
Considerations for the Alternate Assessment based on Modified Achievement Standards

Section II

Test Design: Understanding Content and Achievement Standards and Incorporating Appropriate Item Modifications

Chapter 5: Understanding the Content

**Chapter 6: Developing Items and Assembling Test Forms for the
Alternate Assessment Based on Modified
Achievement Standards (AA-MAS)**

**Chapter 7: Developing Modified Achievement Level Descriptions
and Setting Cut Scores**

The contents of this publication were developed under cooperative agreement S283B050019 with the U. S. Department of Education. However, the contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

For the full version of this white paper, visit:

http://nycomprehensivecenter.org/initiatives/inits_sp_assessment



University of the
State of New York
State Education
Department

SECTION II

TEST DESIGN: UNDERSTANDING CONTENT AND ACHIEVEMENT STANDARDS AND INCORPORATING APPROPRIATE ITEM MODIFICATIONS

This section moves the discussion from one about the students—who they are, how they learn, and how they should be instructed—to the assessment itself. Once we have a grasp on which students might be best served by an alternate assessment based on modified achievement standards, we need to determine how to take our understanding of the students and apply it to good test design. Of critical importance is to understand how to cover the same breadth and depth as a general assessment and yet make it less difficult. These modified achievement standards can be less rigorous, but what does that truly mean?

Chapter 5, by David Pugalee and Bob Rickelman, bridges us from the discussions of Section I to lay the foundation for good test design. It focuses on content standards and curriculum and describes how content standards are developed. Then it moves to the key issue of how to maintain the same content, only modifying the achievement standards. It ends with some suggestions on ways to enhance or revise items to provide scaffolding for students who may need additional supports in order to show what they know and can do. The authors point out that the scaffolds described only work if they are incorporated both in instruction and assessment.

Chapter 6, by Cathy Welch and Steve Dunbar, picks up where Chapter 5 leaves off, focusing on types of modifications that can be made to the general assessment to make it more appropriate for low-achieving students with disabilities. It also provides an overview of best item and test development practices and uses these considerations to frame the discussion of areas for modification. The authors then address the psychometric consequences of test modifications as they play out in the assembly of test forms and in the analysis of technical characteristics of items and test forms.

Then, in Chapter 7 by Marianne Perie, the focus turns to the modified achievement standards. Here, the issue of rigor and what standard students are measured against is addressed. This chapter focuses on the main components of achievement standards—numbers and names of levels, achievement level descriptors, and cut scores—and provides guidance on how to develop each component. The theory is brought back to match both the test design from Chapter 6 and student cognition, discussed in Chapter 4.

This section also benefited from several helpful reviews from the expert panel members who reviewed these chapters. In particular, comments from Howard Everson, Suzanne Lane, Brian Gong, and Claudia Flowers were especially insightful and helped inform the final chapters.

CHAPTER 5

UNDERSTANDING THE CONTENT

David K. Pugalee
Robert J. Rickelman

In order to understand the assessment process, whether discussing general assessments or alternate assessments, it is essential to have a good basic understanding of the content learning that is being assessed. The content domain, as explicated in state standards, must be the continued focus of assessment and the underlying force that drives instruction. For alternate assessments based on modified achievement standards (AA-MAS), students' work must align with the published state grade-level standards. But how are these standards developed? How do they link to the curriculum approved for use in schools? How do these standards come into play when developing IEP goals? In the previous chapter, Pellegrino described how cognition plays a major role in student learning and assessment. In Chapter 3, Karvonen discussed the IEP process in detail, and suggested that the "Opportunity to learn requires a curriculum that is well-aligned to state standards and assessment" (p. 51). This chapter will further these arguments by focusing on how the content standards reflect these content domains and provide a framework for both testing and instruction for students who meet the AA-MAS criteria, as discussed in Quenemoen (Chapter 2, this volume).

This chapter defines curriculum, explains the link to content standards, and describes how content standards are developed by states. It is important to understand the difference between *content standards* and *achievement standards*, so this differentiation will be made. Finally, issues surrounding links between the general content and modified assessments will be discussed, including examples related to mathematics and English/Language Arts. A discussion related to the effects of scaffolding on instruction and assessment will conclude this chapter.

What is the Curriculum?

At a very basic level, a curriculum is a set of planned instructional activities that are designed to allow students to document achievement of their knowledge and skills, including how these skills can be applied to real-life situations. The goal of a curriculum is to provide a comprehensive focus for instruction and learning within a school, a school system, and/or across a state. They also provide a scope and sequence of skills within and across grade levels. The curriculum generally drives the important factor of materials that will be used to implement the curriculum, and often there are several choices among state-approved materials developed by different sources within the state or at the national level. These materials are showcased in teacher's manuals and related materials, detailing the overarching philosophy and theory behind the development of the curriculum, and how it can be used across grade levels. These philosophy statements are generally written by an expert editorial team—often including university faculty who are experts in the field and by school personnel or state-level curriculum experts. In short, the curriculum is the glue that holds the pieces together, informing both teaching and learning, which should then link to the content standards assessed and subsequent interpretation of the assessment, which, in turn, should then drive instruction. In this continuous improvement loop (instruction – assessment – interpretation – instruction . . .), the curriculum lays out the scope and sequence of skills aligned to the content standards that will be taught and assessed in different subjects across different grade levels.

For example, a state may approve five different programs to be used to inform the mathematics curriculum across the K-12 grade levels. In order to purchase materials with state funding, a school district would have to choose from among the state-approved materials. It is common for states to have textbook adoption committees, made up of experts within the content field, who make decisions about the quality of the options and how well they align to the state curriculum goals and approved content standards.

A less common option is that a set of materials can be developed at a much smaller scale, to be used with smaller populations of students. For instance, in a grade level where a specific state history is the focus for a course, related materials will likely be developed by in-state experts, since the content would not appeal to anyone in other states. It is common for individuals to develop these materials around a curriculum that will guide their decision making at a local level.

In alternate assessments based on modified achievement standards, students must document that they can meet the state grade-level content standards. What this means is that they must be given the opportunity to learn and be assessed using the same curriculum as the general population of students who are not using modified assessments. In other words, they cannot be accountable for learning a different set of standards or for using a curriculum that is not available to the general population of students within that state. Not only must they have access to the general curriculum, but they must take part in a modified assessment that is aligned to grade-level content standards. So, a student in the ninth grade could not be assessed on standards established for fifth grade, even though that might be the grade level most representative of that student's observed skill.

How are Content Standards Developed?

Content standards are generally developed in one of two ways. In some subject areas, standards are established at the national level, and these subsequently drive the development of individual state standards. For instance, in mathematics, the National Council of Teachers of Mathematics (NCTM, 2000) has developed a set of standards for grades pre-kindergarten through 12th grade. According to the Executive Summary of the 2000 *Principles and Standards for School Mathematics* (available at http://www.nctm.org/uploadedFiles/Math_Standards/12752_exec_pssm.pdf), these standards were developed by a set of national content experts, with broad opportunities for input from

teachers and others, based on an extensive study of curriculum materials, state documents, and best practice research, to:

- Set forth a comprehensive and coherent set of learning goals for PK-12 math
- Serve as a resource for educators and policymakers
- Guide the development of curriculum frameworks, assessments, and instructional materials, and
- Stimulate ideas and ongoing conversations at the national, state, and local levels about how best to help students gain a deep understanding of important mathematics (p.1).

These standards are used extensively to guide state standards committees, which shape the NCTM standards to the needs of the specific state, including aligning them to state content and assessments required of all students. So, while there may be minor differences in the details of the state standards across states, the general standards themselves are very consistent across states.

The second way that content standards are developed is within states, when national-level standards have not been developed, or when national-level content standards are fairly generic, and perhaps not specifically aligned to grade levels. These standards are considered to be more generic guiding principles, and can be helpful in developing an overall philosophy of the goals for standards; but more specific, focused, and assessable statements must be crafted by state-level experts to make sense of the continuous improvement cycle mentioned earlier.

The Reading and English/Language Arts (ELA) content area provides an example of this type of standards development. In 1996, the National Council of Teachers of English (NCTE) and the International Reading Association (IRA) published *Standards for the English Language Arts*. This document was the result of five years of collaboration between these two professional organizations. However, unlike the NCTM standards which will be discussed in detail later in this chapter and which are broken down into specific content and process standards across

grade levels, these ELA standards are more generic. For instance, Standard One states that a goal for ELA instruction should be that,

Students read a wide range of print and nonprint texts to build an understanding of texts, of themselves, and of the cultures of the United States and the world; to acquire new information; to respond to the needs and demands of society and the workplace; and for personal fulfillment. Among these texts are fiction and nonfiction, classic and contemporary works (p.3).

It is up to each individual state to determine how this general recommendation is actualized within and across grade levels within the state. So these can be considered more like guiding principles for content standard development rather than actual grade-level content standards.

It is easy to see that, unlike the NCTM standards, these are not grade-band specific standards, but rather 12 general standards supported by research and classroom vignettes of what might happen in a classroom in which the standards were being implemented. Since these are more recommendations than standards, states generally must craft their own standards using committees made up of experts in the field who work in state departments of public instruction, colleges and universities, and public and private schools. These committees meet for several days to write the PK–12 state standards in each subject area, using (much like NCTM did on a national level) previous state standards and current state and national policy, along with scientific research in best practices, to put together a detailed list of standards for each grade level. These standards are then generally widely disseminated among stakeholders for additional input before being officially approved and implemented by the state department.

States vary greatly in how often these committees meet to update standards, whether they come from national or state sources. Many states update grade-level content standards every five years, but this cycle is often interrupted by national or state mandates and/or new federal laws, so that standards may be changed more than once every five years to adhere to

new mandates, or even less than every five years if major mandates are expected to be forthcoming and states want more guidance before proceeding with this tedious task.

