After you have completed steps 1 to 3 for all three grades:**
   a. examine the descriptions for the proficiency levels for all the three grades at the same time;
   b. refine the descriptions, if necessary, to reflect consistency and development across grade levels.
   c. finalize the notes to be "attached" to the descriptions (c-2).

NOTES:

These descriptions should:

I. be written in positive terms (i.e., what students can do);
II. reflect the Maryland Learning Outcomes by using terms from the Maryland Learning Outcomes statements;
III. be parallel in organization, language, and style;
IV. be written in clear and concise language without using unmeasurable qualifiers such as thoroughly, often, seldom, etc.;
V. be based on only those activities that are located at a proficiency level. If an outcome is not represented in a proficiency level, do not describe the proficiency level in terms of that outcome.[6.1]

[6.1] *Establishing Proficiency Levels and Descriptions for the 1992 Maryland School Performance Assessment Program (MSPAP).* (1993, June). Rockville, MD: Westat, Inc., Appendix F: "Step-by-Step Procedures for Committee Two."

Figure 6.2: 1993 MSPAP proficiency levels: Grade 5 reading

At proficiency levels 1, 2, 3, 4, and 5 students construct, extend, and examine the meaning of fifth-grade appropriate texts.

- Students at an MSPAP proficiency level are likely to be able to display most of the knowledge, skills, and processes at that level and lower proficiency levels.

LEVEL 1

When *reading for literary experience*, readers:

- Demonstrate a comprehensive understanding of the text.
- Make clear multiple connections and extensions of meaning between elements of the author's craft and the meaning of the text.
- Make clear connections and extensions between their ideas and the text.
- Support their opinions with relevant, explicit text-based information.
- Demonstrate a comprehensive understanding of literary elements.

When *reading to be informed*, readers:

- Demonstrate a comprehensive understanding of the text.
- Establish clear connections between their ideas and the text.
- State relevant opinions or personal judgments and support them with extensive, explicit references to the text.

When *reading to perform a task*, readers:

- Demonstrate a complex understanding of the text with evidence of connections, extensions, and examinations of meaning.
- Support inferences with explicit connections between their ideas and the text.

LEVEL 2

When *reading for literary experience*, readers:

- Demonstrate a developed understanding of the text.
- Establish clear connections between their ideas and the text.
- Support their responses with relevant, explicit text-based information.
- Demonstrate a developed understanding of literary elements.
- Make clear connections between elements of the author's craft and the meaning of the text.

When *reading to be informed*, readers:

- Demonstrate a developed understanding of informational sources.
- Make clear connections between their ideas and the text.
- State relevant opinions, personal judgments, or interpretations and support them with explicit references to the text.

When *reading to perform a task*, readers:

- Demonstrate a developed understanding of the text with evidence of connections.
- Establish connections between their ideas and the text.
- Support inferences with connections between their ideas and the text.

LEVEL 3

When *reading for literary experience,* readers:

- Demonstrate an understanding of the text.
- Make some connections and extensions between their ideas and the text.
- Support responses with text-based information.
- Demonstrate an adequate understanding of literary elements.
- Establish connections between elements of the author's craft and the meaning of the text.

When *reading to be informed,* readers:

- Demonstrate an adequate understanding of informational sources.
- Suggest connections between their ideas and the text.
- State opinions, personal judgments, or interpretations and provide some support for them with limited references to the text.

When *reading to perform a task,* readers:

- Demonstrate an adequate understanding of the text.
- Provide adequate evidence of constructing meaning.
- Apply graphic information.
- Integrate information from one or more texts.
- Use personal experience to elaborate ideas from the text.

LEVEL 4

When *reading for literary experience,* readers:

- Demonstrate a little understanding of what they read.
- Make minimal connections between their ideas and the text.
- Attempt to support their responses with minimal text-based information and/or personal experience.
- Demonstrate minimal understanding of literary elements.

When *reading to be informed,* readers:

- Demonstrate a superficial understanding of informational sources.
- Provide limited relevant connections between their ideas to the text.
- State relevant but unsupported inferences in connecting their ideas and the text.

When *reading to perform a task,* readers:

- Demonstrate little understanding of the text.
- Provide limited evidence of connection of meaning.
- Make limited extension between their ideas and the text.

LEVEL 5

When *reading for literary experience,* readers:

- Demonstrate inadequate understanding of fifth-grade appropriate texts.

When *reading to be informed,* readers:

- Demonstrate inadequate evidence of constructing the meaning of fifth-grade appropriate texts.

When *reading to perform a task,* readers:

- Demonstrate inadequate understanding of fifth-grade appropriate texts.

Students at Level 5 are likely to have provided some responses to assessment activities at Level 4, but not enough assessment activities to place them at proficiency Level 4.[62]

[62] *1996 MSPAP and Beyond Maryland School Performance Assessment Program: Score Interpretation Guide.* (1996, December). Baltimore, MD: Maryland State Department of Education, pp. 18-20.

Figure 6.3: 1993 MSPAP proficiency levels: Grade 5 mathematics

All 13 mathematics outcomes are assessed in the MSPAP at grades 5 and 8. All outcomes except algebra are assessed at Grade 3. Differences in the content assessed at each grade level result from the level of complexity of the concepts that are assessed and the language used in the tasks.

- Students at an MSPAP proficiency level are likely to be able to display most of the knowledge, skills, and processes at that level and lower proficiency levels.

LEVEL 1

Students at Level 1:

- Justify and explain results or solutions to open-ended problems and multi-step problems.
- Reason mathematically to solve problems involving geometric data using shapes and dimensions and to make predictions from patterns in a data chart.
- Apply mathematical thinking to real-world problems.
- Determine if a solution is sensible based on given criteria.
- Solve problems involving money, time, and elapsed time, and demonstrate understanding of the meaning of the operations.
- Distinguish among kinds of polygons, design geometric patterns, solve geometric problems involving measurement and spatial reasoning.
- Apply estimation of perimeter to real-world problems.
- Collect, organize, and display data for given situations using appropriate displays such as line plots, stem-leaf plots, bar graphs, pictographs, or glyphs.
- Make predictions using basic concepts of probability in abstract settings.
- Write a rule based on patterns; create and describe a geometric pattern.
- Write an algebraic equation based on a geometric model.
- Use an algebraic expression to explain a rule.

LEVEL 2

Students at Level 2:

- Justify and explain one or more results based on data from charts, graphs, or text that are organized and constructed by the students.
- Use mathematical language to interpret and support scientific conclusions and to communicate problem-solving strategies.
- Use diagrams to model situations.
- Reason mathematically to solve problems involving the concepts of proportion, area, and spatial reasoning.
- Use deductive reasoning to make predictions using data from a chart.
- Multiply and divide whole numbers and solve problems involving money and time.
- Distinguish among kinds of polygons.
- Construct a polygon with a given area.
- Apply estimation strategies to real-world problems involving measurement such as rate/distance.
- Collect, organize, and display data as line plots, charts, and tree diagrams and bar graphs; interpret bar graphs and stem and leaf plots.
- Use a function table to create a rule with an algebraic expression.
- Describe patterns involving connections between numbers and geometry.
- Evaluate algebraic operations.
- Make predictions using number theory and basic concepts or probability.

LEVEL 3

Students at Level 3:

- Justify and explain a result based on interpretation of data.
- Explain number relationships, geometric relationships, number concepts, and use of operations.
- Reason mathematically to make comparisons using information from a graphical display, to solve problems, to make predictions, and to compare the basic concepts of probability.
- Use statistical data to build an argument.
- Use all four arithmetic operations of whole numbers, fractions and decimals, including money. Choose an appropriate operation to solve a problem.
- Demonstrate an understanding of fractions and the relationship between whole numbers, fractions, and decimals.
- Demonstrate an ability to perform computations using numbers in a variety of equivalent forms.
- Identify symmetry and construct a circle given its radius.
- Use arithmetic operations to find area and perimeter.
- Select the appropriate tool of measurement and measure accurately.
- Measure angles and apply knowledge of congruency.
- Collect, organize, and display data in tables or charts; interpret data from glyphs and line plots; find the mean.
- Describe how a change in one variable results in a change in another variable.
- Develop a probability model for a real-world situation.

LEVEL 4

Students at Level 4:

- Use reasoning processes to interpret data from a chart and support a position.
- Identify symmetry and geometric patterns.
- Distinguish among kinds of triangles and quadrilaterals.
- Construct a polygon given the name and dimensions.
- Apply mathematical reasoning to a geometric configuration.
- Use arithmetic operations to solve real-life problems.
- Collect, organize, and display data in a table and model concepts of averaging.
- Describe relationships among data in a chart/table.
- Demonstrate the understanding of basic concepts of probability.
- Complete geometric patterns.
- Compare and order numbers.
- Create an hypothesis which reflects the understanding of the relationships between two variables.
- Solve for a missing number in a number sentence.

LEVEL 5

Students at Level 5:

- Use arithmetic operations to solve problems.
- Add and subtract whole numbers.
- Generalize a rule from a simple pattern.
- Interpret mathematical data and write a conclusion.
- Collect, organize, and display data in a table.

Students at Level 5 are likely to have provided some responses to assessment activities at Level 4, but not enough assessment activities to place them at proficiency level 4.[62]

[62] *1996 MSPAP and Beyond Maryland School Performance Assessment Program: Score Interpretation Guide.* (1996, December). Baltimore, MD: Maryland State Department of Education, pp. 28-30.

References

Atash, N. (1994, January). *Establishing Proficiency Levels and Descriptions for the 1993 Maryland School Performance Assessment Program (MSPAP).* Rockville, MD: Westat, Inc.

Establishing Proficiency Levels and Descriptions for the 1992 Maryland School Performance Assessment Program (MSPAP) (1993, June). Rockville, MD: Westat, Inc.

Governor's Commission on School Performance (1989). *The Report.* Baltimore, MD: Governor's Commission on School Performance.

Maryland State Department of Education (1995, December). *One System for All Children: The Alignment of Title I Assessment and School Improvement Requirements with the Maryland School Performance Assessment Program.* Baltimore, MD: Division of Compensatory Education and Support Services and the Division of Planning, Results, and Information Management, Maryland State Department of Education.

Maryland State Department of Education (1996, December). *1996 MSPAP and Beyond Maryland School Performance Assessment Program: Score Interpretation Guide.* Baltimore, MD: Maryland State Department of Education.

Maryland State Department of Education (1995, May). *Maryland's Schools for Success/Goals 2000 Plan.* Baltimore, MD: Maryland State Department of Education.

Maryland State Department of Education (1996, May). *State of Maryland: Final Consolidated State Plan.* Baltimore, MD: Maryland State Department of Education.

Rosenberger, K. (1997, September). *The Maryland School Performance Assessment Program (MSPAP).* Baltimore, MD: Maryland State Department of Education.

Thorn, P., Moody, M., McTighe, J., Kelly, & Peiffer, R. (1990). *Establishing Standards for Maryland's School Systems: A Systemic Approach.* Baltimore, MD: Maryland State Department of Education.

Chapter 7. The Oregon State Assessment System

Background/Content

The Oregon State Assessment System has its roots in legislative action. The Oregon State Board of Education adopted common curriculum goals, content standards, and benchmarks at grades 3, 5, 8, 10, and 12 in December 1996. These documents outline the Certificate of Initial Mastery (CIM) and the Certificate of Advanced Mastery (CAM) content domains to be assessed in English (reading, writing, and speaking), mathematics, science, social sciences, the arts, and second languages. Additionally, with the adoption of career-related learning standards in December 1996, workplace skills to be assessed for the CAM were defined. Together with benchmark monitoring assessments at grades 3, 5, and 8, the CIM and CAM are the primary tools for determining whether Oregon students meet rigorous academic content standards and are being prepared to move easily into a variety of career pathways.

The proposed assessment system is intended to provide individual student data to obtain a CIM and a CAM. This is a movement away from the former purpose of the assessment system, which was to gather information to inform program evaluation. The three types of assessment included in this new system are multiple-choice tests, on-demand performance assessments, and classroom-based work samples. Criterion-referenced multiple-choice tests based on Oregon content standards are currently administered by the state in mathematics and reading and will be administered in science in 1998. The social sciences test is under development. Students will be tested in writing and mathematics problem solving through large-scale, on-demand performance tasks. Classroom work samples will be required of students in all of the content domains identified in the standards document. Speaking, the arts, second languages, and career-related learning standards will be assessed locally. In 1999, students at the tenth-grade level will have the first opportunity to attain a CIM in English and mathematics. The other content domains will be added to the CIM certification at the rate of one per year until the CIM covers all six content areas included in the standards document by 2003.

Performance standards define how well students must perform on classroom and state assessments leading to the CIM. The standards are composed of two elements: (1) number, type, and minimum scores required on classroom assessments; and (2) minimum scores required on state assessments. To demonstrate achievement of the performance standards, students must complete classroom and state assessments showing what they know and can do in the required subject areas.

Classroom assessments vary from teacher to teacher and school to school. Local teachers and schools choose the resources, materials, and methods used to teach and assess students. Students are required to complete a set number and type of classroom assessments. A state scoring guide sets forth the requirements necessary to achieve the standards. Classroom assignments may be used as the required classroom assessments if they are complex enough to be scored on all dimensions of the scoring guide and require students to apply what they have learned in a new situation.

State assessments at grades 3, 5, 8, and 10 contain multiple-choice questions, essay questions, and/or mathematics problem-solving questions requiring students to solve problems and show their work.

There are two scoring systems: one for state multiple-choice assessments, and one for classroom assessments and state essay and problem-solving tests. These two systems are described below.

Multiple-choice questions on the state test have a single correct answer. Students receive a scale score based on the number of correct answers compared to the total number of questions on the test, taking into account the difficulty of the questions on the test. Classroom assessments and state essay and problem-solving assessments require students to produce original work. Students are scored along a scale of 1 to 6 in several different areas:

6	**Exemplary**	Work at this level is both exceptional and memorable. It shows a distinctive and sophisticated application of knowledge and skills.
5	**Strong**	Work at this level exceeds the standard. It shows a thorough and effective application of knowledge and skills.
4	**Proficient**	Work at this level meets the standard. It is acceptable work that demonstrates application of essential knowledge and skills. Minor errors or omissions do not detract from the overall quality.
3	**Developing**	Work at this level does not yet meet the standards. It shows basic but inconsistent application of knowledge and skills. Minor errors or omissions detract from the overall quality. Work needs further development.
2	**Emerging**	Work at this level shows a partial application of knowledge and skills. It is superficial, fragmented or incomplete and needs considerable development. Work at this level contains errors or omissions.
1	**Beginning**	Work at this level shows little or no application of knowledge and skills. It contains major errors or omissions.