Several important points must be made concerning the development of content standards, especially in subject areas for which national standards have not been developed. First, the published content standards, critically important to the assessment process, are assumed to be the “gold standard” within states—reference points of knowledge to which students must achieve. But often there is no clear, scientific methodology behind the development of these standards, as mentioned earlier. In fact, when one state was recently working on its modified achievement standards and developing learning maps for each content standard in each grade level, the teams working on developing the maps struggled to understand what some of the state content standards actually meant, and how they could be taught and assessed and interpreted across the general, modified, and alternate assessment systems. After much struggle, frustration, and consultation, the teams decided that some of the state standards were just poorly written, and some team members expressed a strong interest in being named to content standard writing committees in the future. Another point is that when these content standards are being developed, there is often no deliberate thought about how each content standard will be assessed, which was one problem that these teams discovered

In other words, the development of this “gold standard” is sometimes quite unscientific and can be heavily influenced by one or two strong committee members who may or may not espouse a certain ideology of what should be taught and learned within the state. So, while the process of developing content standards obviously has to take place, understanding both how these standards are developed and how they may be unduly influenced by individual committee members is important to keep in mind. Sending out drafts of standards to a broad variety of stakeholders can be helpful in terms of quality control, but it remains, by nature, an imperfect system.

As mentioned in the previous chapter (Pellegrino, chapter 4, this volume), learning progressions can provide guidance about how a typical skill will be developed on a theoretical level, but these learning progressions have limited usefulness, since some students (perhaps many) do not follow typical learning patterns. Some states use the term “learning maps”, which should not be confused with learning progressions, to offer suggested pathways in which a student can learn and be assessed across achievement standards, with the understanding that, just like a road map, these pathways are meant to offer guidance, but can (and often must) allow for deviation to account for individual differences. As mentioned by Pellegrino (chapter 4, this volume) this flexibility is especially important for students being assessed against alternate and modified achievement standards, since they may be more atypical than their peers in adhering to both theoretical and practical expectations for learning and developing expertise for both declarative and procedural knowledge.

In other words, there is little solid evidence that there is only one way in which all students acquire knowledge. There are, more likely, multiple pathways, and learning maps offer practical guidance into how these might develop, especially in terms of depth of knowledge of the standards. These would generally be developed for each grade-level standard, and appropriate assessment measures would need to be developed to allow students the opportunity to show performance across the levels. These maps also link to achievement standards, often by taking into account depth of knowledge, with the assumption that more depth can demonstrate a higher achievement level. The effectiveness of such maps in positively impacting student learning is dependent on a rich system of formative assessment processes, aligned to instruction that provides pictures of what students are able to do on multiple tasks related to the standard. Examples of learning or progress maps in ELA and mathematics can be seen in the following:

Published State Standard: Student will be able to understand simile and metaphor.

<u>Developing</u>	<u>Proficient</u>	<u>Target</u>	<u>Advanced</u>
Student will be able to correctly label figurative language and literal language given lists of statements.	Student will be able to correctly identify a simile and a metaphor embedded in a paragraph of text.	Student will be able to use an appropriate simile or a metaphor in their own written work.	Student will be able to use a simile or a metaphor in their own written work and will be able to discuss the relevant characteristics of the word(s) being compared.

Published State Standard: Patterns and Functions: Demonstrate and explain the difference between repeating patterns and growing patterns. (See Ban, Holt, & Kurizaki, 2008).

<u>Less Complex</u>	<u>More Complex</u>	<u>Proficient</u>
Student can describe a growing pattern by using objects, pictures and numbers. Student can use appropriate vocabulary to describe the growing pattern.	Student can describe repeating AND growing patterns by paying attention to how each element in the pattern relates to each other. Student can use appropriate vocabulary to explain/justify the growing pattern.	Student can describe repeating AND growing patterns. Student can use appropriate vocabulary to explain/justify the growing pattern. Student can use comparison/contrast and cause-effect language to describe similarity and differences among patterns.

Difference between Content Standards and Achievement Standards

The *Modified Academic Achievement Standards* document (U.S. Department of Education, 2007) defines academic content standards as “statements of the knowledge and skills that schools are expected to teach and students are expected to learn (p. 12-13).” These content standards are mandated for all students, regardless of ability, and are meant to drive instruction and assessments. These are the content standards discussed earlier, established at the national or state level by teams of experts and stakeholders. On the other hand, academic

achievement standards “are explicit definitions of how students are expected to demonstrate attainment of the knowledge and skills of the content standards (p. 13).” They further state that academic achievement standards must have the following elements:

- At least three achievement levels, which are labels that convey the degree of achievement in a given subject area (e.g., proficient, developing, not proficient, etc.)
- Achievement descriptors, which are descriptions of content based competencies associated with each of the achievement levels established (what students at each level know and can demonstrate), and
- Cut scores, which separate one level of achievement from another (how is a proficient student different from a developing student, etc.)

More will be said about establishing achievement standards in Chapter 7 (Perie, this volume). This differentiation is important, because within an AA-MAS system, ONLY the academic achievement standards may be modified, NOT the content standards. In order for assessments to provide meaningful information about students’ academic progress and promote accountability, there must be a clear alignment between the assessments and academic content standards. This process has been discussed in much more depth by Pellegrino (Chapter 4, this volume) with the discussion of the Assessment Triangle.

Barriers in Providing Access to the General Curriculum

Ideally, all students must have access to the general curriculum used in their school, and be able to be assessed with the same assessment that all students use, in such a way as to document learning through performance within a general assessment system. But, with the broad diversity of skills and language that are typical in most schools in every state, this ideal is very difficult to fully implement, which is why states are allowed alternate assessments for a limited population of students. In the not so distant past, students with significant disabilities were not allowed to attend school. When the IDEA was originally passed by Congress in 1975,

school-based placements were mandated, but these were generally done in segregated settings within the school. More recently, students with disabilities (including those with significant and mild disabilities) were able to be excluded from the general assessments used for other students in the school who did not have documented disabilities. However, in the most recent era of high-stakes assessments and *No Child Left Behind* legislation, school administrators can no longer have free reign to exclude certain students from being counted in the standardized assessment system. All students have to be included on Adequate Yearly Progress (AYP) reporting to the state and federal governments.

These changes are helpful in that all students count in terms of AYP. Schools are no longer able to exclude a student because of severe disabilities, and they must document that the content being taught and learned in the school setting has direct or indirect links to the general curriculum being studied by peers. In the past, some administrators thought of some students in terms of a “test score,” and tried to exclude any student thought to be detrimental in bringing down the schools overall achievement level, which may impact their ability to meet their AYP goals. Now schools must find ways to create meaningful links to the curriculum for students with even the most significant disabilities. But this requirement also created challenges, or barriers, that were brought to the fore after the new federal legislation was implemented.

One barrier is the way that many teachers are trained, both at the in-service and pre-service levels. In a nutshell, special education teachers (especially those working with the most severe students) are not trained (with rare exceptions) to think about the general state standards and required curriculum and are not shown how to link their teaching and student learning directly to those standards, since students with disabilities were traditionally expected to work on nonacademic skills, which are important in terms of day-to-day living. Many students, including those with more mild intellectual disabilities, were generally functioning below grade level. So much of the focus of instruction was spent on teaching and learning content standards

at lower grade levels, in addition to nonacademic content, such as life skills. On the other hand, general education teachers and teacher candidates, who were generally much better versed in understanding how their teaching was driven by the state standards expected within the different content subjects, had little or no idea how to apply that information to students with disabilities. In fact, it was common in most schools to find that general education and special education teachers rarely, if ever, interacted professionally within a school setting. This is not surprising, considering that these teachers are generally trained in segregated classes within teacher training programs. Even in graduate schools, special education teachers rarely take classes with general education peers. So the information necessary to successfully navigate the new federal mandates were generally not shared among the two sets of teachers.

Why is this important? General education and special education professionals must find ways to align their points of view, with general education teachers providing help in understanding mandated grade-level standards for learning and special education teachers bringing expertise in how to teach the necessary skills to achieve these standards to students with disabilities. There is a synergy when special and general education teachers work together that is not possible when they work in isolation. General education teachers are able to address the “What” questions surrounding student learning—what is meant by a state standard? What are higher level thinking skills? Process writing skills? What does it look like in a classroom setting? And special education teachers are able to better address the “How” questions—how do I teach a nonverbal student to understand algebraic principles? How can I teach a blind or deaf student to read? Both kinds of expertise are needed in order to ensure that all students have access to good teaching and learning. But still, in most teacher training schools and many PK-12 schools, these teachers continue to work alone, with rare exceptions.

So, how do we create professional development models that break these barriers? What processes need to be in place for this to happen? First, it is essential that teacher training institutions consider developing teacher candidates with dual expertise in both general

education and special education. Regardless of what subject each individual teacher is expected to teach, this dual expertise will be helpful, especially in a diverse society, where a “typical” classroom includes students with intellectual disabilities, learning differences, language differences, etc. Imagine programs where elementary and middle/secondary teacher candidates work alongside special education teacher candidate colleagues. Imagine graduate programs where candidates for general and special education degrees collaborate with colleagues in educational administration and school counseling preparation programs, working through common scenarios and learning how to cooperate in the world in which they will eventually be expected to collaborate. For schools, this could open a broad door to professional development opportunities, using case studies and real-life scenarios to promote school climates where all teachers understand both the “what” and the “how,” regardless of their assigned grade level or content expertise, and they are supervised and supported by professionals with similar training and expertise.

In addition, these partnerships between general and special educators must be fostered to allow for the design of meaningful assessments, including the tricky work of designing an assessment that is less difficult but that maintains the same levels of breadth and depth. General education teaching specialists should have the content knowledge related to the domain being assessed, to be able to ensure that depth and breadth are maintained, and special education teaching experts generally can address alternate (and perhaps less difficult) ways of allowing students to “show what they know.” In addition, these teams can help develop learning maps and/or curricula that utilize best practices in teaching, thereby allowing for alternate ways to teach students skills and processes related to content standards.

By working together and negotiating this fine line between content integrity and less difficult assessments, the synergy mentioned earlier can be maintained. And these experts must work with assessment experts, as well, with the general education teachers making judgments about fidelity of assessment items to the learning domain, and special education experts making

judgments about accessibility to testing for students who may need supports in order to show what they know.

The Content Standards for Mathematics

The majority of states have mathematics content standards that align to *Principles and Standards for School Mathematics*, the content standards document published by the National Council of Teachers of Mathematics (2000). This document presents key mathematics goals for students in pre-K through twelfth grade. The document describes ten standards, five content and five process standards, that represent a comprehensive and connected organization of key mathematical understandings and competencies of what students should know and be able to do. The content standards include number and operations, algebra, geometry, measurement, and data analysis and probability.

The five process standards underscoring ways of acquiring and using mathematics content knowledge include problem solving, reasoning and proof, communication, connections, and representation. For the five content standards, broad goals are presented for all students preK–12 with specific expectations explicated for the various grade bands: pre-K through grade 2, grades 3 through 5, grades 6 through 8, and grades 9 through 12. The following table presents the goals for each of the content standards for all students grades pre-K through 12.

Table 5-1. Goals for Pre-kindergarten through Grade 12 for Five Content Standards

Number and Operations	Algebra	Geometry	Measurement	Data Analysis and Probability
Understand numbers, ways of representing numbers, relationships among numbers, and number systems; Understand meanings of operations and how they relate to one another; Compute fluently and make reasonable	Understand patterns, relations, and functions; Represent and analyze mathematical situations and structures using algebraic symbols; Use mathematical models to represent and understand quantitative relationships;	Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships; Specify locations and describe spatial relationships using coordinate	Understand measurable attributes of objects and the units, systems, and processes of measurement; Apply appropriate techniques, tools, and formulas to determine measurements.	Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them; Select and use appropriate statistical methods to analyze data; Develop and evaluate inferences and predictions that are based on data;

estimates	Analyze change in various contexts.	geometry and other representational systems; Apply transformations and use symmetry to analyze mathematical situations; Use visualization, spatial reasoning, and geometric modeling to solve problems.		Understand and apply basic concepts of probability
-----------	-------------------------------------	---	--	--

Each of the five content standards is broken down by grade-level bands providing greater specificity as to what is expected of students at that particular grade level. This elaboration is provided for each of the goals listed in the above table. For example, the algebra standard includes “analyze change in various contexts” as one of the goals. This following presents the expectations at various grade bands for this goal.

Table 5-2. Grade Band Expectations for the Algebra Standard

Pre-K through Grade 2	Grade 3 through Grade 5	Grade 6 through Grade 8	Grade 9 through Grade 12
Describe qualitative change, such as a student’s growing taller; Describe quantitative change, such as a student’s growing two inches in one year.	Investigate how a change in one variable relates to a change in a second variable; Identify and describe situations with constant or varying rates of change and compare them.	Use graphs to analyze the nature of changes in quantities in linear relationships.	Approximate and interpret rates of change from graphical and numerical data.

Similarly, the process standards for mathematics provide general guidelines that assist in describing the types of mental processes that are inherent in a well-balanced mathematics curriculum. Though more difficult to specify in terms of concrete and measurable behaviors, the process standards present key ideas about what is valued in the discipline. The following table lists the goals for each of the five process standards for pre-kindergarten through grade 12. The *Principles and Standards for School Mathematics* (NCTM, 2000) does not further

delineate grade band expectations for the goals. Later in this section, the process standards will be revised in reference to their use in designing state assessments.

Table 5-3. Process Standards for Pre-K through Grade 12

Problem Solving	Reasoning and Proof	Communication	Connections	Representation
Build new mathematical knowledge through problem solving; Solve problems that arise in mathematics and in other contexts; Apply and adapt a variety of appropriate strategies to solve problems; Monitor and reflect on the process of mathematical problem solving.	Recognize reasoning and proof as fundamental aspects of mathematics; Make and investigate mathematical conjectures; Develop and evaluate mathematical arguments and proofs; Select and use various types of reasoning and methods of proof.	Organize and consolidate their mathematical thinking through communication; Communicate their mathematical thinking coherently and clearly to peers, teachers, and others; Analyze and evaluate the mathematical thinking and strategies of others; Use the language of mathematics to express mathematical ideas precisely.	Recognize and use connections among mathematical ideas; Understand how mathematical ideas interconnect and build on one another to produce a coherent whole; Recognize and apply mathematics in contexts outside of mathematics.	Create and use representations to organize, record, and communicate mathematical ideas; Select, apply, and translate among mathematical representations to solve problems; Use representations to model and interpret physical, social, and mathematical phenomena.

Recognizing that states and local educational agencies were often challenged in implementing rigorous assessment and accountability systems and to assist teachers in identifying consistent priorities and focus, the NCTM (2006) developed *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics*. In this document, NCTM asserts that:

...organizing a curriculum around these described focal points, with a clear emphasis on the processes that *Principles and Standards* addresses in the Process Standards—communication, reasoning, representation, connections, and, particularly, problem solving—can provide students with a connected, coherent, ever expanding body of mathematical knowledge and ways of thinking. (p.1)

It is clear that these documents offer a comprehensive picture of the domain of school mathematics. It is further evident that such documents provide the core for developing curriculum and informing instructional and assessment priorities.

Professional specialty organizations, such as the National Council of Teachers of Mathematics and various government sponsored enterprises including the National Mathematics Advisory Panel, provide an articulation of the content domain for mathematics. This articulation is a broad framework providing state educational agencies with a launching point from which to develop grade-level specific mathematics competencies. States use different processes to develop academic standards for grade-level content. The resulting documents become the critical focus as states develop assessments, including modified achievement standards, to assess students' proficiency towards meeting those state competencies.

Sampling Mathematics

Important mathematics that should be reflected in assessments includes “both the necessary content and the interconnectedness of topics and process” (Mathematical Sciences Education Board [MSEB], 1993, p. 42). The National Assessment of Educational Progress [NAEP] employs a new way to characterize the learning domain and the corresponding assessment that utilizes a lattice structure allowing a more interconnected view of mathematics. Since 1995, items reflect five content categories: number and operations; measurement; geometry; data analysis, probability, and statistics; and algebra and functions. Also included are mathematical abilities categories: conceptual understanding, procedural knowledge, and problem solving. These ability categories are considered in the final stage of development to confirm that there is balance among the three categories though not necessarily within each content category (MSEB, 1993).

New York, for example, has test blueprints for mathematics that assess a range of mathematics skills and abilities. The items are also aligned with one content-performance indicator for reporting purposes and are also aligned to one or more process-performance indicators (New York State Education Department, 2007a). The alignment to both content and process strands is intended to provide tests which “assess students’ conceptual understanding, procedural fluency, and problem-solving abilities, rather than solely addressing their knowledge of isolated skills and facts” (p. 5). New York includes five content strands: number sense and operations, algebra, geometry, measurement, and statistics and probability. The distribution of score points across the strands was determined during specifications meetings with panels of New York State educators during blueprint specifications meetings. The 2007 Blueprint, for example, indicates that at grade 5 for the content strand of algebra that the target number of points would be 5 (6 points were selected for the test) comprising 11% of the test (13% was the percentage of items selected for the test).

The Content Standards for English/Language Arts

As mentioned at the beginning of this chapter, there are no specific grade-level content standards in ELA at the national level, as there are in mathematics. So the manner of teaching, learning, and assessing ELA will, not surprisingly, also be different. Part of the issue was discussed earlier in Chapter 4 (Pellegrino). While it is fairly easy to observe and draw inferences about how a child might develop skills in and learn long division, it is much more difficult to make inferences about how a child might learn to comprehend information and, as also mentioned earlier, even these inferences from commonly used measures can be misleading. This is likely one reason why there is less consensus among experts about how ELA develops, than there is in mathematics. Rather, there are the general recommendations from the NCTE/IRA publication mentioned earlier, which are somewhat outdated but still sound.

It might be easy to assume that the ELA standards across different states are quite different, especially compared to areas like math where there are national guidelines for standard setting. But an examination of ELA standards by grade level and across different states reveals that they are actually comparable. Part of the reason for this general consensus relates to the extensive research that is available in ELA, outlined in more detail in Chapter 4. So there *does* tend to be much general consensus, framed by this research, about the standards that need to be taught in different grades, with an initial focus on “learning to read” being gradually replaced by an emphasis on “reading to learn.”

This distinction is important, and first surfaced in the early 1970s (Herber, 1970). In the first few years of school, especially at the preschool through second grade levels, much ELA instruction is focused on “learning to read,” which involves learning the requisite skills that lead to more complex reading skills, in phonics, phonemic awareness, etc., and also developing fluency practicing and using these skills on both narrative and informational texts. As students begin to move into third grade, in general, and even more dramatically as they move into the middle grades and then into high school, the emphasis shifts further and further away from these core basics, since the assumption is that they have been taught and learned in the earlier grades. The emphasis then shifts to “reading to learn.” This means that students become more and more accountable for using the earlier developed skills to read, in order to learn content information—mathematics, science, social studies, etc. While there is not a clean cut break between the two, and there is indeed much overlap as students work to learn to read and read to learn at the same time, as students progress to the higher grades, the skills expected to be learned become much more intangible, making them more difficult to assess. While it is fairly simple to teach and assess a student’s ability to put letters and sounds together to make words, it is much more difficult to teach and assess a student’s ability to utilize higher order thinking and critical comprehension skills. So the task of developing content standards, curricula and assessments tends to be much more straightforward in the earlier grades, where the skills can

be demonstrated in a much more concrete way, and much more difficult as students move into the higher grade levels, and documenting learning of the skills becomes much less of a concrete process (as Pellegrino discussed in Chapter 4, this volume).

Much attention has been paid to the report of the National Reading Panel (2000) (www.nationalreadingpanel.org). For instance, the findings of the panel were used in establishing the Reading First program, a \$5 billion dollar initiative introduced during the latest Bush administration. Many curriculum materials and school professional development activities are developed around the “Big Five” or the “Essential Five” skills highlighted in the report—phonemic awareness, phonics, fluency, vocabulary, and comprehension. Some advocates of the report suggest that these “big five” are the only skills that are “research based,” and so these are the only ones that should be a part of the state ELA standards.

Timothy Shanahan, a member of the National Reading Panel, tried to dispel myths surrounding the panel report (2003), discussing both what the report said and what it did not say. He stated that, in determining which areas to study in the report, “. . . we arrived at more than 30 topics that we thought might merit review—and even that list was not complete with regard to all topics that have been researched or that have been discussed as having potential importance (p. 649–650).” Some of these topics were not studied because the panel felt that they had been adequately reviewed elsewhere in the professional literature. Some were not studied because the panel felt that there was not enough evidence (previously published research) available to include it within the framework of the meta-analysis. Some specific studies were not included in the report, even though the topic of the study WAS included in the report, because the studies did not meet the criteria established by the panel for inclusion in the meta-analysis, not because, as some experts suggest, the research was not scientific in nature.