Title I Environment

Oregon is poised to meet Title I requirements with the assessment system as it is currently described. Content standards have evolved and have been framed by large representative state groups. They will be revisited periodically to better align them with evolving assessments and appropriate classroom practices. The multiple-choice tests reflect knowledge of the content standards at the benchmark levels; the performance standards expressed as cut scores are set to measure student progress toward a proficient level of performance. In addition, standards are being set to make sure that performance in content processes, skills, and applications are addressed and measured through performance tasks and work samples. These are evaluated using six-point scoring guides that describe the type of evidence students must exhibit to be judged proficient (at level 4 on the scoring guides). Student work is scored using an analytic trait model, and results are expressed in terms of whether the student meets the standard in each of the content area criteria. For example, in science there have been five criteria identified: application of scientific content and concepts, formulation of questions, design of investigation, collection of data, and analysis and interpretation.

Current Program Status

The Oregon Educational Act for the 21st Century, passed by the state legislature in 1991, required Oregon to set much higher standards for all students in English (reading/literature, writing, and speaking), mathematics, science, social studies (history, civics, geography, and economics), arts, and second language. Students in the class of 2001 will have the first opportunity to earn a CIM in English and math when they are sophomores in the 1998–1999 school year. The CIM will be awarded to students who achieve the Grade 10 standards on state tests and classroom work samples. Although school districts will continue to issue diplomas, certificates will be even more meaningful to colleges and future employers because they represent achievement of a high level of knowledge and skills.

The new standards describe what students should know and be able to do in English, mathematics, and other basic subjects. The following information summarizes grade level expectations.

GRADE 8

Reading/Literature benchmark = 231 scale score
A student who achieves a score of 231 or above on the Grade 8 state reading/literature tests is able to distinguish between novels, short stories, plays, poetry, and nonfiction; determine the meaning of unfamiliar words by contextual clues and other means; and distinguish between facts and opinions. The student can draw conclusions about the meaning of relationships, images, patterns, or symbols in the writing; and analyze whether the writer's conclusion is validated by evidence of character, plot and setting.

Mathematics multiple-choice benchmark = 231 scale score
A student who achieves a score of 231 or above on Grade 8 state multiple-choice math tests is able to perform calculations with whole numbers, fractions, decimals and integers; recognize and use percents, scientific notation, square roots, and exponents; use scale drawings and apply formulas to calculate distance, perimeter, area, volume and angle. The student can compare and make predictions using experimental and theoretical probability; solve linear and nonlinear equations; and use coordinate geometry to solve problems.

Mathematics problem-solving benchmark = 4 in each of four areas
A students who achieves a score of 4 in each of four areas on Grade 8 state math problem-solving tests is able to solve problems accurately; use models, diagrams and symbols accurately; use models, diagrams, and symbols to show mathematical concepts in problems; apply graphic and/or numeric models to solve problems; organize and explain reasoning; review work and evaluate its reasonableness.

Writing benchmark = 4 in each of four areas
A student who achieves a score of 4 in each of four areas on Grade 8 state writing tests is able to convey clear, focused main ideas with accurate and relevant supporting details; develop a clear beginning, middle, and end; use complex sentences; and use correct spelling, grammar, and punctuation.

GRADE 10

Reading/Literature benchmark = 239 scale score
A student who achieves a score of 239 or above on Grade 10 state reading/literature tests is able to analyze whether the writer's argument, action or policy is validated by evidence cited; evaluate the effectiveness of theme, conflict, and resolution in the piece; identify and examine the treatment of similar themes in various works; analyze the impact of literary devices such as figurative language, allusion, dialect, and symbolism; and identify the writer's purpose.

Mathematics multiple-choice benchmark = 239 scale score
A student who achieves a score of 239 or above on Grade 10 state multiple-choice math tests is able to perform numeric and algebraic calculations; convert numbers to decimals, fractions, percentages, exponents in scientific notation or integers; and use formulas, scale drawings or maps to calculate measurements. The student also can use experimental or theoretical probability to solve problems; use recursive relationships and matrices to represent and solve problems; use linear, exponential and quadratic functions; and apply the Pythagorean Theorem and other properties of figures to solve geometric problems.

Mathematics problem-solving benchmark = 4 in each of four areas
A student who achieves a score of 4 in each of four areas on Grade 10 state math problem-solving tests is able to select and use relevant information in the problem to solve it accurately; apply graphic and/or numeric models to solve the problem; organize and explain reasoning; review work and evaluate its reasonableness.

Writing benchmark = 4 in each of four areas
A student who achieves a score of 4 in each of four areas on Grade 10 state writing tests is able to convey clear, focused main ideas with accurate and relevant supporting details; develop a clear beginning, middle and end with logical transitions between ideas and paragraphs; use parallel and other appropriate sentence structure; and use correct spelling, grammar, and punctuation.

As the assessment pieces are developed and implemented, there is ongoing work to ensure the quality and integrity of the system. Because Oregon has a Proficiency-Based Admissions Standards System (PASS) under development for admission to state colleges and universities, it is critical to align the systems relative to content standards, performance standards, and assessment requirements. The CIM assessment system has been defined and is under development. The CAM and PASS assessment systems are not yet finalized. Therefore, collaborative efforts are still under way to ensure that the systems overlap and integrate with one another. For example, a test given for CIM may satisfy a requirement for CAM and/or PASS. The English and mathematics testing components, developed by Oregon teachers and a contractor, are in place. However, the science test was administered for the first time in the spring of 1998; the social sciences test is under development. Performance standards have not yet been set for either science or social sciences. In the area of science, there is also work currently under way to help teachers identify quality tasks and become acquainted with the scientific inquiry scoring guide. Working from this basis of understanding, teachers can begin to develop a bank of possible tasks and models for developing and scoring tasks to meet classroom work sample requirements for the CIM.

To meet the requirements of Title I, there is still work to be done to create and/or identify multiple-choice test items for a wider range of student ability levels. Currently, tests are constructed to measure whether students are performing at the benchmark level of proficiency. Therefore, the band of confidence for the scores is not wide enough to determine accurately the level of performance of many Title I students. This presents a problem for determining adequate yearly progress, unless off-level testing (tests given at other than benchmark years) is done or items to test a wider range of abilities are included on the state test.

There is also a need to look critically at the work sample component of the assessment system to see what information about Title I students can be provided from this source. Oregon has a technical advisory committee of national experts who are conducting research on several aspects of the program. An in-state assessment advisory panel continues to inform and refine the evolving standards-based system in Oregon.

Additional References

Oregon has produced a variety of documents to communicate and further describe and explain the content and performance standards and assessment system. These include content and performance standards documents, eligible content, test specifications, a variety of teacher packets, and several documents intended for the public.

Chapter 8. The Wyoming Comprehensive Assessment System

Background

Of the 48 states participating in the 1992 National Assessment of Educational Progress mathematics test, Wyoming's fourth-grade students ranked tenth. By 1996, Wyoming's fourth-graders dropped in rank to twenty-third, despite scoring higher than the 1992 cohort. The pattern for Wyoming's eighth-grade students was similar. An examination of states that ranked below Wyoming in 1992 but above Wyoming in 1996 (e.g., Michigan, Montana, North Carolina, Texas, Vermont) indicated that many such states had implemented massive statewide education reforms, including standards-based, large-scale assessment programs.

Although local school districts in Wyoming had been developing academic content standards since 1990, it was not until 1997 that the Wyoming Supreme Court mandated that the state develop a standards-based assessment system to monitor the academic achievement of Wyoming's students. The statewide assessments are to be focused on measuring student progress in mathematics, reading, and writing at grades 4, 8, and 11.

At the time of the Supreme Court ruling, Wyoming had already begun the process of developing state model content standards to meet the requirements of the Improving America's Schools Act. In response to state legislation (the Wyoming Enrolled Act 2), the state built upon the local efforts, using a "bottom up" approach. In this approach, regional groups were convened to draft sets of standards in language arts (i.e., reading and writing) and mathematics for grades 4, 8, and 11. The regional groups consisted of representatives from local school districts within the region as well as representatives from community colleges, universities, and business communities. The regional groups were charged with drafting consensus-based standards, using the standards previously developed at the district level and considering national standards.

Once consensus was reached within each of the regional groups, a district representative was selected by the group to participate in a state-level committee charged with drafting content standards based upon a consensus process and using the regional consensus-based drafts. Again, national standards were considered in order to verify and support a case for rigor in the Wyoming standards.

Program Description

The proposed assessment system provides results at the state, district, school, and student levels. The concept of a statewide assessment system is a new one in Wyoming. Its design recognizes that the consequences associated with public reporting may vary by jurisdiction, resulting in a variety of individual and community responses. Therefore, a comprehensive system that could allow for profiling strengths and weaknesses over time at a variety of levels (e.g., student, school, district, state) and for differing populations (e.g., students of different genders and levels of poverty, minorities, students with disabilities, English language learners) was considered highly desirable.

Although the assessment results and subsequent school improvement decisions are crucial aspects of Wyoming's school accreditation review process, the system is not designed for high stakes such as promotion or graduation at the individual student level. In other words, it is primarily intended as a school improvement model.

To assure that reading/language arts and mathematics content standards are adequately covered, multiple assessment formats will be used in the new comprehensive system. Multiple-choice items will be used to provide broad, efficient coverage across the content standards. Constructed-response tasks will be used to provide more in-depth coverage of the content standards by assessing higher level skills. Extended response tasks will be used to assess students' highest levels of [illegible]ting relative to advanced content and skills. For example, in reading/language arts, extended [illegible]nse tasks might be used to assess comprehension in ways that ask students to move beyond

traditional or typical comprehension questions by finishing a story or developing an alternate ending to a passage. In mathematics, extended response tasks might require students to solve multistep problems, produce mathematical proofs, or provide detailed explanations or justify their approach to solving a particular mathematical problem.

In order to capture enough detailed information at the school level for making school improvement decisions, a mix of matrix-sampled tasks and common tasks will be used for the extended response and constructed-response portions of the assessments. The multiple-choice sections of the tests will include enough multiple-choice items for each student so that a matrix-sampling approach will not be necessary.

The writing assessment will require students to write extended responses to each of two prompts, one common to all students and one as part of a matrix sample. This methodology will allow for reporting information at both the individual student and school levels with a minimum of testing time.

The standards-based assessments in reading and mathematics are expected to require no more than 2.5 hours per content area. The writing assessment is anticipated to require less time.

In order to allow for national comparisons — which is not possible with the unique-to-Wyoming standards-based assessments — a nationally norm-referenced, standardized test will be included as part of Wyoming's Comprehensive Assessment System. The intent is for the norm-referenced testing in mathematics and reading/language arts to require no more than 2.5 hours total.

Process: Performance Descriptors, Benchmarks, and Standards

Once content standards were drafted and a comprehensive assessment system was proposed and designed, performance standard descriptors were drafted. These descriptors were drafted by the state committees for each of the major strands contained in the content standards. The descriptors included three levels: advanced, proficient, and partially proficient. The descriptors aim to describe how well students must perform the content standards to be judged at each of the three levels.

Upon completion of the initial draft, the content standards and performance standard descriptors were forwarded to all school districts for comment. These comments informed subsequent revisions that were also sent to local school districts for response. Focus group techniques were used to gather feedback from content organizations such as state affiliates of the National Council of Teachers of Mathematics, Association of Supervision and Curriculum Directors (national and state chapters), National Council of Teachers of English, Math and Science Coalition (national and state chapters), and business/community representatives.

The state standards-setting committees then incorporated the results of the focus groups and local data collection efforts in a revised version disseminated in the spring of 1998 for public comment. Public comment is being solicited through written response, telephone, the Internet, and e-mail. Drafts are also available at public schools, public libraries, and the Wyoming Department of Education Web site.

Another round of revisions to the performance standards, as well as benchmark setting, is anticipated following the pilot administration of the new statewide standards-based assessments. The target date for piloting is spring 1999.

Because the state does not yet have a standards-based assessment system in place, performance descriptors were developed for each major content area strand. Once the actual assessments have been administered, performance levels can be refined, using student work as exemplars, and cut scores defining each of the performance levels will be set.

Title I Requirements

Although economics and financial factors have been major drivers in Wyoming's movement to measure student performance progress, federal Title I legislation has provided an important impetus to Wyoming's effort. In fact, Title I legislation has encouraged Wyoming to continue to

move forward with its push for a comprehensive assessment system for all students and has simplified development of the standards and assessment components of the state's consolidated plan. Many of the requirements of the Title I law are also requirements of Wyoming's Enrolled Act 2, which calls for the development of state and local standards and assessments with data aggregations/disaggregations similar to those required by Title I (e.g., data for traditionally underserved students).

Summary/Next Steps

Content standards have been developed in reading/language arts and mathematics. These standards will be adopted by June 1998. To date, performance descriptors have been developed for each content area and at three benchmark levels (grades 4, 8, and 11). The legislature approved funding for the assessment in the winter 1998 legislative session. The performance levels selected in Wyoming are advanced, proficient, and partially proficient.

Next steps in the assessment development process:

- issued a request for proposals in April 1998, with responses due in May and a target date for awarding a contract in July 1998;
- pilot the new standards-based assessments in the spring of 1999 along with the nationally norm-referenced test; and
- fully implement the Wyoming Comprehensive Assessment System by the year 2000.

Section III:
Technical Advances in Developing Performance Standards

Although research on methods of setting cut scores on selected-response assessments is plentiful, only recently have research efforts addressed assessments that include constructed-response and performance-based items and tasks. As part of a research project sponsored by the Council of Chief State School Officers (CCSSO) and funded by the National Science Foundation (NSF), Ronald Hambleton, Richard Jaeger, Craig Mills, and Barbara Plake are conducting a series of studies investigating standard-setting procedures for such assessments. The researchers have worked closely with state assessment staff (CCSSO's Technical Guidelines for Performance Assessment collaborative) to design studies that address critical questions in developing performance standards for standards-based assessment techniques.

This section contains two reports resulting from the research studies. Craig Mills and Richard Jaeger report on a method for developing performance descriptors (i.e., statements that describe student performance at various levels). Their initial findings are described in Chapter 9. In Chapter 10, Ronald Hambleton describes work in the area of setting cut scores on complex performance assessments. His report is also a result of the larger NSF-funded research project with Richard Jaeger, Barbara Plake, and Craig Mills.

As noted earlier, the terminology used to describe the various components of performance standards is not consistent across, or even within, different educational specialties. In the following research reports, the authors use terms as they are commonly used in the area of educational measurement.

Chapter 9. Creating Descriptions of Desired Student Achievement When Setting Performance Standards[9.1]

Craig N. Mills
Educational Testing Service
and
Richard M. Jaeger
University of North Carolina at Greensboro[9.2]

When tests are used as gatekeepers to occupational licensure or professional certification, defining acceptable levels of performance is an integral part of test development and use. Test developers define minimum test scores that examinees must achieve in order to receive licensure or certification. Similar standard setting takes place in screening applicants for driver's licenses and applicants for enlistment in the armed forces.

More recently, setting performance standards has become prominent in educational testing. Two situations provide ready examples. In the first, adequate performance on one or more tests has been deemed requisite to students obtaining important educational rewards, such as promotion to the next grade or receiving a high school diploma. In the second, named performance categories are used as descriptors in an attempt to communicate the results of student assessments, but important consequences for individual students may not be associated with their test performances. Various statewide student assessment programs (e.g., those of Connecticut and North Carolina) and the National Assessment of Educational Progress (NAEP) illustrate the second situation.

Since 1990, NAEP results have been reported in terms of "achievement levels" defined by the National Assessment Governing Board (NAGB), a practice that has been praised by many users of NAEP results but is a subject of controversy among specialists in the field of educational measurement (Cizek, 1993; Forsyth, 1991; Kane, 1993; Linn & Dunbar, 1992; Phillips, Mullis, Bourque, Williams, Hambleton, Owen, & Barton, 1993; Shepard, Glaser, Linn & Bohrnstedt, 1993; U.S. General Accounting Office, 1993). More specifically, NAEP reports have included data on the percent of students in each tested grade level whose test scores resulted in their achievement being classified as "below basic," "basic," "proficient," or "advanced." The press has focused on these statistics in their articles on NAEP results (Jaeger, 1996, March).

When students are classified on the basis of their test scores, it is necessary to determine the range of test scores associated with each defined achievement category and, more fundamentally, to define the meaning of the labeled categories. It is this latter task that is the subject of this chapter.

As part of a study on the methodology of setting performance standards on assessments that include multiple-choice items and performance exercises, we worked with colleagues to develop several new standard-setting methods (Hambleton & Plake, 1997, March; Jaeger & Mills, 1997, March). Because it provided an excellent vehicle for exploring the issues under study, we applied our methods to the 1996 NAEP Science Assessment for students in Grade 8. In conducting our research, expert science teachers and science curriculum supervisors with whom we worked noted that some parts of the description of advanced student performance adopted by the NAGB for the 1996 Grade 8 Science Assessment was not covered by the content of the single NAEP test booklet used in our research. We later found that this problem was not unique to NAEP's 1996 Science Assessment but has been encountered earlier in NAEP's history as well, prior to the adoption of currently-used standard-setting procedures. As noted by Phillips, et al. (1993),

[9.1]This material is based upon work supported by the National Science Foundation under Grant No. 9554480. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This paper was completed while the second author was a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University. Financial support was provided by The Spencer Foundation under Grant Number 199400132.

> The historical approach to establishing achievement levels for [NAEP] mathematics and reading has been a model where content is first specified through frameworks, item specifications and assessment items and tasks are written consistent with the frameworks, the assessments are administered, and then achievement levels are developed. One of the problems this approach creates is that there may be a lack of consistency between the achievement level standards and what the assessment actually measures. (As an example, the eighth-grade mathematics basic achievement level description indicates that students should be able to solve problems using computers. Computer usage is not a part of the NAEP mathematics assessment.) The question arises, what are the validity issues for assessing progress toward standards when the assessment instrument may not reflect the standards? (p. 78)

This is precisely the issue that led us to examine an alternative approach to development of definitions of achievement levels, and to evaluation of the effect of using such descriptors on resulting cut scores and on the percent of students who would be classified as "below basic," "basic," "proficient," or "advanced." The alternative descriptors we developed differed from those specified by NAGB in that they were grounded explicitly in the item and exercise content of the NAEP Grade 8 Science Assessment booklet used in our research. We want to emphasize that the methods described in this chapter, although evaluated in the context of NAEP, have broad applicability. Similarly, the problem that gave rise to this research is unlikely to be exclusive to NAEP.

In the following sections, we describe the methods used to develop alternative achievement level descriptors for the 1996 National Assessment in Science for students in Grade 8, provide preliminary results of an analysis of the effects of using these alternative descriptors, and comment on the significance of these preliminary findings and their implications for setting performance standards in other settings and venues.

Methodology

The strategy used to develop achievement level descriptors that were grounded in the content of the booklet used in this research consisted of seven steps:

1. A panel of subject matter experts was convened and provided with instruction on the task to be completed.
2. Panelists studied the exercises and items contained in one booklet of the 1996 National Assessment of Science for students in Grade 8.
3. Panelists reviewed the framework prepared by NAGB for the 1996 National Assessment of Science for students in Grade 8.
4. Panelists reviewed the generic descriptors of student performance at the basic, proficient, and advanced levels that had been adopted by NAGB as a matter of policy.
5. Panelists identified elements of test content in the 1996 NAEP Grade 8 Science booklet that they judged to be examples of student performance associated with NAGB's generic descriptors of student performance at the basic, proficient, and advanced levels.
6. Panelists defined student performance on elements of test content that were requisite to students' classification as basic, proficient, and advanced in Grade 8 science.
7. Panelists reached consensus on content-grounded descriptions of student performance at the basic, proficient, and advanced levels.

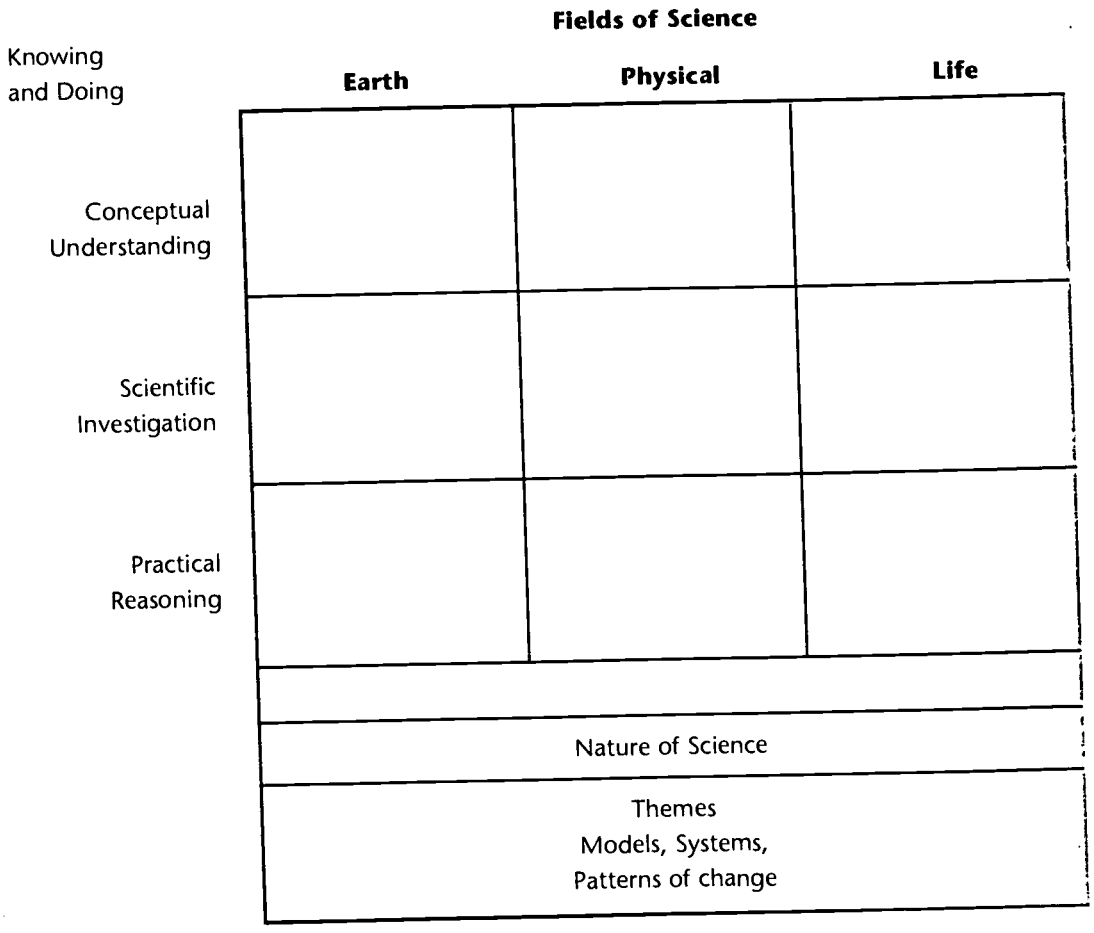
1. *Convening and instructing the panel.* A panel of six teachers was convened to develop performance descriptors. All were female. Three were African American and three were white. All six had participated in a standard-setting study approximately five months earlier in which they judged student performance on a single booklet in the NAEP Grade 8 Science Assessment using either an anchor-based or a holistic rating method. Panelists met for one day.

Following a welcome and introductions, panelists were informed that they had been convened to develop performance descriptors for the performance categories advanced, proficient, and basic on a single booklet of the NAEP Grade 8 Science Assessment. An overview of the NAEP testing design was provided, including a reminder that the assessment consists of multiple test booklets that are administered to disjoint samples of students throughout the nation. Panelists were reminded of the study in which they had participated and of the difficulty they had expressed relating test performance to the achievement-level descriptors provided during that study.

2. *Review of the items and exercises in a NAEP Science Assessment booklet.* Each panelist received a copy of Booklet 226F of the NAEP Grade 8 Science Assessment. Panelists were instructed to review the booklet independently to familiarize themselves with its content. Scoring guides were not provided. Although a discussion of scoring would have been appropriate, it was deemed unnecessary in this study since panelists had previously received extensive training on scoring guides and sample responses at each score level for each item and exercise in the booklet. Panelists spent approximately 30 minutes reviewing the exercises and items in the test booklet. Had they not been familiar with the booklet on the basis of their earlier participation in our research, we suspect that more time would have been required.

3. *Review of the test framework.* Panelists were given a handout showing the content matrix for the 1996 NAEP Science Assessment (NAGB, no date). This handout is shown in Figure 9.1. Each panelist was instructed to review the test booklet once again, marking the cells in the content matrix that were assessed by at least one item or exercise in the booklet. Panelists worked independently. Some simply marked cells as they noted content in the booklet that represented a cell. Others recorded item numbers. When all panelists had completed this review, they discussed the content coverage of the test booklet as a group. All panelists agreed that all cells in the content matrix were represented on the test booklet with one exception: No items on the test booklet were judged to assess scientific investigation in life sciences. Life sciences was judged to be the least represented content area overall. Panelists agreed that some items and exercises could be classified in more than one cell of the content matrix.

Figure 9.1: Content matrix for the 1996 NAEP science assessment



4. *Review of NAGB's generic definitions.* A handout containing the official NAGB policy definitions of achievement at the basic, proficient, and advanced levels was given to the panelists (see Figure 9.2 for these definitions). These definitions were discussed. Of particular importance in this discussion was the generic nature of NAGB's definitions. NAGB's achievement level policy definitions do not address specific student skills or grade levels. Panelists were told that their job was to write new performance descriptors that would provide a direct link between NAGB's generic definitions and the content of the Grade 8 Science Assessment booklet they had reviewed. At this point, panelists also were given copies of new performance level descriptors that NAGB had recently developed. Panelists were informed that the new descriptors might or might not be more closely aligned with the content of the Grade 8 Science booklet than the ones they had used in the previous standard-setting study. Panelists were invited to use these new descriptors as they developed their own achievement level descriptors, but they were not required to do so. None of the panelists used the new NAGB descriptors during the remainder of the study, so those definitions are not included in this paper.

Figure 9.2: National Assessment of Educational Progress achievement levels policy

The Achievement Levels Policy

The 1988 NAEP legislation creating NAGB directed the Board to identify "appropriate achievement goals...for each subject area" that NAEP measures. The 1994 NAEP reauthorization reaffirmed many of the Board's statutory responsibilities including "developing appropriate student performance standards for each age and grade in each subject area to be tested under the National Assessment." Following this directive and striving to achieve a primary mandate of the 1988 statute, "to improve the form and use of NAEP results," the Board has been developing student performance standards (called achievement levels) for NAEP since 1990. The Board has adopted achievement levels in mathematics, reading, U.S. history, world geography, and science.

The achievement levels adopted by the Board and used here to report the performance of students on the 1996 NAEP Science Assessment are developmental, and as such, are currently being evaluated by the National Academy of Sciences (NAS). The NAS findings will be available in late 1998.

The Board has framed the policy for the achievement levels to help answer the questions "How good is good enough?" The goal is to report NAEP results in terms of the quality of student achievement by defining levels of learning linked to a common body of knowledge and skills that all students should attain, regardless of their backgrounds. The Board defined three levels for each grade: Basic, Proficient, and Advanced. These levels are cumulative in nature, that is, it is assumed that students at the Proficient level are likely to be successful at the Basic and Proficient content and students at the Advanced level are likely to be successful at the Basic, Proficient, and Advanced content. Table 1 presents the policy definitions of the achievement levels.

Table 1
Policy Definitions of NAEP Achievement Levels

Achievement Level	
Advanced	Superior Performance
Proficient	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	Partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Source: National Assessment Governing Board

5. *Linking of test content with NAGB's generic definitions.* Panelists were instructed to work independently to write a description of proficient Grade 8 Science performance that linked NAGB's generic definition of proficient student performance with the content of the test booklet. Panelists began with the proficient achievement level so as to define an anchor point that could be modified when they developed descriptions of basic and advanced student performance on the Grade 8 Science Assessment.

Each panelist received a blank form on which they were to record their description of proficient performance (see Appendix 9.1). Copies of each panelist's description were made and distributed to all panelists. A group discussion was then held with the intent of developing a consensus description of proficient Grade 8 Science performance. However, individual panelists differed substantially in the specificity of their descriptions. Some panelists had written very specific iptions, while others had provided far more general statements. As a consequence, panelists ged in a discussion of the level of specificity that was most appropriate to the standard-set-

ting task in which their descriptors would be used. Some panelists favored developing a checklist for each performance level while others felt that this approach was infeasible in light of time constraints. Panelists ultimately agreed to develop somewhat general descriptions of requisite student performance at each achievement level.