In terms of how the work of the National Reading Panel influences state standards in ELA, it is safe to say that the “Big Five” should certainly be included, since there is strong evidence that they are indeed essential for allowing students to access the curriculum in other

content subjects. However, it is not safe to assume that if a topic was not included in the report, there is no scientific evidence that it is worth learning.

Aligning State Standards to Assessments

Using curriculum documents, such as content maps, developed at the state level, assessments are designed to represent the content domain. Content match and depth match are two dimensions on which to consider how the assessments align to the curriculum (LaMarca, 2001). Content match has to do with the degree to which the assessment content matches to the subject area content, as identified in the state academic standards. Content match may be further delineated through analysis of broad content coverage, range of coverage, and balance of coverage. Broad content coverage, or the categorical congruence of the assessment, addresses whether the test content links to the broad content standards. Range of coverage asks whether the test items address the specific objectives related to each content standard. Balance of coverage is concerned with whether the assessment items reflect the major emphases and priorities found within the academic standards. The second dimension of alignment, depth match, is related to the degree to which the test items match the skills and knowledge specified in the state's academic standards in terms of cognitive complexity. Once items are developed, there should be a systematic analysis of the alignment that includes a determination of what objective an item measures and the degree of cognitive complexity for that item (LaMarca, 2001; Webb, 1999).

Guidance on Test Specifications

Ketterlin-Geller (2008) proposes a model of assessment development which extends the assessment model created by the National Research Council (Pellegrino, Chudowsky, & Glaser, 2001) in order to better meet the needs of students with cognitive disabilities. Ketterlin-Geller's model is motivated by the concept of *universal design* as applied to educational testing. Students with cognitive disabilities may not interact with a test in the same way as general

population students. This causes “construct-irrelevant variance” that prevents an accurate assessment of “domain-specific knowledge” (Ketterlin-Geller, 2008, p.4). Universally designed tests assess the same constructs but have flexibility in the format or delivery of the test, thus rendering them more useful and accessible to a greater percentage of the student population. Ketterlin-Geller (2008) argues that “applying the principles of universal design to academic assessments provides a mechanism for reducing the impact of construct-irrelevant variance on test-score interpretation, thereby, increasing the validity of the uses of test results” (p.4).

Assessment must consider the interaction between observation, cognition, and interpretation in the assessment design (Pellegrino, Chudowsky, & Glaser, 2001). Ketterlin-Geller (2008) elaborates on some basic ideas on how this model informs the design of assessment tools. In this model, within the assessment triangle detailed in Chapter 4, the cognition aspect represents the theories and beliefs of learning within the domain. Cognitive models should reflect the ways that children learn content within the targeted domain. Such targets include broad constructs such as analytic reasoning as well as narrow components such as the unit of length in mathematics. The observation aspect involves collecting student behaviors, which become the basis for interpretations about the cognitive targets. The features of assessments should reflect and align with the construct. Students with significant disabilities may interpret and respond to items as a result of their disability contributing to construct-irrelevant variance. Assessment features such as test platform, item format, problem context, administration procedures, and scoring systems should be considered when determining the characteristics of assessment tools. The interpretation aspect grounds decisions made about student skills and knowledge in the domain. Student characteristics, the cognitive model, and the observational tool interact in ways that influence the interpretation of student performance. Failure to consider these interactions may lead to problems with the validity of score-based interpretations.

Part of creating a universally designed test is to incorporate a cognitive task analysis for each test item (alongside the content of the domain targeted). Ketterlin-Geller proposed a cognitive task analysis along four levels of cognitive engagement: knowledge and application of general facts and procedures, knowledge and application of concepts and procedures, strategic thinking, and extended thinking. Furthermore, delineation between target and access skills should be clear. Target skills include both the cognitive and content components that the test is designed to actually measure. Access skills, on the other hand, include cognitive and content components that are needed to attain the target skills, but which the test is not designed to measure. Explicitly analyzing and articulating the cognitive tasks underlying a given problem can lead to better test accommodations for students with cognitive disabilities. For instance, a cognitive task analysis for a given mathematics problem may reveal that a test taker must be familiar with the concept of a calendar. If familiarity with a calendar is classified as an access skill, and the student has a limited concept of a calendar, an accommodation may be made (e.g. eliminate calendar reference or include an explanation of concept of calendar in problem).

As the assessment system is further developed, review of items for assessments should follow a structured protocol and should be reviewed by content and grade-level experts. Such review should be sensitive to the interaction between cognition, observation, and interpretation. Item review might include the following elements:

- accuracy and grade-level appropriateness
- mapping of the items to performance indicators
- accompanying exemplar responses (for constructed-response items)
- appropriateness of the correct response and distracters
- conciseness, preciseness, clarity, and readability
- existence of ethnic, gender, regional, or other possible bias. (NYSED, 2007a, p.16)

Such procedures are imperative, particularly for alternate assessments based on alternate and modified achievement standards, so that students have access to the range of academic content specified in the state's academic grade-level curriculum. Procedures minimize clustering of assessment items in isolated content strands and further guarantee that assessments are not over reliant on items that align to processes such as procedures in mathematics and decoding skills in reading.

It is essential that reviews consider the interpretation aspect of the assessment model. Messick (1989) puts forth the idea of a unified view of validity which takes into consideration the ethical underpinnings of the test interpretation and use. He argues that the way content validity has been defined, as "evidence in support of the domain relevance and the representativeness of content of the test instrument" (p.7) does not consider the inferences that may be made from the test. Messick argues that "we must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values" (p.8). Only through systematic and comprehensive analysis of the assessment program will all of the issues related to the assessment model (observation, cognition, and interpretation) be an integral part of the test design process.

Judging the Alignment between Expectations and Modified Assessments

Webb (1997) offers five categories for judging the alignment between expectations and assessments. The first, content focus, states that the focus should consistently be on developing students knowledge of content. This consistency is primary emphasized in four components: categorical concurrence, depth of knowledge consistency, range of knowledge correspondence, and balance of representation.

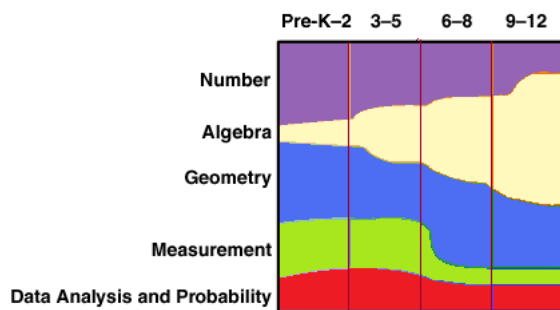
1. Categorical concurrence allows for differences in the level of detail but expects the same categories of content (such as content headings and their subheadings) to appear in the

expectations and the assessment. For example, an assessment in mathematics would need to reflect the five content strands from the NCTM.

2. Depth of knowledge consistency can vary on a number of dimensions, such as level of cognitive complexity, and describes how well students should be able to transfer the knowledge to different contexts and how much prerequisite knowledge is necessary in order to understand more difficult concepts. For example, the New York grade 5 core mathematics curriculum, for standard indicator 5.A.7, states that the student will “Create and explain patterns and algebraic relationships (e.g., 2, 4, 6, 8...) algebraically: $2n$ (doubling)” (New York State Education Department, 2006). If students are only required to identify the next item in the pattern, the depth of knowledge is not aligned for this performance indicator. The learning map related to figurative language presented earlier in this chapter provides another example of how the depth of knowledge can vary across content achievement standards.
3. Range of knowledge correspondence refers to the degree to which expectations and assessments cover comparable topics and ideas within categories. For example, the New York (New York State Education Department, 2007b) grade 4 performance indicators for ELA include the following: “Standard 2: Students will read, write, listen, and speak for literary response and expression: Make predictions, draw conclusions, and make inferences about events and characters.” Assessments that only focus on prediction would not meet the range of knowledge correspondence, since the remainder of the skills are left out. While it is tempting to create learning maps related to achievement standards that are additive (i.e., if student can make predictions, they are in the developing level, if they can make predictions AND draw conclusions, they are at the target level, etc.), this methodology does not adhere to the intent of the content standard, which requires knowledge of all of the skills mentioned.

4. Balance of representation means that similar emphasis is given to different content topics, instructional activities, and tasks. Assessments must reflect shifts in emphasis in content. The following visual from the National Council of Teachers of Mathematics (2000) emphasizes this shifting emphasis for the five content strands. Typically alternate assessments have focused on number and measurement even for students at the middle and secondary levels. Such emphasis would not be consistent with the shifting emphasis in content. Similarly, in language arts, the example stated previously relating “learning to read” and “reading to learn” document this shifting emphasis across grade levels.

Figure 5-1. Shift in Emphasis in Content Pre-K through Grade 12



The second category for determining alignment is articulation across grades and ages. Expectations and assessments should reflect views about how students develop and learn at different stages. This view includes ‘cognitive soundness as determined by best research and understanding’ and ‘cumulative growth in knowledge during students’ schooling’. The underlying structure of knowledge in content domains influences how instructional experiences for students should be organized. Specialty professional associations such as the National Council of Teachers of Mathematics, the National Council of Teachers of English, and the International Reading Association exist as organizations to support the articulation of this body of research and understanding.

The third component is equity and fairness. Assessments that align to this criterion provide every student with a reasonable opportunity to demonstrate their level of attainment

relative to what is expected. High expectations are reflected in all learning standards. Multiple forms of assessment provide a better alignment based on students' level of knowledge, culture, social background, and experiences.

The fourth category is pedagogical implications. Classroom practice is related to the learning of students. Review of assessments must consider the implications for classroom practice. Teachers might be asked to interpret expectations and assessments and consider how their classroom practices fit with their interpretations. Two critical elements to consider when taking pedagogical influences into account are the active engagement of students in learning and effective classroom practice including the use of technology, materials, and tools. Assessment is not a stand-alone component of educational practice. Curriculum, instruction, and assessment should be linked in a coherent and meaningful fashion. Further, assessment is an ongoing process that should inform instruction; therefore, effective practices align all three components so that student learning is promoted as a coherent whole.

The fifth category for determining alignment of expectations and assessments is system applicability. Programs must be realistic and manageable. Policy must be constructed so that it is applicable to teachers and administrators in their day-to-day efforts and not present an additional burden outside of what is considered "normal" school activities.