In order to define better the level of generality to be reflected in their performance descriptors, panelists again reviewed the test booklet as a group. They developed two lists on the basis of the items and exercises in the test booklet. The first list was titled "What do students have to do?" This list was composed entirely of verbs such as "recall," "recognize," "analyze," and "interpret." The second list was titled "What do they do it to?" and identified content areas covered by the test booklet. This list was divided into three content categories: Earth Science, Physical Science, and Life Science, consistent with the dimensions of the content matrix discussed previously. The more specific content areas identified by the panel were listed under these three generic headings. They were:

Earth Science
- Renewable and Non-Renewable Resources
- Earth's Place in the Universe
- Changes in the Earth's Surface
- Composition/Formation of the Earth

Life Science
- Ecology
- Factors/Needs Influencing Living Things
- Organization of Living Things

Physical Science
- Energy (heat, light, sound)
- Forces of Nature (gravity)
- Properties of Matter (density)

6. *Defining student abilities associated with basic, proficient, and advanced performance.* Panelists were divided into three pairs, one for each content area. Assignment was not random. Panelists indicated their preferences and volunteered for content areas. Each pair of panelists received a blank matrix (see Appendix 9.2) containing cells in which they could record content areas and write descriptions of abilities (knowledge and skills) that would be exhibited by students who performed at the basic, proficient, and advanced levels in those content areas. Panelists were instructed to complete descriptions for all three achievement levels for each content area and then to proceed to the next content area. More specifically, panelists were instructed to:

1. Record the first content area in the column labeled "Content Framework."
2. Record the performance a proficient student would exhibit on this content, considering the items and exercises in the test booklet, in the column labeled "Proficient."
3. Record the performance a basic student would exhibit on this content, again as assessed by the items and exercises in the test booklet, in the column labeled "Basic."
4. Record the performance an advanced student would exhibit on this content, again as assessed by the items and exercises in the test booklet, in the column labeled "Advanced."
5. Repeat Steps 1 through 4 for each content area.

When each pair of panelists had completed this form for all content headings within its assigned content area, copies of the completed forms were made and were distributed to all panelists. A group discussion followed. Modifications were made until there was group consensus that the descriptions accurately represented the content of the test booklet and appropriately represented basic, proficient, or advanced student performance on that content. In some cases, descriptive statements were moved from one performance category to another. In one case, panelists were unable to provide a description of performance at the advanced level. They noted that the test booklet did not include any advanced material for that content area.

During this group discussion, panelists identified another set of skills, "Scientific Processes," that was assessed by the items and exercises in the test booklet but not reflected in their descriptors of student performance. Two areas of scientific process, "Using Scientific Equipment" and "Designing Experiments," were identified. Descriptors of basic, proficient, and advanced student performance in these two areas were developed through group discussion.

7. *Development of consensus descriptions of basic, proficient, and advanced student performance.* Narrative descriptors were then developed using the statements for each performance level within each content area. The narrative descriptors were developed by the researchers by grouping together the consensus statements for each performance level from all content areas on the test. This was done following the meeting.

Results

Panelists' consensus definitions of content-grounded student performance at the basic, proficient, and advanced levels are as follows:

Basic: Students performing at the Basic level should be able to recall the definition of natural resources and recognize methods for conserving them. They should recall basic facts about the solar system. Basic students can recall the definitions of earthquakes, erosion, weather, pollution, and agriculture. They also recall the theories of the earth's formation and recognize the commonly occurring elements and compounds in the earth. They are able to recall the definition of energy and its source and can describe visible changes that result from the activities of man or nature. Students performing at the basic level can recall the basic requirements for living organisms and the structural organization in living things from the cell organelle to organism. These students recognize that energy can be transferred. They are able to describe gravity as a force of nature. They can recall the existence of a relationship between material and its displacement. They have the ability to manipulate equipment and follow directions, although they may make some procedural errors. They can complete, read, and/or locate information in simple graphs and tables and recognize the presence of variables in an experimental design.

Proficient: Proficient students are able to recall which natural resources are renewable and which are non-renewable, predict how living things can influence natural resources, and explain how to conserve natural resources. They can comprehend the organization and formation of solar systems, galaxies, and the universe as well as explain how the forces of the universe govern heavenly bodies. A proficient student can interpret data to conclude that forces are changing the earth's surface and predict how and to what extent forces change the earth. Students performing at the proficient level can comprehend (by restating or explaining) theories of the formation of the earth and interpret charts and graphs showing the composition of the earth. These students have the ability to describe the transfer of energy through an ecosystem. They can explain the effect of changes in the environment on both living and non-living things. Proficient students can analyze a set of data to determine the effects of the basic requirements of organisms and explain the result of deficiencies of those requirements. They can describe structural organization from cell organelle through community and the relationships between and among levels. Students at this level can comprehend that energy occurs in many forms that can be transferred to do work. They also comprehend the relationship of gravity to mass and/or distance. They are able to comprehend and graph the relationship of density to the displacement of matter. They can manipulate equipment and follow directions correctly and thoroughly and are able to design simple experiments. Proficient students can complete, read, and/or locate data in complex graphs and tables.

Advanced: Students performing at the advanced level have the ability to analyze the value of natural resources. They can also analyze the implications of forces in the universe as they

apply to objects. Students at the advanced level can analyze how changes may affect life forms and how life forms may cause changes. They can predict and make conclusions based on data collected and analyzed from ecological activities. Advanced students are able to interpret information about the factors influencing organisms in various situations. These individuals can analyze the structural organization and the interaction of living things, from cell organelle through biome. They can predict and analyze information related to energy and its transformation. They have the ability to predict and calculate the effect of changing mass and distance on gravity. These students can interpret a situation concerning density and the displacement of matter. They can make predictions and draw conclusions based on data presented or entered in tables and graphs and can design comprehensive, controlled experiments to test hypotheses including incorporation of the concept of repeatability in the design.

These descriptors, based on the content of the booklet used in this study, can be compared with the more general, total-framework-based definitions of basic, proficient, and advanced performance on the NAEP Grade 8 Science Assessment that we obtained from the National Assessment Governing Board and used in our first study involving this assessment. Those definitions follow:

Basic. Students should possess fundamental knowledge concerning both the structure and function of human anatomy. They should know the main causes of common diseases. In addition, basic students should be aware of their immediate environment including concepts of the diversity of living things and food chains. In the physical world, they should be able to distinguish states of matter and understand the basic properties and characteristics of matter. They should be able to identify common energy sources and methods for transforming energy. Basic students should be able to make accurate measurements and display the data. At the basic level, students should be able to infer information from the simple tests they make and be able to identify forces that alter the Earth's surface; describe the composition of the Earth, its atmosphere and climate; and describe the major features of the solar system and universe.

Proficient. Students should know and/or be able to collect basic information and apply it to the physical, living, and social environments. They should be able to link simple ideas in order to understand payoffs and tradeoffs. Proficient students should be able to understand cause and effect relationships such as predator/prey and growth/rainfall. Proficient students should be able to design experiments to answer simple questions involving two variables, to isolate variables, and to collect and display data and draw conclusions from them. They should be able to draw relationships between two simple concepts; they should be starting to understand relationships (such as force and motion and matter and energy) and they should be beginning to understand the laws that apply to living and nonliving matter.

Advanced. In addition to the ability to infer cause and effect relationships from data, advanced students should be beginning to visualize interacting systems and subsystems at various levels. For instance, they should be able to relate several factors (variables) to explain a phenomenon. They should be able to describe many elements of a system, select a particular example, and explain its limits. They should be able to understand more complex models and know that scientific models have limits. Advanced students should be beginning to understand the nature and limits of science and that science is subject to change. Students at the advanced level should have a knowledge of genetics, cells and the communications systems; know basic laws of probability; and be able to express them quantitatively. They should be able to describe basic chemical processes and how chemicals and classes of chemicals interact and account for physical properties in terms of their physical state and atomic structure. They should be able to understand more abstract concepts/theories related to the Earth's climate, atmosphere, the solar system, and the universe. The advanced student should be able to manipulate variables and form hypotheses in the abstract as well as in concrete settings.

The Effect of Achievement-Level Descriptors on Assessment Results

The two sets of achievement level descriptors shown above are clearly different. However, comparison of these descriptors will not reveal the effect of either set on resulting performance standards or on the percentages of students who would be classified as below basic, basic, proficient, and advanced were one or the other set of descriptors to be used.

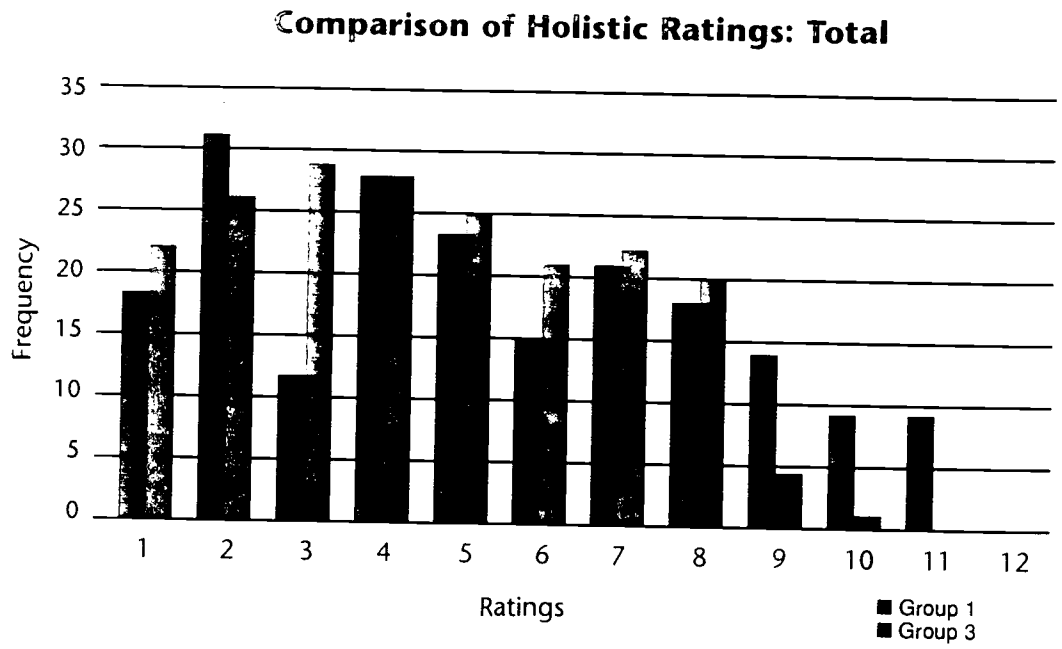
Following the development of the new descriptors, a standard-setting study was conducted in which, as one component of the design, one panel of four science teachers based their classification of students' responses to the items and exercises in the NAEP Grade 8 Science test booklet on the new descriptors and an independent panel of four science teachers based their classification of students' responses on the descriptors used on our initial standard-setting study. The panels were composed through random assignment of teachers. All other standard-setting conditions (orientation, training, review of test content, tests reviewed, etc.) were held constant for the two panels.

The results of this study have not been fully analyzed. However, preliminary results suggest that the definitions of achievement levels have important effects on cut scores and on the percentages of students classified at each achievement level. Table 9.1 contains preliminary results for both panels, based on the teachers' classifications of the same sample of 50 students' responses to the exercises and items in the test booklet. The distribution of the two panels' classifications of the 50 students is shown in Figure 9.3. A twelve-point scale was used in the classification exercise.

Table 9.1: Comparison of standard-setting results using old and new descriptors

		Performance Category			
		Below Basic	**Basic**	**Proficient**	**Advanced**
Number of Ratings in Category	Old	77	74	46	1
	New	62	66	53	18
Raw Score Cut Scores	Old		29	42	49
	New		27	38	47
% At or Above Category	Old		56	25	8
	New		57	35	15

Figure 9.3: Distribution of ratings using new (Group 1) and old (Group 3) descriptors on a 12-point scale



It is clear that different results were obtained for the two panels. Panelists using the new descriptors classified proportionately more students at the high end of the scale than did panelists who used the old descriptors. The resulting cut scores also are different. At all three achievement levels, the new descriptors resulted in lower cut scores. Although corresponding cut scores differed by only two to four raw-score points, they have a pronounced effect on the percentages of students in this sample who were placed in the proficient or advanced categories.

Conclusions and Implications

In this study, we developed and pilot tested methodology designed to ground performance descriptors directly in the content of a test booklet. Cut scores established with such descriptors were compared to those based solely on an assessment framework. The methodology proved to be practical and feasible. Panelists reported that the content-grounded descriptors were better suited to standard setting than were descriptors based on an assessment framework. Independent panels developed cut scores using one or the other set of definitions. Preliminary results suggest that the difference in the descriptors had a noticeable effect on panelists' assignment of students to performance categories and on the percentages of students whose performances were described as below basic, basic, proficient, and advanced.

Since different students complete different NAEP booklets, the standard-setting approach evaluated here—grounding achievement-level definitions in the content of a booklet—would have to accommodate differences in test booklet difficulty within a given NAEP assessment and across assessments. One possibility would be use of equating adjustments for differences in difficulty. We have not examined the degree to which cross-booklet variation within NAEP administrations or across administrations is a problem of practical consequence, or how it should be addressed, if at all. Since all NAEP exercises are scaled to a common set of banked parameters within an assessment, and equating adjustments are routinely used across assessments, grounding achievement-level descriptors in the content of a single booklet for the purpose of setting achievement levels might not prove to be problematic. This issue clearly requires additional consideration and research.

The quality of public schooling in the United States is evaluated in the press and by many affected consumers and taxpayers in terms of the percentages of students whose achievement is classified as basic, proficient, and advanced on NAEP assessments. The sensitivity of such percentages to procedures used to define these achievement levels suggests the need for thoughtful research on the methodology of performance standard setting in general, and on the methodology of developing achievement level definitions in particular. We regard this study as a beginning, and would conclude that it provides a useful illustration of the need for additional inquiry.

References

Cizek, G. J. (1993). *Reactions to National Academy of Education Report, "Setting Performance Standards for Student Achievement."* Unpublished manuscript.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice,* 10(3), 3–9, 16.

Hambleton, R. K., & Plake, B. S. (1997, March). An anchor-based procedure for setting standards on performance assessments. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago.

Jaeger, R. M. (1996, March). Reporting large-scale assessment results for public consumption: Some propositions and palliatives. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, New York.

Jaeger, R. M., & Mills, C. N. (1997, March). A holistic procedure for setting performance standards on complex large-scale assessments. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago.

Kane, M. (1993). *Comments on the NAE Evaluation of the NAGB Achievement Levels.* Unpublished manuscript.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement,* 29, 177-194.

National Assessment Governing Board. (no date). *Science Framework for the 1996 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP Scales.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting Performance Standards for Student Achievement: An Evaluation of the 1992 Achievement Levels.* A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment. Stanford, CA: National Academy of Education.

U.S. General Accounting Office. (1993). *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* (Rep. No. GAO/PEMD-93-12). Washington, DC: U.S. General Accounting Office.

Appendix 9.1

Panelist: _____

Individual Definition of "Proficient" Performance on the Grade 8 NAEP Science Assessment 226F

Appendix 9.2

Content Area: _____

Analytical Breakdown of the Definitions of Basic, Proficient, and Advanced

Content Framework	BASIC Students at this level should be able to:	PROFICIENT Students at this level should be able to:	ADVANCED Students at this level should be able to:

Chapter 10. Setting Performance Standards on Achievement Tests: Meeting the Requirements of Title I

Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

Probably the most challenging problem today in educational assessment concerns setting performance standards on the test score scale to separate students into performance categories (e.g., certifiable and not certifiable). In the case of Title I programs, the problem is even more challenging because at least two performance standards must be set for classifying students into three performance categories: advanced, proficient, and partially proficient.

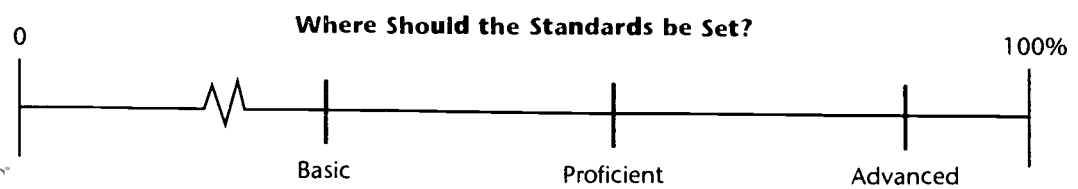
It is well known that there are no true performance standards waiting to be discovered through research studies. Rather, setting performance standards is ultimately a judgmental process that is best done by appropriate individuals who (1) are familiar with the test purpose and test content and knowledgeable about the standard-setting method they will use, (2) have access to item or task performance data as well as test score distribution data to set a framework for their judgments, (3) understand the social and political context in which the tests are being used, and (4) are aware of the consequences of their decisions (e.g., the passing rate associated with the possible standards that might be chosen). The goals of this chapter are to provide descriptions of some of the issues and the most common methods for setting performance standards and to offer a checklist for implementing a standard-setting method.

Introduction

One of the primary purposes of educational assessments is to make mastery/non-mastery decisions about students in relation to well-defined domains of content. In one important example, the task is to determine whether students have achieved a sufficiently high level of performance on the educational assessment to receive a high school diploma. This requires a performance standard or cutoff score on the test score scale, or some other scale on which achievement is reported, to separate students into two performance categories, often labeled "masters" and "non-masters" or "certifiable" and "not certifiable."

With the National Assessment of Educational Progress (NAEP), students are separated, based upon their performance, into four performance categories called "advanced," "proficient," "basic," and "below basic." In the context of Title I programs, the expectation is that multiple standards will be used to separate students into more than two performance categories. Three standards must be set to separate students into four mastery categories; for example, "advanced," "proficient," "partially proficient," and "below partially proficient." Figure 10.1 highlights the performance standards on a typical test score scale. Many Title I programs will use two or three performance standards.

Figure 10.1: A typical test score scale and three performance standards—Basic, Proficient, and Advanced



In this chapter, issues and methods associated with setting performance standards on educational assessments will be addressed. In addition, a checklist is provided to assist school districts and state departments working through the process of setting performance standards on their educational assessments. The chapter has been organized into several sections: In the next section, 11 steps for setting standards on an educational assessment are offered. Each step is described briefly. In subsequent sections, readers will learn about some of the common methods for setting standards — several that apply to selected-response items such as those in the multiple-choice format, and several that apply to constructed-response items such as writing samples and other performance tasks. Finally, some of the most important issues that arise in standard setting will be addressed, and a checklist for conducting a standard-setting study will be presented.

Three points are important to make at the outset. First, it is important to clearly distinguish between *content standards* and *performance standards*. Content standards refer to the curriculum and what students are expected to know and to be able to do. Students, for example, might be expected to carry out basic mathematics computations, read a passage for comprehension, or carry out a science experiment to identify the densities of different objects. Performance standards refer to the level of performance that is expected of students to demonstrate, for instance, basic, proficient, and advanced level performance in relation to the content standards. In other words, performance standards reflect how well students are expected to perform in relation to the content standards (Linn & Herman, 1997).

For example, we might require students to solve 10 of 20 basic mathematics computations to be judged as basic, whereas we may require that students solve 14 of 20 problems to be judged as proficient. In reading comprehension assessment, students may be expected to answer correctly 60 percent of the questions to be judged as basic, 80 percent to be judged as proficient, and 90 percent to be judged as advanced. Content standards should be thought of as what we expect students to learn, whereas performance standards indicate the levels of expected performance of students on the educational assessments constructed to assess the content standards. But performance standards may not always be scores on a test score scale. They may correspond to verbal descriptions that can be used in classifying student test performance into performance categories.

Second, all standard-setting methods in use today involve judgment and they are arbitrary. Some researchers have argued that arbitrary standards are not defensible in education (Glass, 1978). Popham countered with this response:

> Unable to avoid reliance on human judgment as the chief ingredient in standard setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as arbitrary, hence unacceptable.
>
> But *Webster's Dictionary* offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is, "Determinable by a judge or tribunal." The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, "selected at random and without reason." In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.
>
> But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd (Popham, 1978, p. 168).

Performance standards for educational assessments used in Title I programs will be set arbitrarily, but the goal is to set them in the best sense of the word "arbitrary." The method used to set the performance standards should be carefully planned, and it should be carried out by persons who are qualified to set the standards. Many teachers, curriculum specialists, and administrators would be well qualified to participate in the standard-setting process for Title I programs. Sometimes, representatives of the public may be asked to participate too.

Sometimes performance standards may be set too high or too low quite unintentionally by a panel. Through experience and carefully designed follow-up validity studies of the scores and decisions that are made, performance standards that are not well positioned can be identified and revised.

Finally, on the one hand, methods for setting standards on educational assessments using the multiple-choice item format are well developed and validated, and steps for implementation are clear. Most districts and states have set defensible performance standards using one of the acceptable methods (e.g., Angoff, Ebel, contrasting groups) which will be described later in this chapter. On the other hand, standard-setting methods for performance assessments such as writing samples and performance tasks are not well developed at this time, and certainly none of them have been fully researched and validated. It is simply not possible at this time to advance fully validated methods. In this chapter, several of the most promising methods will be described. Readers are referred to Hambleton, Jaeger, Plake, and Mills (1998) for a review of methods and issues for setting standards on performance assessments.

Typical Steps in Performance Standard Setting

Perhaps the best way to defend a particular set of performance standards on an educational assessment is to demonstrate that a reasonable process was followed in arriving at the final standards (Hambleton & Powell, 1983). If the process reflects careful attention to (1) selection of panelists, (2) training, (3) aggregation of data into a final set of standards, (4) validation of performance standards, and (5) careful documentation of the process, the defensibility of the resulting standards is considerably increased. If, on the other hand, panelists are chosen because (1) they live near the meeting site, (2) they are willing to work over a weekend, or (3) they happen to be known by the coordinator of the meeting, questions should be raised about the resulting standards. Other common problems that, when present, reduce the validity of the performance standards include the use of ambiguous descriptions of the performance standards, failure to train panelists fully on the standard-setting method, failure to allow sufficient time for panelists to complete their ratings in a satisfactory manner, and failure to validate and document the process that was implemented to set performance standards.

A presentation of 11 steps follows:

1. Choose a panel (large and representative of the stakeholders).

   Discussion. Who are the stakeholders in the decisions that will be made with the educational assessments? These are the persons who should be involved in the standard-setting process. In the case of NAEP, teachers, curriculum specialists, policy makers, and the public (30 percent by law) make up the standard-setting panels. Fifteen to 20 persons are often placed on a panel to provide the diversity that is needed (geographical, cultural, gender, age, technical background, educational responsibilities) and to provide stable estimates of the performance standards (Jaeger, 1991). In the case of Title I programs, many of the same groups would seem to be relevant for inclusion on a performance standard-setting panel.

2. Choose one of the standard-setting methods, prepare training materials, and finalize the meeting agenda.

   Discussion. There are many acceptable methods for setting performance standards, and several of these will be considered later in this chapter (see, for example, Hambleton, Jaeger, Plake, & Mills, 1998; Jaeger, 1989; Livingston & Zieky, 1982). Some of these methods focus panelists' attention on the items and tasks in the assessment, and other methods focus panelists' attention on the students and their work on the items and tasks in the assessment.

89

It is especially important to use training materials that have been field tested. For example, a miscalculation on the time required to complete various steps in the process may result in panelists needing to rush their ratings to complete their work on time.

3. Prepare descriptions of the performance categories (e.g., advanced, proficient, partially proficient).

Discussion. In recent years, time spent on defining the performance level descriptions has increased considerably in recognition of the importance of the descriptions. In setting performance standards on the NAEP, for example, more than two full days are spent on the activity of preparing descriptions. If panelists are to set defensible standards, performance levels need to be clearly articulated. Panelists are requested to consider the performance of borderline students on the assessment material, or they may be required to classify student work using the performance level descriptions. When these descriptions are unclear, the whole process is flawed, and the resulting standards can be questioned. A critical step in the process, then, is for the panel (or a prior panel) to develop descriptions of students in each performance category. Recently, Mills and Jaeger (1998) produced the first published set of steps for producing test-based descriptions of performance levels, and these steps will be of interest to readers.

In the first example below, the descriptions used recently in the setting of Grade 4 performance standards in the area of reading on the NAEP appear. These descriptions provide an idea of the level of detail that it is assumed panelists need to complete their rating tasks:

> *Basic.* Demonstrates an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences, and extend the ideas in the text by making simple inferences.
>
> *Proficient.* Demonstrates an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.
>
> *Advanced.* Generalizes about topics in the reading selection and demonstrates an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought.

As a second example, descriptions of four levels of performance on the Pennsylvania Grade 8 mathematics assessment are provided below:

> *Novice.* Novice students demonstrate minimal understanding of rudimentary concepts and skills. They occasionally make obvious connections among ideas, providing minimal evidence or support for inferences and solutions. These students have difficulty applying basic knowledge and skills. Novice students communicate in an ineffective manner.
>
> *Apprentice.* Apprentice students demonstrate partial understanding of basic concepts and skills. They make simple or basic connections among ideas, providing limited supporting evidence for inferences and solutions. These students apply concepts and skills to routine problem-solving situations. Apprentice students' communications are limited.

Proficient. Students performing at the proficient level demonstrate general understanding of concepts and skills. They can extend their understanding by making meaningful, multiple connections among important ideas or concepts, and provide supporting evidence for inferences and justification of solutions. These students apply concepts and skills to solve problems using appropriate strategies. Proficient students communicate effectively.

Advanced. Students at the advanced level demonstrate broad and in-depth understanding of complex concepts and skills. They make abstract insightful, complex connections among ideas beyond the obvious. These students provide extensive evidence for inferences and justification of solutions. They demonstrate the ability to apply knowledge and skills effectively and independently by applying efficient, sophisticated strategies to solve complex problems. Advanced students communicate effectively and thoroughly, with sophistication.

These descriptions provide a good idea of the level of detail it is assumed panelists need to set performance standards.

4. Train the panelists to use the method (including practice in providing ratings).

   Discussion. To achieve this goal, effective training and practice exercises will be needed. Effective panelist training would include (1) explaining and modeling the steps to follow in setting standards, (2) showing the scoring keys and/or scoring rubrics and ensuring they are understood, (3) completing easy-to-use rating forms, (4) offering practice in providing ratings, and (5) explaining any normative data that will be used in the process.

   In addition, panelists need to be informed about factors that may affect student performance and should be considered in the standard-setting process — for example, (1) the role of time limits for the assessment, (2) the artificiality of educational assessments (panelists need to remember that when a student chooses to write a story, the student will often select the topic, have unlimited time to complete the work, and will often prepare several drafts — characteristics that are often not present in the typical writing assessment), (3) distractors in multiple-choice items that may be nearly correct (and, therefore, increase the difficulty of the item for students), and (4) the role of guessing behavior on performance on multiple-choice items.

   Finally, administering the assessment to panelists is often an effective way to demonstrate to them the knowledge and skills that students must possess to obtain a high score. It is assumed that panelists are likely to set more realistic performance standards if they have experienced the assessment themselves. The assessments always appear more difficult to panelists when they are completed without the aid of the scoring keys and scoring rubrics!

5. Compile item ratings or other data from the panelists (e.g., panelists specify expected performance of borderline basic students).

   Discussion. This step is straightforward if the training has been effective. A summary of the panelists' ratings can be prepared. For example, suppose panelists are asked to judge the minimum expected performance of proficient students on a task with a five point scoring rubric (e.g., 0 to 4). The median or typical rating and the range of ratings of the panelists could be calculated. Later (step 6), this information can be provided to the panelists and used to initiate discussion about the performance standard for proficient students.

6. Conduct a panel discussion: Consider actual performance data (e.g., item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on inter-panelist and intra-panelist consistency.

91

Discussion. With several of the test-based standard-setting methods that will be described in the next two sections, panelists are asked to work through the method and set preliminary standards and then to participate in a discussion of these initial standards and actual student performance data on the assessment. The purposes of the discussion and feedback are to provide the opportunity for panelists to reconsider their initial ratings and to identify errors or any misconceptions or misunderstandings that may be present.

The precise form of the feedback depends on the method, but, with several methods, the feedback might include average performance and student score distributions on the items or tasks of the assessment and descriptive statistics of the panelists' ratings.

More elaborate forms of feedback are also possible. For example, it is possible to determine the extent to which panelists are internally consistent in their ratings (van der Linden, 1982). Panelists who set higher performance standards on difficult tasks than easier tasks would be identified as being "inconsistent" in their ratings. They would be given the opportunity to revise their ratings or explain the basis for their ratings. Sometimes the so-called "inconsistencies" in the ratings can be defended, but, regardless, panelists would rarely be required to revise their ratings if they were comfortable with them. For a full review of factors affecting ratings, readers are referred to Plake, Melican, and Mills (1991).

7. Compile item ratings a second time (could be followed by more discussion and feedback). This iterative process is common but not essential. Typically, a two-stage rating process is used: panelists provide their first ratings (independent of other panelists or performance data of any kind), discussion follows, and then panelists complete a second set of ratings.