Considering Cognitive Complexity in Assessments

The Council of Chief State School Officers recognizes three models for evaluating the alignment between curricular expectations and assessments: Webb's alignment model, the Surveys of Enacted Curriculum model, and the Achieve model (Roach, Niebling, & Kurz, 2008). The Webb alignment model is the primary model that has been applied to alternate assessments and will be the focus of this discussion (see also Chapter 9 by Abedi); however, Surveys of Enacted Curriculum provides an additional framework for considering cognitive complexity and is also described in the following paragraphs.

The Surveys of Enacted Curriculum [SEC] includes a common language framework for examining the content and visual displays of alignment analysis (see Porter & Smithson, 2001). The common language framework provides general categories under which a series of topics is organized. For example, addition and subtraction of whole numbers would be under the larger category of “Operations”. Other topical content categories for K-12 include number sense/properties/relationships, measurement, consumer applications, basic algebra, advanced algebra, geometric concepts, advanced geometry, data displays, statistics, probability, analysis, trigonometry, special topics, functions, and instructional technology. Content areas for reading and language arts for K-12 include phonemic awareness, phonics, vocabulary, awareness of text and print features, fluency, comprehension, critical reading, author’s craft, writing processes, writing components, writing applications, language study, listening and viewing, and speaking and presenting. All content areas will not be present at every grade level. Comparing the content categories with levels of cognitive demand (see tables below) allow for a coarse-grain view of what students are expected to do with their knowledge of content. A fine-grained view breaks the content into more discrete descriptions. Algebra, for example, includes such components as absolute value, multi-step equations, factoring, etc.

Table 5-4. Surveys of Enacted Curriculum Cognitive Demand Categories for Mathematics

Memorize	Perform Procedures	Demonstrate Understanding	Conjecture, Generalize, Prove	Solve Non-routine Problems, Make Connections
Recite basic mathematical facts	Use numbers to count, order, denote	Communicate mathematical ideas	Determine the truth of a mathematical pattern or proposition	Apply and adapt a variety of appropriate strategies to solve non-routine problems
Recall mathematics terms and definitions	Do computational procedures or algorithms	Use representations to model mathematical ideas	Write formal or informal proofs	
Recall formulas and computational procedures	Follow procedures / Instructions	Explain findings and results from data analysis strategies	Recognize, generate or create patterns	Apply mathematics in contexts outside of mathematics
	Solve equations/ formulas/routine word problems	Develop/explain relationships between concepts	Find a mathematical rule to generate a pattern or number sequence	Analyze data, recognize patterns
	Organize or display data		Make and	Synthesize content and ideas from several sources
	Read or produce			

	graphs and tables Execute geometric constructions	Show or explain relationships between models, diagrams, and/or other representations	investigate mathematical conjectures Identify faulty arguments or misrepresentations of data Reason inductively or deductively	
--	--	--	--	--

Table 5-5. Surveys of Enacted Curriculum Cognitive Demand Categories for ELA/Reading

Memorize/Recall	Perform Procedures/ Explain	Generate/ Create/ Demonstrate	Analyze/ Investigate	Evaluate
Reproduce sounds or words Provide facts, terms, definitions, conventions Locate literal answers in text Identify relevant information Describe	Follow instructions Give examples Check consistency Summarize Identify purpose, main ideas, organizational patterns Gather information	Create / develop connections among text, self, world Recognize relationships Dramatize Order, group, outline, organize ideas Express new ideas (or express ideas newly) Develop reasonable alternatives Integrate with other topics and subjects	Categorize / schematize information Distinguish fact and opinion Compare and contrast Identify with another's point of view Make inferences, draw conclusions Predict probable consequences	Determine relevance, coherence, internal consistency, logic Assess adequacy, appropriateness, credibility Test conclusions, Hypotheses Synthesize content and ideas from several sources Generalize Critique

SEC involves raters, including individual teachers and an alignment panel of three or more content area specialists. Teachers complete surveys at the end of the year, rating level of coverage for topics and subtopics and the level of cognitive demand for tasks in each of the topic areas. The model provides useful descriptors of cognitive demand that can serve as a guide in considering the design of assessments.

Application of Webb's model requires members of a trained alignment panel, consisting of educators and curriculum experts, to:

1. Recognize and apply depth-of-knowledge (DOK) level rating for each objective in the state content standards.
2. Rate the DOK level for each assessment task.
3. Identify the objective(s) from the content standards to which the assessment item corresponds.

The central feature of this model is the Depth of Knowledge rating given to each assessment item. There are four depth-of knowledge levels: recall, skill/concept, strategic thinking, extended thinking. Once this task is completed, analysis of the ratings allow for computing descriptive statistics for each of the four criteria in Webb’s alignment model: categorical concurrence, range of knowledge, balance of representation, and depth of knowledge which were described earlier in this chapter.

Table 5-6. Webb’s General Descriptions for Depth-of-Knowledge Level

Level	Description
Level 1: Recall	Recalling information such as facts, definitions, terms, or simple procedures; performing simple algorithms or applying formulas
Level 2: Skill/Concept	Requires some decision as to how to approach a problem or activity; classifying, organizing, estimating, making observations, collecting and displaying data, comparing data
Level 3: Strategic Thinking	Requires reasoning, planning, using evidence, and a higher level of thinking than recall or skill/concept; Explaining one’s thinking, making conjectures, determining solutions to a problem with multiple correct outcomes
Level 4: Extended Thinking	Requires complex reasoning, planning, developing, and thinking often over an extended period of time. Cognitive demand for tasks is high and work is complex. Requires making connections within and between subject domains. Includes designing and conducting experiments, making connections between a finding or outcome and related concepts, combining and synthesizing ideas into new concepts, critiquing literary pieces and designs of experiments.

These models provide useful descriptors for developing modified achievement standards and alternative assessments based on those standards. The descriptors can also guide

assessment development and ensure the assessments cover the same breadth and depth of content.

Levels of Cognitive Complexity in Mathematics

State assessment frameworks articulate performance indicators listed for content strands and are intended to provide teachers with guidance in determining the outcomes of instruction. The following discussion illustrates how state standards can be addressed through items with different depth of knowledge levels with the items still directly related to the content standard.

The New York Mathematics, Science, and Technology: Standard 3 states “*Students will:*

- understand the concepts of and become proficient with the skills of mathematics;
- communicate and reason mathematically;
- become problem solvers by using appropriate tools and strategies;

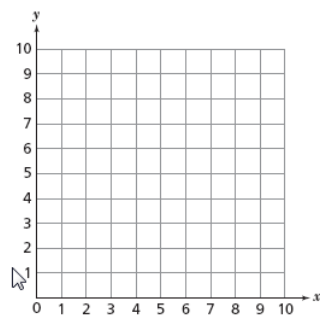
through the integrated study of number sense and operations, algebra, geometry, measurement, and statistics and probability” (NYSED, 2005).

The NYSTP Mathematics Tests is designed to assess students on the content and process strands of this standard. Items are aligned to one content-performance indicators, but is also aligned to one or more process-performance indicators as appropriate for the concepts that are embodied in the task (NYSED, 2007a) though the procedure used for determining alignment to the process performance indicators is not described in the technical documents.

An Illustration Based on Standard 3 for Grade 5 Mathematics. Consider an example based on the NY Standard 3 for grade 5 mathematics. The Geometry Strand includes the goal that “Students will apply coordinate geometry to analyze problem solving situations” and more specifically to “plot points to form basic geometric shapes (identify and classify)” which is indicator 5.G13 (NYSED, 2005).

The guiding principle is that assessment tasks must be aligned to content for the grade-level standards. This indicator requires that students demonstrate skills and understanding of coordinate geometry by plotting points in the context of identifying and classifying basic geometric shapes. The indicator contains multiple content-related targets suggesting that a modified standard could be constructed by breaking relevant tasks into multiple components. Depth of knowledge can be considered by changing the complexity of assessment tasks related to the indicator. How specific assessment items contribute to a student's level of proficiency is discussed later in this chapter.

The following illustration would include the appropriate information displayed in a coordinate grid such as the one that follows:



Depth of knowledge illustrations are presented for each of the four levels offered by Webb.

1. *Recall and Reproduction.* Present student with the following points graphed on a coordinate plane: A (1, 5), B (3, 2), C (6, 2), and D (? , ?). Students are asked to give the coordinates for point D. Next, present a similar diagram with points A (1, 1), B (1, 5), C (5, 5), and D (5, 1) connected to form a quadrilateral. The student is required to identify the type of quadrilateral formed by connecting the points. Students might be given possible choices such as square, rectangle, and trapezoid.

2. *Skills and Concepts/Basic Reasoning.* Presented with a coordinate grid, students are asked to plot points A (1, 1), B (1, 5), C (5, 5), and D (5, 1). They are then asked to connect A, B, C, and D in order and to identify the quadrilateral which is formed. Students might be asked

to explain or describe how they determined the type of quadrilateral that was formed. The assessment item might include scaffold to guide the students in this process such as “Describe the characteristics of the sides and angles that helped you decide what type of figure was formed”.

3. *Strategic Thinking/Complex Reasoning*. The student is asked to plot points A (1, 1), B (1, 5), and C (5, 5). They are then asked to plot point D such that the figure formed by connecting the points A, B, C, and D, in order, forms a rectangle. Name the coordinates for point D. Give two reasons why the figure has to be a rectangle.

4. *Extended Thinking/Reasoning*. The student is asked to plot point A (1, 1). Students are then asked to plot three additional points and connect them such that the figure formed is a rectangle. To extend their thinking, the student is asked to describe a process for forming a rectangle in a coordinate grid given one point as a vertex. Instead of a rectangle, the student might be asked to discuss the process for constructing a trapezoid given one point as a vertex.