   Discussion. Following the discussion phase of the process, panelists are instructed to provide a second set of ratings. It is not necessary that panelists change any of their initial ratings, but they are given the opportunity to do so. Sometimes this iterative process is continued for another round or two. For example, in some of the NAEP standard-setting work that has been done (Hambleton & Bourque, 1991), panelists went through five iterations of ratings and discussions.

   Not all standard-setting researchers are committed to the use of discussion and feedback in the process. For example, with performance assessments, their argument is that better (i.e., more stable) performance standards will result if panelists spend their time rating more student responses, because the main effect of discussions and feedback is to achieve a consensus in ratings, but rarely are the performance standards set initially with the first set of ratings changed to any substantial degree. The competing argument is that it is important for panelists to discuss their ratings and receive feedback. Sometimes discussion and feedback will alter the performance standards, and even small changes can be of practical consequence; standard errors are almost certainly lower, and discussion and feedback may increase panelist confidence and acceptance of the resulting performance standards.

   Panelists like this step very much (or at least they report that they do on post evaluations), appreciate the opportunity to discuss their ratings with their colleagues, find the feedback valuable, and sometimes, performance standards do shift significantly up or down, especially when the feedback is a surprise to panelists (Hambleton & Plake, 1997).

8. Compile panelist ratings and average to obtain the performance standards.

   Discussion. At this stage, panelists' ratings are compiled to arrive at the performance standards. Often, this is simply an average of the performance standards set by each panelist. Median ratings may be preferable with small samples or non-symmetric distributions of ratings.

9. Present consequences data to the panel (e.g., passing rate).

Discussion. One step that is sometimes inserted into the process involves the presentation of consequential data to panelists. Panelists are informed about the percentage of students who would be located in each performance category. For example a panel might be shown the following chart.

Category	Percent of Students
Advanced	7.0%
Proficient	33.2%
Partially Proficient	42.5%
Below Partially Proficient	17.3%

If these findings were not consistent with the panelists' experiences and sense of reasonableness, they could be given the opportunity to revise their performance standards. Panelists may feel that a performance standard that resulted, for example, in 80 percent of the students being classified as below partially proficient is simply not reasonable or consistent with other available data about the students, and they may want to lower the standard for partially proficient students. And, in so doing, the number of partially proficient students would be increased, and the number of below partially proficient students would be decreased.

10. Revise, if necessary, and finalize the standard(s), and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards.

Discussion. Again, panelists are given the opportunity to revise their ratings to increase or decrease their performance standards. In addition, a panelist evaluation of the process should be conducted. One sample evaluation form appears in Appendix 10.1 (this is a modified version of an evaluation form used by Hambleton, Jaeger, Plake, and Mills) and can be used as a basis for generating an evaluation form for particular standard-setting initiatives.

11. Compile technical documentation to support the validity of the standards.

Discussion. It is important not only to be systematic and thoughtful in designing and carrying out a performance standard-setting project but it is also necessary to document the work that was done and by whom. Such a document will be valuable in defending the performance standards which have been set. A good example of documentation is provided in the report by Hambleton and Bourque (1991). Often the group setting the performance standards is advisory to a board that ultimately must set the standards. Technical documentation of the process is valuable information for the board.

Performance Standard-Setting Methods

There are several well-established methods in the measurement literature for setting performance standards on achievement tests, and they can be organized into two main categories: methods in which panelists are focused on a review of test content, called "test-based methods," and methods that are focused on the students themselves, called "student based methods." A brief description of the methods that are applicable to selected-response items follows. Follow-up references for readers include Berk (1986) and Jaeger (1989). In the literature, nearly all of these methods are described in terms of a single standard. These methods can be extended to setting multiple standards by simply repeating the process itself for more than one performance standard.

93

Test-Based Methods

With the test-based methods, individual items are studied in order to judge how well a borderline student will perform on the test items or tasks. The borderline student is someone who has a proficiency score located right at the performance standard. In the case of Title I, there may be three borderline students: one at partially proficient, one at proficient, and one at advanced. The ratings process described next is repeated for each borderline student.

Panelists are asked to assess how or to what degree a student who could be described as borderline would perform on each item or task. The choice of method is inserted into steps 2 and 4 in the standard-setting process.

Nedelsky Method

With the Nedelsky (1954) method, panelists are asked to identify distractors in multiple-choice test items that they feel the borderline student will be able to identify as incorrect. The assumption is then made that the borderline student would be indifferent to the remaining answer choices, and therefore he or she would choose one of the remaining choices at random. The minimum passing level or performance standard for that item then becomes the reciprocal of the number of remaining answer choices. For example, suppose a panelist reviews a test item and feels that a borderline student would recognize that two of the available five choices are incorrect. The expected score for this borderline student (i.e., the performance standard) then is 0.33, since the assumption is made that all remaining choices (three remain) are equally plausible to the borderline student. This rating process is carried out for borderline partially proficient, proficient, and advanced students.

The panelists proceed with each test item in a similar fashion and, on completion of the rating process, each panelist sums the minimum passing levels across the test items to obtain a performance standard. A panelist's standard is the expected score on the test for the borderline student. Individual panelists' performance standards are averaged to obtain a standard that is considered to be the best estimate of the standard.

Often a discussion of the panelists' ratings will then take place (see the section, *Typical Steps in Performance Standard Setting*), and panelists will have the opportunity to revise their ratings if they feel revisions are appropriate. And often panelists do make revisions, since misreading of test items, overlooking of important features of test items, and even some carelessness in making the ratings are common in the item-rating process. After panelists provide a second set of ratings, again, each panelist's item ratings are summed to obtain a standard on the test, and then the panelists' standards are averaged to obtain a standard based upon the ratings of all of the panelists.

The standard deviation of the panelists' standards is often used as an indicator of the consensus among the panelists (the lower the standard deviation, the more consensus there is among the panelists on the placement of the standard). When the variability is large, confidence in the standard produced by the panelists is lessened. Very often the goal in standard setting is to achieve a consensus among the panelists.

Ebel's Method

With the Ebel (1972) method, panelists rate dichotomously scored test items along two dimensions: relevance and difficulty. There are four levels of relevance in Ebel's method: essential, important, acceptable, and questionable. These levels of relevance are often edited or collapsed into two or three levels when the method is used in practice. Ebel used three levels of item difficulty: easy, medium, and hard. These levels of relevance or importance and difficulty can be used to form a 4 x 3 grid for sorting the test items. The panelists are asked to do two things:

1. Locate each of the test items in the proper cell, based on their perceived relevance and difficulty.

2. Assign a percentage to each cell representing the percentage of items in the cell that the borderline students should be able to answer.

The number of test items in each cell is multiplied by the percentage assigned by the panelist and the sum of these products, when divided by the total number of test items, yields the performance standard. As with all of the judgmental methods, the standards set by the individual panelists are averaged to obtain a final standard. An example of a 3 x 3 rating form is displayed in Appendix 10.2. The method can be generalized to polytomously scored test items.

Angoff's Method

When using Angoff's method (Angoff, 1971), panelists are asked to assign a probability to each dichotomously scored test item directly, thus circumventing the analysis of a grid or the analysis of answer choices. Each probability is to be an estimate of the "borderline student" answering the test item correctly (for example, the borderline basic student). Individual panelist-assigned probabilities for items in the test can be summed to obtain a standard, and then the panelists' standards can be averaged to obtain a final standard. This process is repeated for each performance standard of interest. A sample rating form appears in Appendix 10.3.

Here is one example of the Angoff method instructions to panelists who set the 1990 performance standards on the NAEP Mathematics Assessment, from the *Handbook for Panelists:*

> For the *Borderline Basic* student, your task is to specify the probability that this borderline student should answer each item in the assessment correctly. This chance or probability for each test item can range from zero (where you would be specifying that the borderline student should have no chance of giving a correct answer) to 1.00 (where you would be specifying that the borderline student should, without a doubt, answer the item correctly). After specifying the performance level for the Borderline Basic student on an item, you should provide estimates on the same item for the *Borderline Proficient* and *Borderline Advanced* students. (Hambleton & Bourque, 1991, p. 114)

Panelists would then work their way through the complete set of test items. Sometimes panelists are encouraged to think of 100 borderline students, and then estimate the number of these borderline students who should answer an item correctly. For many panelists, this seems to be an easier task than estimating the probability of correct performance on an item by the borderline student.

As with the other judgmental methods, common practice is to repeat the probability assignment process following discussions among the panelists about their assigned probabilities. Often, too, panelists are provided with item statistics, or information that addresses the consequences (i.e., passing and failing rates) of various standards to aid them in the standard-setting process. Item statistical information often has a substantial effect on the resulting standards (Taube, 1997).

Table 10.1 displays the hypothetical ratings of a panelist in setting performance standards for basic, proficient, and advanced students. The performance standards set on the second round of ratings are averaged over all panelists in arriving at the final set of recommended performance standards.

The method has been applied successfully to multiple-choice test items, and in a modified form to performance data (see, for example, Hambleton & Plake, 1995). For example, suppose a standard for separating below partially proficient and partially proficient on a performance task is the goal. Panelists, using a variation on the Angoff method, might be asked to specify the expected number of score points on the performance task (i.e., the standard) for the borderline student. Performance standards from each of the panelists can be averaged to obtain a final performance standard that would be used for classifying students on the performance task.

Table 10.1: Calculation of performance standards for a single panelist using the Angoff method for two sets of ratings

	Basic		Proficient		Advanced	
Item	R-1	R-2	R-1	R-2	R-1	R-2
1	.30	.35	.70	.65	.80	.80
2	.40	.40	.65	.65	.85	.85
3	.25	.28	.50	.45	.70	.65
4	.60	.55	.70	.70	.95	.90
5	.70	.65	.80	.80	.90	.90
6	.30	.30	.40	.45	.75	.80
7	.20	.20	.40	.45	.70	.70
8	.50	.50	.60	.65	.85	.85
9	.60	.55	.70	.75	.90	.90
10	.45	.45	.75	.80	.85	.85
Performance Standard	4.30	4.23	6.20	6.35	8.25	8.20

Student-Based Methods

With these methods, judgments are made about the mastery status of a sample group of students from the population of interest. For example, suppose the goal was to classify students into one of four performance categories: below partially proficient, partially proficient, proficient, and advanced. In the school context, these judgments would come from the teachers. The choice of method determines the nature of the required judgments. Next, the members of the groups for whom mastery determinations have been made are administered the test. Details are offered next for analyzing the judgmental data and the test scores.

Borderline-Group Method

This method requires that a description be prepared of each performance category. Several examples were presented earlier for step 3 in the section, *Typical Steps in Performance Standard Setting*. In practice, teachers who are familiar with the academic accomplishments of the students are asked to submit a list of students whose performances would be so close to the standard or borderline that they could not be reliably classified. The test is administered to these "borderline" groups, and the median test score for each group (e.g., "borderline partially proficient," "borderline proficient," and "borderline advanced") may be taken as the standard. Alternately, other decisions may be taken for arriving at the standards.

Contrasting Groups Method

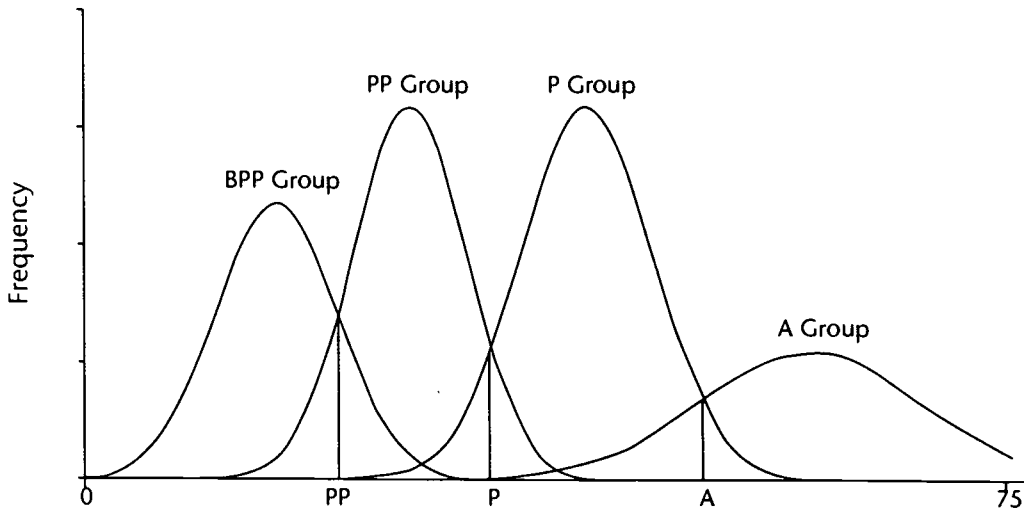
Working with the description of students in each performance category, teachers in, for example, a random sample of schools in a state are asked to classify their students into these performance categories or groups. The test is administered to the groups, and the score distributions for the groups are compared. The point of intersection is often taken as the initial standard (Berk, 1976). An example is given in Figure 10.2. With four groups, first the point of intersection of the advanced and proficient distributions is determined to select the advanced performance standard. Then the proficient and partially proficient distributions are compared to determine the proficient standard, and so on.

A standard can be moved up to reduce the number of false positive errors (students identified as advanced by the test but who were not in the advanced group formed by the teachers) or down to reduce the number of false negative errors (students identified as proficient by the test but who were in the advanced group formed by the teachers). The direction to move each performance standard will depend on the relative seriousness of the false positive and false negative errors. For example, which is the more serious error: to deny a high school certificate to a student who deserves it or to award a certificate to a student who does not? The answer to this question will influence the final placement of the performance standards. To minimize false negative errors, performance standards should be lowered. To minimize false positive errors, performance standards need to be raised.

If the score distributions overlap completely, no classifications of students can be made reliably. The ideal situation would be one in which the two distributions did not overlap at all. Then, the performance standard can be positioned between the two distributions, and the assignment of students to performance categories would be in complete agreement with the teachers' assessments.

The validity of this approach to standard setting depends, in part, on the appropriateness of the panelists' classifications of students. If the teachers tend to err in their classifications by assigning students to higher groups than they belong, the result is that standards from the contrasting groups method are lower than they should be. On the other hand, the standards tend to be higher if teachers err by assigning students to lower performance groups than they belong. Like the Angoff method, or the modified Angoff method, the contrasting groups method can be applied to performance assessment data, too.

Figure 10.2: Application of the contrasting groups standard setting method



BPP—Below Partially Proficient
PP—Partially Proficient
P—Proficient
A—Advanced

Some Practical Guidelines for Setting Performance Standards

A number of researchers have suggested guidelines to follow in setting and/or reporting performance standards (Cizek, 1996a, 1996b; Hambleton & Powell, 1983; Livingston & Zieky, 1982; Plake, 1997). An updated list of guidelines follows for setting performance standards via *test-based methods:*

1. The importance of the classifications of students to performance categories should impact substantially on the effort that is committed to the standard-setting process. With important tests, such as those used in awarding high school diplomas, and assigning students to Title I programs, substantial effort should be committed to producing defensible standards and this effort would include compiling evidence to support the validity of the standards.

2. The design of the standard-setting process should be influenced by the panelists (and their backgrounds), test length, and test item formats. For example, inexperienced panelists may require substantial amounts of training, long tests may require that sub-groups of panelists be formed with each group assigned a different portion of the test, and some assessment formats such as performance measures will require modifications to the common methods for setting standards (Hambleton, Jaeger, Plake, & Mills, 1998).

3. With important standard-setting initiatives, the full process should be field tested prior to using it operationally. Serious errors can often be avoided with carefully conducted and evaluated field tests of the standard-setting process.

4. The selection and number of panelists should be given considerable attention. Do the panelists represent the main constituencies, and are there enough panelists to produce stable standards? The defensibility of the resulting standards depends very much on how this question is answered.

5. The panelists should take the test (or a part) under testlike conditions. Familiarity with the test and its administration will enhance the validity of the resulting standards. This reduces the common problem of panelists underestimating the difficulty of the test and setting unreasonably high expectations for student performance.

6. The panelists should be thoroughly trained in the standard-setting process and be given practice exercises. The panelists' understanding of the process is critical to their confidence in the process and the acceptability of the standards that are produced.

7. It is often desirable to provide an opportunity for panelists to discuss their first set of ratings with each other prior to providing a final set of ratings. The second set of ratings will often be more informed and lead to more defensible standards because many sources of error due to misunderstandings, carelessness, inconsistencies, and mistakes can be removed. It has also become common to provide panelists with item statistics and passing rates associated with different performance standards so that they have a meaningful frame of reference for providing their ratings.

8. The full process of standard setting should be documented so that it is available if challenges to the performance standards arise. Every detail, from who determined the composition of the panel, to the choice of method, to the resolution of differences among the panelists, to the rationale for any adjustments made to the final performance standards, should be documented for possible use later.

Some New Advances in Performance Standard Setting

Standard setting has always been the Achilles' heel of educational testing. At the best of times there has been a concern for both the most suitable method for setting performance standards and evidence for their validity. Now, there is a new challenge for setting standards: performance assessments. In both educational testing and credentialing exams, performance assessments that require students to construct answers, write essays, or conduct science experiments, etc., are becoming more frequent. For example, the Kentucky Department of Education has moved to a total performance-based assessment system for school accountability. Most other states are using performance assessments in student accountability as well (Bond, Braskamp, & Roeber, 1996).

Performance assessments are often associated with complex and polytomous (i.e., more than two score points per task) scoring rubrics, multidimensionality in the response data (i.e., the tasks require multiple skills for successful completion), interdependencies in the scoring rubrics (sometimes, if students miss one part of a task, then they are unable to complete the remainder of the task because of the absence of a key piece of information), and low score generalizability at the task or exercise level (this means that students who perform well on one group of tasks cannot be assumed to be high performers on another set).

These features of performance assessments create special problems for standard-setting methods. For example, several of the popular standard-setting methods (Zieky, 1995) such as the Nedelsky and Angoff methods, are not even applicable with performance assessments that are polytomously scored. The challenge is to adapt old standard-setting methods or develop new methods to meet the current characteristics of performance assessments and that can meet existing standards of quality and defensibility.

This section of the chapter describes and comments on a number of standard-setting methods that can be applied to performance assessments. With several methods, follow-up references are provided. Although most of the discussion that follows applies to applications with a single performance standard (e.g., pass/fail decision point), the arguments are easily extended to setting multiple performance standards on an educational assessment (e.g., novice, apprentice, proficient, and advanced). Readers are encouraged to read Hambleton, Jaeger, Plake, and Mills (1998); Jaeger, Plake, and Hambleton (1993); and Mills, Plake, Jaeger, and Hambleton (1997) for discussions of additional standard-setting methods and issues specific to performance assessments.

Contrasting Groups

This method was described earlier and is one of the few methods in the literature that can be extended easily to performance assessments. Still, this method has some shortcomings in each context. Students are classified based on an external criterion (often teacher judgments) to performance categories. One problem with this method is that it is not always possible to classify students independent of a test and then obtain representative samples from each of the populations to derive performance standards. Without representative sampling, any resulting performance standard would be sample dependent and, therefore, of limited value. Additional concerns about the contrasting groups method are presented by Kane (1994).

One promising exception is the work of Clauser and Clyman (1994), who asked panelists to identify passing and failing students based on their holistic review of the student test booklets and without knowledge of the student test scores. The score distributions of these two groups of students were then used in deriving the performance standard (in this case, looking for the test score that optimally separates the students into the same classifications as those made by the panel). This method is limited, however, by its use of an internal criterion (i.e., students' overall test performance). On the positive side, the method is easily extendable to multiple performance standards.

Extended Angoff

Consistent with the traditional Angoff methodology, panelists estimate performance of borderline (or minimally competent) students. Panelists are trained to estimate the number of score points on performance exercises or tasks that likely would be obtained by borderline students. Additionally, under this variation of the Angoff method, panelists can set weights for exercises or tasks for the total assessment for use in computing the composite performance standard. Exercises or tasks judged as more important can be assigned higher weights in the setting of performance standards.

This method appears to have some promise with performance assessments (Hambleton & Plake, 1995). This method is popular with panelists and can lead to performance standards that they find acceptable and that are consistent over panelists. However, external validity evidence has not been compiled with this method to date. The method has the additional desirable feature of being compensatory. When there are multiple exercises or tasks composing an assessment (this is almost always the case), a compensatory approach is more desirable than a conjunctive approach in setting standards (Hambleton & Slater, 1997) due to the unreliability of individual exercise or task scores. A compensatory approach allows students to compensate for low performance on some exercises or tasks by achieving higher scores on other exercises or tasks. Only total score is considered in assigning students to performance categories in a compensatory approach to standard setting.

In order to use the Extended Angoff method, panelists need to be intimately familiar with the scoring protocols for the performance exercises or tasks in the assessment. In some recent standard-setting work with the National Board for Professional Teaching Standards, as many as two to three days were needed in order to familiarize panelists sufficiently with the scoring protocols for them to set performance standards. The standard-setting process itself may take an additional day or two (Hambleton & Plake, 1995). Therefore, with complex performance assessments consisting of multiple exercises or tasks, the amount of time needed to train panelists on the assessment tasks and scoring protocols should not be underestimated. In addition, considerable time needs to be allowed for panelist ratings and discussions. Exact times might be determined through carefully conducted field tests of the standard-setting process.

This standard-setting method too may have some potential when multiple forms of assessment (such as norm-referenced tests, criterion-referenced tests, quizzes, classroom work, and portfolios) are being combined into a single test score for judging student performance. Performance standards can be set on each component of the total test score, and weights for each component can be established, in arriving at final performance standards for interpreting the total test score. One useful suggestion in implementation is that before any standard-setting work is carried out, scores on each component should be converted to z-scores. In this way, each score component is placed on a common scale before any combining of scores is carried out.

Estimated Mean, Expected Score Distribution

This method has some similarities to the extended Angoff method. Here, panelists are required to estimate not only the minimum number of score points for borderline students (as in the extended Angoff method) but also the distribution of scores of the borderline students. The method was tried by the National Assessment Governing Board (NAGB) in its work to set performance standards on the NAEP (see, for example, Cooper-Loomis & Bourque, 1996).

One advantage of this method, in principle, is that additional relevant information about the performance of borderline students is extracted from panelists. Panelists, who expect the standard deviation of the score distribution for borderline students to be low, are indirectly also expressing considerable confidence in the placement of the performance standard. Higher standard deviations about the performance of borderline students likewise correspond to less confidence on the part of panelists about the proper location of the performance standard.

Student Paper Selection

When using the student paper selection approach, panelists are instructed to review a sample of student papers and identify student work that they believe is associated with borderline students (e.g., borderline advanced) who took the assessment (Hambleton & Plake, 1997; Jaeger & Mills, 1997). Normally, this method is applied to each exercise or task. After discussion among panel members, revised selections can be made. The average score associated with the student work identified as borderline is one way to arrive at the performance standard (other promising ways are described by Jaeger & Mills, 1997) for the exercise or task. The sum of performance standards set for the exercises or tasks in the total assessment provides the performance standard for the assessment. Of course, this process can be repeated for multiple performance standards. This method is being used, on an experimental basis, by the NAGB, some state departments of education, the National Board of Medical Examiners, and the Educational Commission for Foreign Medical Graduates.

One major advantage of this method is that panelists are required to look at student work on the assessment tasks. Often, panelists find this activity to be very interesting and more meaningful than simply looking at the exercises themselves and scoring rubrics. A major disadvantage is that the method can sometimes be very time-consuming and difficult to implement in practice. For example, when a student's work involves a product such as a videotape, report, or project, sorting through student work for examples of borderline work can be very tedious, if not totally impractical. Fortunately, often student responses to performance assessments can be captured in a test booklet or portfolio.

A related disadvantage is that the resulting performance standards may be based on a very small number of students if time does not permit the review of substantial numbers of student papers or sufficient numbers of borderline papers cannot be found. Still, the paper selection method likely deserves considerably more research and development because the face validity of the approach appears high and the approach seems practical in many assessment situations. Also, one of the identified shortcomings might be overcome by incorporating all of the papers reviewed into the standard-setting process, not just the borderline papers.

Holistic or Booklet

This method has some similarities with the student paper selection method. This is a new method suggested originally by the National Academy of Education in its review of the standard-setting work of NAGB and the American College Testing (ACT) (Shepard, Glaser, Linn, & Bohrnstedt, 1993). Basically, panelists are asked to consider the *complete* work (includes all exercises or tasks in the assessment) of a student and decide which student booklets represent those of borderline students (or masters and non-masters, or basic, proficient, and advanced students). This approach also has been suggested as an alternative to the Angoff method with multiple-choice items to counter the criticism that the Angoff method focused at the item level loses the overall impression of a student's performance (see Hambleton & Plake, 1997).

NAGB and ACT have been field testing this method with NAEP data in several subject areas and at grades 4, 8, and 12. It remains to be determined, however, how well the method will work in practice. Certainly the focus on student work seems desirable. The work by Jaeger and Mills (1997) is especially relevant here, as they have conducted several successful field tests of the method.

Dominant Profile

This method is a direct approach to standard setting. A panel, after becoming familiar with the purpose of the assessment and the scoring scheme, attempts to formulate a standard-setting policy such as the following:

101

A student passes the test if he or she (1) has an overall score of 18 on the seven-exercise assessment, (2) scores at least 3 (out of 4) on exercises B and C, the two exercises judged to be most important, and (3) has no scores of 1 on the exercises (a score of 1 indicates disappointing or totally inappropriate performance).

This method may begin with a consideration of which profiles of scores over the exercises are worthy of, for example, promotion to the next grade. Over a series of iterations, the panel tries to arrive at a consensus policy or set of rules for passing and failing students. No limits or restrictions are placed on the final result. It may be compensatory, conjunctive, or some combination of compensatory and conjunctive components.

A major advantage is that the dominant profile method is direct and involves extensive discussions among panelists. From our experiences, panelists find the discussions very helpful. They will sometimes express a lack of trust for methods that they cannot completely control. A major disadvantage is that a single policy for making pass-fail decisions may not emerge from the panel. For example, suppose the panel is fundamentally divided on the desirability of a conjunctive component in the policy (such as components 2 and 3 in the previous example). Unlike the performance standards set with other methods, it may not be possible to average policies to arrive at a group consensus. This method has been studied by Plake, Hambleton, and Jaeger (1997).

Another disadvantage of this approach is that a conjunctive policy can result that is based on unreliable exercise scores (see Hambleton & Slater, 1997). Panelists need to be cautioned about the undesirability of conjunctive standard-setting policies when there are several exercises in the assessment package and the associated levels of exercise reliability are not high.

Policy Capturing

This method involves having panelists consider hypothetical score profiles (across a set of exercises in a performance assessment) and classifying them according to their level of proficiency (e.g., outstanding, excellent, good, fair, poor). Then, a mathematical model (e.g., linear regression model) is fit to a panelist's ratings to determine his or her "latent standard-setting policy." It is latent because the panelist is not able to articulate the policy. A group policy (or standard-setting decision rule) can be obtained by a weighted average of the individual panelists' standard-setting policies. Successive iterations are often used to achieve greater panelist consistency in score profile ratings, and to move the group of panelists toward a consensus policy for making pass-fail decisions. Extensions of the method to multiple performance standards is straightforward.

A major advantage of this method is that a decision policy is assured. Potential disadvantages are that it may be difficult to find statistical models to fit individual panelists' ratings of the score profiles, the method requires the setting of a standard or standards on the dependent variable (i.e., the rating scale used in sorting score profiles), and the mathematical manipulations of the data make it difficult to explain to panelists how their ratings affect the overall decision rule or policy. Some researchers believe that panelists ought to completely understand the process used in arriving at the standard. This method has been under development by Richard Jaeger for several years and the results, to date, are encouraging (see, for example, Jaeger, 1995; Jaeger, Hambleton, & Plake, 1995).

Item Mapping Method

This method, which is quite new, presents panelists with a scale for reporting achievement and highlights the performance of students on the assessment material at different places on the reporting scale (Lewis, Mitzel, & Green, 1996). This is accomplished with "item characteristic curves" and item statistics from the measurement field of item response theory (Hambleton, Swaminathan, & Rogers, 1991). Obviously, students with more ability will be able to perform better on more of the assessment material than students with lower ability. The important question for panelists is to decide the level of performance expected of partially proficient, proficient, and

advanced students. This judgment is made easier by the ordering of assessment material by its level of difficulty over the reporting scale.

One of the unknowns in this method is the role that the approach to displaying the performance data on the reporting scale plays in the final determination of performance standards. For example, items might be identified on the reporting scale by the ability level at which a student has a 50 percent chance of success. In one variation, items might be identified on the reporting scale by the ability level at which a student has a 75 percent chance of success. There is a suspicion that this simple variation may impact considerably on performance standards, and therefore more research on this part of the standard-setting methodology is needed.