Item Modifications

Additional modifications can be accomplished by changing the format of the assessment items, reducing the complexity of the language used in the item, and providing additional information or scaffolding to reduce the cognitive load for the student; however, the items must maintain alignment to the grade-level content in the standard, as discussed earlier. Hess, McDivitt, and Fincher (2008) conducted a pilot study about the effects of providing scaffolds for students within test items and across state assessments, to see if these scaffolds allowed students to better document knowledge that they possessed related to content standards. Some of the scaffolds that they studied included restricting the use of pronouns, using graphic organizers, chunking or segmenting longer texts into shorter pieces, left justifying text, shortening or simplifying test item stems, adding graphics to illustrate a term, and paying attention to the physical presentation of the assessment material by examining typeface,

spacing on the page, line length, and the use of blank space (or leading) around paragraphs or between columns of numbers to make them more legible. A goal of these modifications was to allow students access to the information on the assessment without cuing them to correct answers. These methods, if effective, could meet the AA-MAS guidelines of “less difficult” but adhering to the fidelity of the grade-level standard. In this study, teachers were also asked to use the scaffolding supports in their lessons, so that students were used to seeing them before the actual assessment. The results of this pilot study indicate that providing scaffolding that supports both teaching and assessments could provide a valid way to assess students on the AA-MAS test. While the scaffolds discussed are research based, there are some inconsistent findings about their effectiveness in improving student performance. For instance, Abedi et al. (2008) found that students with disabilities did not perform significantly better on reading comprehension assessments that utilized segmented text, although the reliability of the tests improved. Further studies and appropriate field testing are necessary to justify the choice and use of proper scaffolds at the state level.

Some more specific examples of scaffolding are shown in the following:

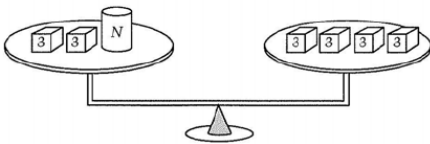
Common Stimulus. The approach of a common stimulus has been used on the National Assessment of Educational Progress (NAEP; Kenney, 2000). The common stimulus might be a table, graph, or chart. A series of items draws from previously presented information or a common context. Anderson and Morgan (2008) offer some guidelines for constructing items that have a common stimulus:

- Items should be independent. A student’s response on one question should not be dependent on getting the correct answer for a previous item.
- Items should refer to a clearly different aspect of the stimulus to avoid overlap.
- Items should assess a range of skills.
- Items should have a range of difficulty with the easier items appearing first.

- Information given in a stem or answer choice should not assist the student in correctly answering another item.
- Items should appear on the same page or on a facing page.

Such simple approaches reduce the cognitive load required to comprehend and process multiple items presented with varying contexts or information.

Replace text with relevant pictures, diagrams, tables, graphics. In the following diagram, the student is presented with a diagram that may assist them in visualizing the relationship between the sides of the scale and representing that relationship symbolically. This NAEP item received mostly exemplary comments from raters because of the scaffolding provided by the visual; however, one rater argued that the item was inauthentic and imposed as a testing convention since the item could be solved by visual inspection and did not require the construction of a number sentence (Daro, Stancavage, Ortega, DeStefano, & Linn, R., 2007). Despite this criticism, the item demonstrates how visuals might assist a student in understanding the relationship and thus able to focus on the task of identifying an equivalent symbolic representation. The diagram did not change the underlying skill of being able to identify a relationship symbolically.



The weights on the scale above are balanced. Each cube weighs 3 pounds. The cylinder weighs N pounds. Which number sentence best describes this situation?

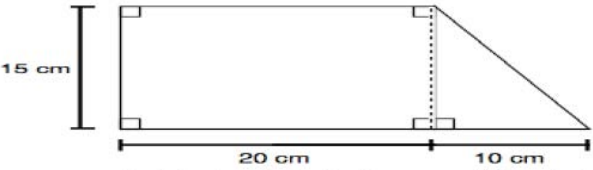
A $6 + N = 12$
 B $6 + N = 4$
 C $2 + N = 12$
 D $2 + N = 4$

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 4, Block B1M7, #4, 2007.

Reduce complexity of stem. The following question and its modification (Elliott, Kurz, Beddow & Frey, 2009) illustrates how items might be modified to reduce the complexity of the

item stem while preserving the alignment to the content standard. The modified item requires the student to evaluate a formula to find the area of a complex figure consisting of a rectangle and a triangle. The revised format removes the requirement that the student either recalls the appropriate formula or identifies it from the list provided in the booklet. Both items assess a student's ability to evaluate a geometric formula using data from a figure. Both items also require the student to differentiate between a rectangle and a triangle and to understand the basic concept of area. The modified item removes extraneous information, i.e. adjacent, while also maintaining depth of knowledge.

Original

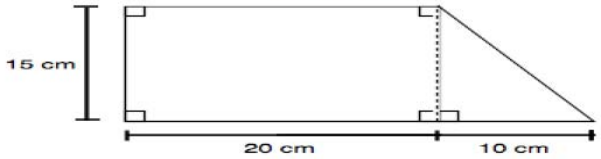


Reminder: Formulas for plane figures are available on the test reference sheet on the back of this booklet.

3. The figure above consists of one rectangle that is adjacent to one triangle. What is its area?

A. 375 cm²
 B. 450 cm²
 C. 83.03 cm²
 D. 600 cm²

Modified



Area of a rectangle: $A = l \times w$
 Area of a triangle: $A = \frac{b \times h}{2}$

3 What is the area of the figure?

A. 375 cm²
 B. 450 cm²
 C. 600 cm²

Hess, McDivitt, and Fincher (2008) provide a similar example of simplifying the stem and making the distractors complete sentences, in the following example.

<u>Original Stem</u>	<u>Modified Stem and Distractors</u>
The United States eventually reduced the number of immigrants allowed to enter the country because:	Why did the United States reduce the number of immigrants?
<ul style="list-style-type: none"> A. the United States already had too many people. B. the immigrants were taking away jobs from American workers. C. the immigrants had too many hardships to face in America. D. the country that the immigrants came from was angry about their leaving. 	<ul style="list-style-type: none"> A. The United States already had too many people. B. The immigrants were taking away jobs from American workers. C. The immigrants had too many hardships to face in America. D. The country that the immigrants came from was angry about their leaving.

Grouping questions by content or topic. Clustering items according to the particular standard that they link back to or by specific learning targets or objects is another way of modifying the assessment without changing the content focus of the item. For example if an assessment in mathematics has four items that link to the state standard “Apply ratios and proportions in solving real-world problems” then clustering those items (preferably by level of difficulty with easier items first) will facilitate the student being focused on specific content for a longer period of time without having to adjust to changes in content from item to item. The student will be working with similar thinking processes as proportional reasoning is applied to situations such as rates of change, percentages, unit pricing or rates, etc. Such measures reduce anxiety and may generate more interest thus improving concentration.

Asking understanding questions. Another set of modifications involves breaking complex tasks into components which may contain hints or supports to assist the test taker. A longer task may be broken into parts that are matched to a specific indicator or expectation (Suurtamm, Lawson, & Koch, 2008). Suurtamm and her colleagues do warn, however, that such modifications potentially lead to specific approaches to solving a problem and may diminish students’ opportunities to participate in complex problem solving. In modified assessments such practices are likely to reduce the overall complexity of the problem-solving situation while retaining a link to the content standard.

Another aspect of this type of modification is providing hints. For example if a student is asked to define a compound word (“noninterference,” for example), they might be prompted to “Break the word into parts.” Hess, McDivitt, and Fincher (2008) show how thought balloons, similar to what you might see in a comic strip, might be used to provide these hints.

These illustrations demonstrate that item modifications can be made while preserving the fidelity of the content. Such modifications reduce cognitive load and simplify language features that sometimes obscure the intent of the assessment item. Simplifying language features is important in making assessment items accessible to a larger population of students including those with learning difficulties and those for whom English is not their native language (also see the chapter by Abedi). Scaffolding and related practices are good instructional tools and should not only be used during assessments. Remember that there should be a coherent link between the curriculum, instruction, and assessment.

How Do We Link Content to Curriculum and Instruction Appropriate for this Population?

Curriculum access, data collection, and instructional effectiveness are issues that have been identified as variables that potentially influence student outcomes (Spooner, Dymond, Smith, and Kennedy, 2006). As emphasized throughout this chapter, linking assessment to content standards increases the likelihood that students with learning difficulties will have access to relevant grade-level academic content. The importance of curriculum access has been the focus of this chapter. Continued monitoring through data collection and analysis of student performance will provide greater alignment between instruction and assessment outcomes. Linking the curriculum and instruction to assessment outcomes is crucial in focusing the instructional design system on planning, implementing, and assessment of student learning.

The teacher is a critical factor in linking curriculum and instruction. Browder, Karvonen, Davis, Fallin, and Courtade-Little (2005) found that when teachers are trained on sound instructional practices, students’ scores on alternate assessments improved. Fuchs et al. (2008)

identified seven instructional principles that promote mathematical learning for students with disabilities. These instructional principles also provide an important set of guidelines for application to other content domains. First, instructional explicitness refers to instruction in which the teacher provides explicit and didactic teaching sharing information focused on the goals of instruction. The authors report that a meta-analysis of 58 math studies show that while developing students advance from programs with constructivist and inductive styles that students with mathematics difficulties do not profit in meaningful ways . Second, instructional design to minimize the learning challenge anticipates and eliminates misunderstandings with precise details, and the utilization of intentionally sequenced and integrated instructions focused on addressing gaps in achievement. The use of learning tools such as manipulatives and visuals enhance mathematics instruction, reducing confusion and the inability to maintain content. Third, a strong conceptual basis situates the procedures being taught in order to provide a strong conceptual foundation. Fourth, drill and practice is critical to maintaining skills through daily lessons, review, and computerized supports. Fifth, cumulative review reinforces practice and review, building a continued reliance on foundational skills being taught. Sixth, instruction must include motivators to help students regulate their attention and behavior and to work hard integrating systematic self-regulation and motivation supports including tangible reinforcers. The seventh principle, considered the most essential, is ongoing progress monitoring to establish whether a treatment is effective for a particular student.

In the next chapter, these ideas regarding modifying the content will be actualized into test design theory. However, it is important to remember that the best test design will not produce the desired results if the understandings about human cognition applied to the item development are not also carried into the classroom through curriculum and instruction.

References

- Abedi, J., Kao, J.C., Leon, S., Sullivan, L., Herman, J.L., Pope, R., Nambiar, V., & Mastergeorge, A.M. (2008). Exploring factors that affect the accessibility of reading comprehension assessments for students with disabilities: A study of segmented text. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. [on-line] Available: <http://www.cse.ucla.edu/products/reports/R746.pdf>
- Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement*. Washington, DC: The World Bank.
- Ban, P., Holt, L., & Kurizaki, V. (2008). Hawaii progress maps. Presentation made at the Council of Chief State School Officers Conference, National Conference on Student Assessment Resources, Orlando, FL.
- Browder, D.M., Karvonen, M., Davis, S., Fallin, K., & Courtade-Little, G. (2005). The impact of teacher training on state alternate assessment scores. *Exceptional Children*, 71, 267-282.
- Browder, D.M., Spooner, F., Wakeman, S., Trela, K., & Baker, J. (2006). Aligning instruction with academic content standards: Finding the link. *Research & Practice for Persons with Severe Disabilities*, 31(4), 309-321.
- Department of Education (April, 2007). *Modified academic achievement standards*. Washington, DC: Authors.
- Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP Mathematics Assessment: Grades 4 and 8*. Washington, DC: NAEP Validity Studies Panel, U.S. Department of Education.
- Elliott, S. N., Kurz, A., Beddow, P., & Frey, J. (2009). Cognitive load theory and universal design principles: Applications to test item development. Paper presented at the annual meeting of the National Association of School Psychologists, Boston, MA.
- Fuchs, L.S., Fuchs, D., Powell, S.R., Seethaler, P.M., Cirino, P.T., & Fletcher, J.M. (2008). Intensive intervention for students with mathematics disabilities: Seven principles of effective practice. *Learning Disability Quarterly*, 31(2): 79–92.
- Herber, H.L. (1970). *Teaching reading in the content areas*. Englewood Cliffs, NJ: Prentice-Hall.
- Hess, K., McDivitt, P., & Fincher, M. (2008). Who are the 2% and how do we design test items and assessments that provide greater access to them? Results from a pilot study with Georgia students. [online] available at: http://www.nciea.org/publications/CCSSO_KHPMMF08.pdf
- International Reading Association and National Council of Teachers of English. (1996). *Standards for the English Language Arts*. Newark, DE and Urbana, IL: Authors.
- Kenney, P. A. (2000). Families of items in the NAEP mathematics assessment. In N. S. Raju, J. W. Pelligrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 5 - 43). Washington, DC: National Academy Press.

- Kenney, P.A. (2000). *Market basket reporting for NAEP: A content perspective*. Paper presented at the March workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.
- Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, 27(3), 3-16.
- LaMarca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21). Retrieved March 20, 2009 from <http://PAREonline.net/getvn.asp?v=7&n=21>.
- Mathematical Sciences Education Board. (1993). *Measuring what counts: A conceptual guide for mathematics assessment*. Washington, D.C.: National Academy Press.
- Messick, S. (1989). Meaning and values in test validation: The sciences and ethics of assessment, *Educational Researcher*, 18(2), 5-11.
- National Center for Educational Statistics. (2003). *Mathematics Framework for the 2003 National Assessment of Educational Progress*. Washington, DC: Author.
- National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Reading Panel. (2000). *Report of the National Reading Panel, teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute for Literacy.
- New York State Education Department. (2006). *Mathematics Core Curriculum: MST Standard 3*. New York: The University of the State of New York & New York State Education Department.
- New York State Education Department. (2007a). *New York State Testing Program 2007: Mathematics, Grades 3-8. Technical Report*. Monterey, CA: CTB/McGraw-Hill.
- New York State Education Department. (2007b). *New York State Testing Program 2007: English Language Arts, Grades 3-8. Technical Report*. Monterey, CA: CTB/McGraw-Hill.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Report Series RR-048). University of Pennsylvania: Consortium for Policy Research in Education.

- Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in Schools, 45*(2), 158-175.
- Shanahan, T. (2003). Research-based reading instruction: Myths about the National Reading Panel report. *The Reading Teacher, 56*, 646-655.
- Spooner, F., Dymond, S. K., Smith, A., & Kennedy, C. H. (2006). What we know and need to know about accessing the general curriculum for students with significant cognitive disabilities. *Research and Practice for Persons with Severe Disabilities, 31*, 277-283.
- Suurtamm, C., Lawson, A., & Koch, M. (2008). The challenge of maintaining the integrity of reform mathematics in large-scale assessment. *Studies in Educational Evaluation 34*, 31-43.
- Webb, N.L. (1997). Determining alignment of expectations and assessments in mathematics and science education. *NISE Briefs, 1*(2), 1-11.
- Webb, N. L. (1999). *Alignment of Science and Mathematics standards and assessments in four states*. Washington, DC: Council of Chief State School Officers.

CHAPTER 6

DEVELOPING ITEMS AND ASSEMBLING TEST FORMS FOR THE ALTERNATE ASSESSMENT BASED ON MODIFIED ACHIEVEMENT STANDARDS (AA-MAS)

*Catherine Welch
Stephen Dunbar*

In many respects, the development of items and assembly of test forms for specific populations involves no special process considerations other than those required of any professional test development activity. This chapter will begin with an overview of best practices in item and test development in K-12 achievement testing in the context of the content domains of English Language Arts (ELA) and mathematics. Although the processes involved may be similar, the specific accountability context established by the AA-MAS guidelines, the potentially diverse characteristics of students in the AA-MAS population, and the fact that states are approaching AA-MAS designs in the presence of existing accountability tests developed under federal guidelines for technical quality means that certain steps in the test development process may deviate from typical best practice.

Nothing in the federal guidelines for the AA-MAS program specifies that the design and development of the two-percent assessment be approached as a modification of an existing general assessment or as an alternate assessment developed as a separate endeavor (U.S. Department of Education, 2007). Given the position of the AA-MAS assessment in a difficult to define gray zone between two existing assessments in each state, one can imagine its design and development to follow an approach already established by a state (or by its contractors) for any existing assessment, including an alternate assessment based on grade-level achievement standards (e.g. Massachusetts). Alternatively, an AA-MAS assessment might be developed as a kind of hybrid, consisting of features and materials from the general assessment and, where required by considerations of accessibility for example, measurement approaches adopted in the state's AA-AAS program. This chapter will discuss test development processes in general

that apply to whatever approach is taken by a state. Professional standards for test development and for the assessment of students with disabilities (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) provide guidance regardless of the approach. Much of the discussion in this chapter, however, presumes the AA-MAS approach taken by most states is to modify the existing general assessment. In most settings, the general assessment (its essential measurement features, its alignment to state content standards, its methods of scaling and reporting student achievement and mapping onto achievement levels) establishes the technical standards to be evaluated for purposes of reliability, validity, and comparability in the federal peer review process. Modification of the general assessment is likely to be cost effective as well. Thus, the argument to support inferences from the AA-MAS for purposes of accountability is likely to be structured with a state's general assessment in mind. This provides states with a logical starting point for developing the many justifications for resource allocation that set the foundation for validity and comparability arguments (Kane, 2006; Marion, 2009; Abedi, Chapter 8, this volume) as articulated in federal guidelines.

This chapter begins with a discussion of best practice in test development. The purpose of this discussion is to clarify key processes in development that contribute to the technical qualities of any assessment. Specific aspects of the process prominent in K-12 achievement testing for NCLB are noted as they represent key procedural steps that may be altered for the AA-MAS. Special considerations for the AA-MAS context then are discussed in order to clarify the implications of modified achievement standards and performance level descriptors (cf. Perie, 2008) for item and test development. In this discussion, the advantages and disadvantages of various options for modification are highlighted. Because a very real aspect of development for the AA-MAS is modification of items from the general assessment, examples of item analysis results from the latter are presented to illustrate approaches to identifying items for modification. Finally, psychometric consequences of test modifications are discussed, as they

play out in the assembly of test forms and in the analysis of technical characteristics of items and test forms. This chapter closes with consideration how best to document modifications during test development so the case can be made for validity and comparability as well as for interpretation of test results for both reporting on the achievement of individual students and the use of results for AYP purposes.

Best Practice in Item Development and Forms Assembly

Test development plays a key role in validation, and validity considerations play a key role in test development. The procedures to develop and revise test materials and interpretive information lay the foundation for test validity. Meaningful evidence related to inferences based on test scores can only provide scores with utility if test development produces meaningful test materials. Content quality is the essence of arguments for test validity (Linn, Baker and Dunbar, 1991). Test development is undeniably important to the proper interpretation of test scores and the inferences that are drawn from them (Kane, 2006). Users of test scores should study the specifications for the test, how they were derived, and the process by which the test is developed. Test development influences many aspects of validity (most importantly content validity) and many types of inferences. The purpose of this section is to discuss the issues and considerations associated with best practice in developing tests. The considerations that are provided in this chapter are consistent with the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999). The *Standards* constitute a seminal guide for proper test design and development.

Key to proper test design and development is the incorporation of universal design principles for all types of test items. These principles suggest an approach to assessment development based on principles of accessibility for a wide variety of end users. These principles are as applicable to the general assessment population as they are to a student population preparing to take an AA-MAS. Thompson, Johnstone, and Thurlow (2002) described

seven elements of universally designed assessments including inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility. Further guidance for states on universally designed assessments is provided by Lazarus, Thurlow, Christensen and Cosmier (2007).

Test Domain

A critical stage in the development of a statewide assessment is the design stage. In the design stage, important overall decisions about the test are made including: establishing a validation foundation for the test based on the state's academic content and achievement standards; designing test specifications that align with those standards; and reviewing, refining, and reaffirming validity evidence for test design.

Before test design can take place, it is important that the test developer understand the link between test purpose and test domain. The NCLB Act requires all public school students to participate in statewide assessments. The primary purposes of these assessments (given annually in certain grades and subjects) is to measure student progress towards state achievement standards and to hold schools, districts, and states accountable to bring all students to a proficient level in reading and mathematics. Test domain is used to refer to the various attributes used to define what a test should measure, including content topics, tasks, and process levels. In the NCLB context, states' content and achievement standards provide this definition of test domain. Understanding this connection between purpose and domain allows developers to determine what should and what should not be included on a statewide assessment. To be able to develop items that measure the test domain, developers need to define the test domain explicitly (i.e., which of the state's content standards are eligible for inclusion on the statewide assessment and which are not).

One of the major struggles with current statewide assessments is the large number of standards that most of the states have adopted and the need to align content standards with curriculum and statewide assessments. There are many issues in alignment, stemming from the wide variation in the specificity and clarity of state standards in defining what students need to know and be able to do, an imbalance between the number of standards and the testing time available, and the lack of agreement about the relative importance of the standards and the emphasis each receives in the statewide assessment. Several methods for evaluating alignment have been developed in recent years (Webb, 1999; Rothman, Slattery, Vranek, Resnick, 2002; Bhola, Impara, & Buckendahl, 2003). Using empirical validation strategies that focus on alignment can help identify, through more objective means, those standards that are the most important priorities for inclusion on the assessment.