At this time, the Extended Angoff Method, the Paper Selection Method, and the Holistic or Booklet Method appear to have the most utility for school districts and states, since all three methods have appeared in the assessment literature and have established some credibility by being used by states and school districts. We expect, however, that within the next five years, there will be several other viable methods in the literature, including several that were introduced in this chapter.

Summary

Many researchers and policy makers are still not comfortable with current performance standard-setting methods. Criticisms center on both the logic of the methods and the ways in which the methods are being implemented. Clearly, there is a need for new ideas and more research. Both new methods and improved implementation of existing methods are needed. On the other hand, performance standards are being set on many educational assessments, and there are methods that appear to lead to defensible standards. Appendix 10.4 provides a set of questions that might be asked during the course of a standard-setting initiative to ensure that the process is not flawed.

One of the special complications of performance standard setting for Title I programs is that often multiple measures are used in arriving at estimates of student levels of accomplishment. These student measures might include both norm-referenced and criterion-referenced test results, state test results, classwork, teacher quiz results, portfolio assessments, and other measures of student performance. To arrive at performance standards for partially proficient, proficient, and advanced students, first, all of the measures of student performance will need to be placed on a common scale. Z-scores would be especially suitable. Next, a weight for each measure could be established for combining these measures into a single total score for students. These weights, set by policy makers, curriculum specialists, and teachers, would reflect the relative importance attached to each measure of student performance in a total score. Highly reliable and valid assessments such as those coming from a state criterion-referenced test would likely be given more weight, than, for example, the results from a number of classroom tests administered over the school year (unless, of course, the state criterion-referenced tests are questionable because of factors such as short time limits or the use of unfamiliar item formats). Finally, performance standards could be set on each measure (using methods and steps described in this chapter), and then these standards can be transformed to z-scores and weighted accordingly, in arriving at performance standards on the total test score scale.

Of course, this is only one approach. Another approach might be to consider not a total score by obtaining a weighted sum of student measures, but rather to consider a vector of scores for each student across the measures of interest. In this approach, judgments about partially proficient, proficient, and advanced student performance would be made by considering the full vector of score information. The Dominant Profile Method might be especially useful for setting performance standards with this approach. Regardless of the approach, standard-setting methods described in this chapter could be modified to fit the choice of model for considering the multiple sources of information about student performance.

The most controversial problem in educational assessment today concerns setting standards [illegible]e test score scale to separate students into performance categories. It is now recognized by [illegible]ers in the educational testing field that there are no true standards waiting to be discovered.

103

Rather, setting standards is ultimately a judgmental process that is best done by appropriate individuals who (1) are familiar with the test purpose and content and knowledgeable about the standard-setting method they will be expected to use, (2) have access to item performance and test score distribution data in the standard-setting process, and (3) understand the social and political context in which the tests will be used.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education,* 45, 4–9.

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

Bond, L., Braskamp, D., & Roeber, E. (1996). *The Status of State Student Assessment Programs in the United States*. Oakbrook, IL: North Central Regional Educational Laboratory.

Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice,* 15, 20–31.

Cizek, G. J. (1996b). Standard-setting guidelines. *Educational Measurement: Issues and Practice,* 15, 13–21.

Clauser, B. E., & Clyman, S. G. (1994). A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine,* 69(10), S42–S44.

Cooper-Loomis, S., & Bourque, M. L. (1996, April). Psychometric considerations of items for achievement levels setting. Paper presented at the annual meeting of the National Council of Measurement in Education, New York.

Ebel, R. L. (1972). *Essentials of Educational Measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement,* 15, 277–290.

Hambleton, R. K., & Bourque, M. L. (1991). *The Levels of Mathematics Achievement: Initial Performance Standards for the 1990 NAEP Mathematics Assessment* (Technical Report, Volume 3). Washington, DC: National Assessment Governing Board.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (1998). *Handbook on Setting Standards on Performance Assessments*. Washington, DC: Council of Chief State School Officers.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education,* 8, 41–56.

Hambleton, R. K., & Plake, B. S. (1997, March). An anchor-based procedure for setting standards on performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation & the Health Professions*, 6, 3–24.

Hambleton, R. K., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education,* 10(1), 19–38.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice,* 10, 3–6, 10.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education,* 8, 15–40.

Jaeger, R. M., Hambleton, R. K., & Plake, B. S. (1995, April). Eliciting configural performance standards through a sequenced application of complementary methods. Paper presented at the annual meeting of the American Educational Research Association and the National Council of Measurement in Education, San Francisco.

Jaeger, R. M., & Mills, C. (1997, March). A holistic procedure for setting performance standards on complex large-scale assessments. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Jaeger, R. M., Plake, B. S., & Hambleton, R. K. (1993). Integrating multi-dimensional performances and setting performance standards. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research,* 64, 425–462.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard-setting: A bookmark approach. Paper presented at the meeting of the Council of Chief State School Officers, Phoenix.

Linn, R. L., & Herman, J. L. (1997). *A Policymaker's Guide to Standards-Led Assessment.* Denver, CO: The Education Commission of the States.

Livingston, A., & Zieky, M. J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

Mills, C. N., & Jaeger, R. J. (1998, April). Creating descriptions of desired student achievement when setting performance standards. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Mills, C. N., Plake, B. S., Jaeger, R. M., & Hambleton, R. K. (1997, March). An evaluation of two promising procedures for establishing performance standards on complex performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 14, 3–19.

e, B. S. (1997). Criteria for evaluating the quality of a judgmental standard setting procedure: t information should be reported? Unpublished paper.

105

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement,* 57, 400–411.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice,* 10, 15–16, 22, 25-26.

Popham, W. J. (1978). *Criterion-Referenced Measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting Performance Standards for Achievement Tests.* Stanford, CA: National Academy of Education.

Taube, K. T. (1997). The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. *Evaluation & the Health Profession,* 20(4), 479–498.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard-setting. *Journal of Educational Measurement,* 19(4), 295–308.

Zieky, M. (1995, August). Aspects of standard-setting methodologies. Paper presented at the annual meeting of the American Psychological Association, New York.

Appendix 10.1

This is an edited version of a sample panelist evaluation form from the *Handbook on Setting Standards on Performance Assessments* by Hambleton, Jaeger, Plake, and Mills (1998).

Grade 8 Science Assessment
Standard-Setting Study
(October 9–10, 1997)

Evaluation Form

The purpose of this Evaluation Form is to secure your opinions about the standard-setting study. Your opinions will provide a basis for evaluating the training and standard-setting methods.

Please do not put your name on this Evaluation Form. We want your opinions to remain anonymous. Thank you for taking time to complete this Evaluation Form.

1. We would like your opinions concerning the level of success of various components of the standard-setting study. Place a "✔" in the column that reflects your opinion about the level of success of these various components of the standard-setting study:

Component	Not Successful	Partially Successful	Successful	Very Successful
a. Introduction to the Science Assessment	_____	_____	_____	_____
b. Introduction to the Science Test Booklet and Scoring	_____	_____	_____	_____
c. Review of the Four Performance Categories	_____	_____	_____	_____
d. Initial Training Activities	_____	_____	_____	_____
e. Practice Exercise	_____	_____	_____	_____
f. Group Discussions	_____	_____	_____	_____

2. In applying the Standard-Setting Method, it was necessary to use definitions of four levels of student performance: Below Basic, Basic, Proficient, Advanced.

   Please rate the definitions provided during the training for these performance levels in terms of *adequacy* for standard setting. Please CIRCLE *one* rating for *each* performance level.

	Adequacy of the Definition				
Performance Level	**Totally Inadequate**				**Totally Adequate**
Below Basic	1	2	3	4	5
Basic	1	2	3	4	5
Proficient	1	2	3	4	5
Advanced	1	2	3	4	5

3. How adequate was the training provided on the science test booklet and scoring to prepare you to classify the student test booklets? (Circle *one*)

   A. Totally Adequate
   B. Adequate
   C. Somewhat Adequate
   D. Totally Inadequate

4. How would you judge the *amount of time* spent on training on the science test booklet and scoring in preparing you to classify the student test booklets? (Circle *one*)

   A. About right
   B. Too little time
   C. Too much time

5. Indicate the importance of the following factors in your classifications of student performance. (Beside each factor, place a "✔" under the appropriate column.)

Factor	Not Important	Somewhat Important	Important	Very Important
a. The descriptions of Below Basic, Basic, Proficient, Advanced	_____	_____	_____	_____
b. Your perceptions of the difficulty of the Science Assessment material	_____	_____	_____	_____
c. Your perceptions of the quality of the student responses	_____	_____	_____	_____
d. Your own classroom experience	_____	_____	_____	_____
e. Your initial classification of student performance on each booklet section	_____	_____	_____	_____
f. Panel discussions	_____	_____	_____	_____
g. The initial classifications of other panelists	_____	_____	_____	_____

6. How would you judge the time allotted to do the *first* classifications of the student performance on each booklet section? (Circle *one*)

   A. About right
   B. Too little time
   C. Too much time

7. How would you judge the time allotted to *discuss* the *first* set of panelists' classifications? (Circle one)

   A. About right
   B. Too little time
   C. Too much time

8. What confidence do you have in the classification of students at the ADVANCED level? (Circle one)

   A. Very High
   B. High
   C. Medium
   D. Low

9. What confidence do you have in the classification of students at the PROFICIENT level? (Circle one)

   A. Very High
   B. High
   C. Medium
   D. Low

10. What confidence do you have in the classification of students at the BASIC level? (Circle one)

   A. Very High
   B. High
   C. Medium
   D. Low

11. What confidence do you have in the classification of students at the BELOW BASIC level? (Circle one)

   A. Very High
   B. High
   C. Medium
   D. Low

12. How confident are you that the *Standard-Setting Method* will produce a suitable set of standards for the performance levels: Basic, Proficient, Advanced? (Circle *one*)

   A. Very Confident
   B. Confident
   C. Somewhat Confident
   D. Not Confident at all

13. How would you judge the suitability of the facilities for our study? (Circle *one*)

   A. Highly Suitable
   B. Somewhat Suitable
   C. Not Suitable at all

Please answer the following questions about your classification of student performance.

14. What strategy did you use to assign students to performance categories?

15. Were there any specific problems or exercises that were *especially influential* in your assignment of students to performance categories? If so, which ones?

16. How did you consider the multiple-choice questions in making your classification decisions about student performance?

17. Please provide us with your suggestions for ways to improve the standard-setting method and this workshop:

Thank you very much for completing this Evaluation Form.

Appendix 10.2

A panelist form for applying the Ebel method.

Importance	Difficulty: Easy	Difficulty: Medium	Difficulty: Hard
Critical			
Important			
Less Important			

Appendix 10.3

A panelist rating form for setting three performance standards using the Angoff method (with two rounds of ratings).

Panelist Name:_____ Date:_____

Subject: _____

	Basic		Proficient		Advanced	
Test Item	**1**	**2**	**1**	**2**	**1**	**2**
1	___	___	___	___	___	___
2	___	___	___	___	___	___
3	___	___	___	___	___	___
4	___	___	___	___	___	___
5	___	___	___	___	___	___
6	___	___	___	___	___	___
7	___	___	___	___	___	___
8	___	___	___	___	___	___
9	___	___	___	___	___	___
10	___	___	___	___	___	___
Total	___	___	___	___	___	___

Appendix 10.4

A checklist for carrying out a standard-setting process.

Question	Answer
1. Has consideration been given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?	
2. Is the final panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on this particular educational assessment?	
3. Was the performance standard-setting method field tested in preparation for its use in the actual study?	
4. Is the chosen standard-setting method appropriate for panelists and the particular educational assessment?	
5. Will panelists be briefed on the purposes of the educational assessment and the uses of the test scores?	
6. Will panelists be administered the educational assessment or at least a portion of it?	
7. Will panelists be suitably trained on the method they will be using to set standards? For example, will they work through a practice exercise?	
8. Will the descriptions of the performance categories be clear to the extent that they can be used effectively by panelists in the standard-setting process?	
9. If an iteration process is being used, will the feedback to panelists be clear, understandable, and useful?	
10. Will the process itself be conducted smoothly? Are the rating forms easy to use? Are documents such as student booklets, tasks, items, etc. simply coded?	
11. Will panelists be given the opportunity to "ground" their ratings? (For example, will panelists be given normative data at the task level, or the full assessment level?)	
12. Will panelists be provided consequential data (or impact data) to use in their deliberations? Will the panelists be instructed on how to use the information?	
13. Will an evaluation of the process be carried out by the panelists at the end of the meeting?	
14. Will any additional evidence be compiled to support the validity of the resulting standards?	
15. Will the full standard-setting process be documented (from the early discussions of the composition of the panel to the compilation of validity evidence to support the performance standards)?	

Acknowledgments

Several people contributed to the development of this handbook. Appreciation goes to Phoebe Winter, Doris Redfield, and Edward Roeber for reviewing earlier drafts of the handbook; Julia Lara, Rebecca Kopriva, and Sharon Saez for providing information regarding how to include all students in the system of standards and assessments; the State Collaborative on Assessment and Student Standards Comprehensive Assessment Systems for IASA Title I/Goals 2000, Performance Standards Study Group; and representatives from states and other organizations who attended the June 1997 ad hoc committee meeting to frame the document and share their stories.

State Collaborative on Assessment and Student Standards Comprehensive Assessment Systems for IASA Title I/Goals 2000 Performance Standards Study Group* and the ad hoc committee on performance standards†

Adrienne Bailey, Council of the Great City Schools*†
Don Burger, Mid-Continental Regional Educational Lab†
Dale Carlson, California Department of Education*†
Roy J. Casey, Nevada Department of Education*
Michael Dalton, Oregon Department of Education†
Mary Beth Fracek, Iowa Department of Education*
David Frisbie, University of Iowa*
Ken Gentry, Kansas Department of Education†
Elaine Grainger, Vermont Department of Education*
James Grissom, California Department of Education*
Stuart Kahl, Advanced Systems in Measurement and Evaluation†
Barbara Kapinus, Council of Chief State School Officers†
John Poggio, University of Kansas†
Jessie Pollack, Maryland Department of Education†
Doris Redfield, Council of Chief State School Officers*
Grace Ross, U.S. Department of Education*†
Tom Stubits, Pennsylvania Department of Education*
Cheryl Tibbals, Council of Chief State School Officers*
Hugh Walkup, U.S. Department of Education*†
Jack Wills, PRC*
Phoebe Winter, Council of Chief State School Officers*†

115

TM029391

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").