Another important consideration at the early stages of test design, and particularly important for the AA-MAS, is defining the examinee population(s) for whom the test is intended (Quenemoen, Chapter 2, this volume). It is important to define the characteristics of the students who will constitute the examinee population for the test. Specifying the examinee population must take into account examinee characteristics that fall outside of the requirements of the test, but may constrain or confound the examinee's performance on the test.

Test Specifications

The next step in test design is to specify the important attributes of the items and test forms. Test specifications are often called blueprints because they specify how the test is to be constructed. Derived directly from test philosophy, test purpose and use, test audience, and empirical validity evidence gathered for the test, test specifications delineate the requirements for the subsequent stages of development, review, field testing, assembly, and evaluation of the end product. The test specifications should identify the content domain, cognitive processes, the

number and balance of items, the technical characteristics of the test, and the appropriate item formats (AERA, APA & NCME, 1999, Standard 3.3). Each of these is presented below.

Content domain. The empirical methods described above can be used to define the content topics to be included in the domain measured by a test. The ideal level of specificity of content topics in test specifications is that which ensures adequate control of all crucial elements of the content standards. In defining the content domain, the test developer must understand the structure of the content domain, how topics within the domain relate to each other, and how students build their knowledge over time.

Processes. It is equally important to specify the process requirements of the items in the test. These cognitive requirements represent the breadth, depth and range of complexity that students have been taught to use within the context of the content domain. There are multiple approaches to the classification of items by cognitive demand (see Pugalee & Rickelman, Chapter 5, this volume, for more detail).

Distribution of content and processes. Each content area and process needs to have a weight assigned to it in the test specifications that represents the relative emphasis to be placed on the topic or skill in a test form or item pool. These weights are important to the assembly of multiple, parallel test forms, and they are especially relevant in establishing comparability of an AA-MAS and the general assessment.

Item formats. Following the specification of test content and processes, the next aspect of the test specifications involves identifying the type(s) of items to be developed. At this point, the test developer needs to define the item format features that are required by the content and process specifications, specify the item types that possess those features, and comparatively evaluate each item type to identify those that might be preferred for reasons of coverage, economy, precision, response time, development and scoring costs, delivery constraints, and feasibility. The content and skills to be measured should drive the choice of item type. Selected-response (aka multiple-choice) items may be significantly more efficient in the amount of

information they gather per unit of testing time, but constructed-response items can add a performance dimension to observed scores and can be scored reliably, although not as inexpensively as selected-response items (Welch, 2006). Lindquist (1951) argued that the item type should match the criterion of interest. He indicated that the test developer should make the item format as similar to the criterion format as possible, recognizing the constraints of efficiency, comparability, and economy. As noted by Pellegrino (Chapter 4, this volume), there is no necessary connection between response format and cognitive level (e.g. multiple choice items can be used to assess higher-order thinking, and some performance tasks only measure surface knowledge).

Test Length. The next substantive consideration under the category of test specifications is length. The optimum test length is one that is accurate enough to support the inferences that will be made on the basis of the test results. Test length is a function of many concerns, most of which have been described including content coverage and item formats. However, in addition to such concerns, test length is also a practical constraint of testing time. Testing time may be influenced by constraints such as the administrative time periods such as class periods or the age of the student examinees

Technical Characteristics. The test developer must consider the specifications for the test(s) as a whole. This includes consideration of statistical specifications such as estimates of reliability, distribution of content and processes across the test form, test organization, administrative plan, and special accommodations.

To the extent the test specifications are well specified, the test forms produced will be far more parallel than they would be if developed from general specifications. By developing detailed specifications, the test developer considers many specifics of the test-development process before that process is begun, thereby resolving many of the issues that will arise during development. By carefully considering the major aspects of the testing process, the test developer can identify inconsistent or conflicting specifications early in the design process.

Well-developed test specifications drive the entire item-development and test-assembly process and serve as helpful directions to item writers, reviewers, and test users. Kane (2006) notes the importance of linking test specifications, item development, and forms assembly with the interpretive argument that will be used in attaching meaning to test scores. In this sense, validity is grounded in the test development process itself.

The content-validation process is also ongoing. Evidence supporting the test specifications should be reaffirmed on a regular basis. If state assessments are to reflect the curriculum and the expectations of teachers as to what their students need to be ready to learn or what they should have learned, test developers need to engage in a regular process to collect evidence to adjust or reaffirm the test specifications. Test design is at best an iterative process, one that repeatedly cycles through information gleaned through item development, test administration, and evaluation.

Item Development

Sound test development depends on well-defined, defensible item development. Sound item development is critical for providing the quality and consistency necessary to produce reliable test scores upon which validated test-score inferences can be made.

The development process should include considerations of universally designed assessments. Thompson, Johnstone & Thurlow (2002) identify specific questions for test developers to take into account as they develop items and design assessments. The considerations of universal design appropriate for all stages of test development include:

1. Incorporate elements of universal design in the early stages of test development.
2. Include disability, technology, and language acquisition experts in item reviews.
3. Provide professional development for item developers and reviewers on use of the considerations for universal design.
4. Present the items being reviewed in the format in which they will appear on the test.

5. Include standards being tested with the items being reviewed.
6. Try out items with students.
7. Field test items in accommodated formats.
8. Review computer-based items on computers.

Item Writing. Item writing is very much an iterative process, but it can be undertaken in a standardized manner. Item-development processes need to establish principles and procedures that take into account the various audiences and purposes of the program. Item-development processes for constructed-response items may also include the initial drafting of the scoring rubric simultaneously with the item writing. The qualifications of the item writers, the security of the process, and the training are all essential considerations for the item-development process.

The process adopted for developing items in any testing program is critical and must be considered in relation to issues of validity, reliability, and interpretability. The determination of the source of the item content depends upon test purpose and the inferences that need to be made based upon that content. Identifying those individuals who are qualified to develop items will be dependent upon the requirements of a particular assessment. Common procedures reflect a concern for demographic characteristics such as representation of the racial/ethnic backgrounds and gender of the examinee population.

Item Reviews. Once items have been developed, they should be subjected to a multistage, multipurpose review for content accuracy, fairness, universal design, and psychometric concerns. As with item writers, it is critical that item reviewers be experts in the area for which they are being recruited, that they be representative of the examinee population, and that they receive standardized training on the item attributes they are being recruited to evaluate.

The content reviewers should then be asked to review the items according to a set of established criteria. These criteria include scrutinizing items to ensure that they:

1. Align with the specified content standards,

2. Match to the specified processes,
3. Are technically correct,
4. Include effective distractors for multiple-choice items
5. Include draft scoring rubrics for constructed-response items,
6. Show clarity in response options (keyed option correct, distractors incorrect), and
7. Adhere to the specified item format.

Reviewers should also provide guidance on how to rephrase item stems, propose alternative keys and distractors, clarify scoring criteria, and identify ambiguous or confusing language in order to improve item quality. This guidance could be informed by cognitive interviews, think-alouds, and piloting the items in individual administrations to examine their cognitive demands.

All item development should be attended to fairness both in principle and in practice. Both the *Code of Fair Testing Practices in Education* (2004) and the AERA/APA/NCME *Standards* include obligations for ensuring fairness to test takers. The *Standards* also address obligations to ensure fairness through all stages of test development, test administration, and test use.

Assessments should also be reviewed for consistency with universal design principles to help ensure that optimal, standardized conditions are available for *all* students and that the test materials students encounter do not present unnecessary complexity in surface appearance.

Although content reviews are critical in consideration of internal qualities of a test, fairness reviews are equally essential in large-scale assessment programs as they are designed to ensure that all test takers have a comparable opportunity to demonstrate what they know and can do. Test fairness starts with design of the test and its specifications. It then continues through every stage of the test-development process, including item writing and review, item field testing, item selection and forms construction, and forms review. These reviews help to

ensure that items are evaluated from diverse viewpoints, not least of which are based on multicultural and gender-related perspectives (Camilli, 2006; Schmeiser & Welch, 2006).

Field Testing. Once the items have been reviewed and problems with them addressed, the items are typically prepared for field testing. Following the field test, item evaluations should be conducted using the field test data. Statistical analyses of field-test data typically include item analysis that is used to identify items that may be problematic. For constructed-response items, analyses may include: (a) descriptive statistics, such as the mean performance, standard deviation of the mean performance, range of responses, and frequency distribution of responses; (b) rater consistency and reliability estimates; and (c) correlations with multiple-choice items. For multiple-choice items, analyses may include difficulty and discrimination indices. It may also include an analysis of the distractors, student response patterns, and indications of speediness.

Items that appear statistically flawed should be carefully reviewed for possible content-related problems and for structural problems with the item (e.g., inadvertent cues to the key or distractors that are too close to the key).

Test Assembly. This stage of the development activity includes the process of selecting and organizing a particular set of items that will constitute a given form of a test. Test form assembly requires expert-level knowledge and skills in test construction, including an understanding of the relationships between the content and statistical characteristics of the items in a test and the test's measurement properties. Though test assembly is guided by test specifications, it also requires the well-reasoned decisions of a test developer who understands the relevant measurement principles and the judgments of content experts.

Test Specifications, Item Development, Forms Assembly, and Item-Level Statistics for the AA-MAS

The primary purpose of this section is to describe approaches and strategies and identify various considerations that test developers of AA-MAS should take into account as they develop

or modify general state assessments to create new alternate assessments. The intention is to focus on various steps of the development process, as outlined in the section on best practice, particularly those that are most relevant or available for modifications.

Several major assumptions guide the discussion in this section:

1. Modified achievement standards do not imply that the content standards are being modified. Rather, AA-MASs must adhere to the state content standards and must cover the same breadth and depth as the general assessment. The AA-MAS must be aligned to the content standards with respect to the content and process specifications but may be less difficult.
2. Given the substantial investment that states have made in the design and development of their testing programs, states may elect to modify existing assessments as a preferred approach to developing an AA-MAS.
3. The AA-MAS must satisfy reasonable technical requirements in terms of validity and reliability (Sato, Rabinowitz, Worth, Gallagher, Lagunoff & Crane, 2007). In order to maximize the validity of the AA-MAS, test developers must follow the same rigorous and iterative approach that has been established as best practice in test development.

Figure 6-1. Schematic Diagram of Interplay among Test Design Components

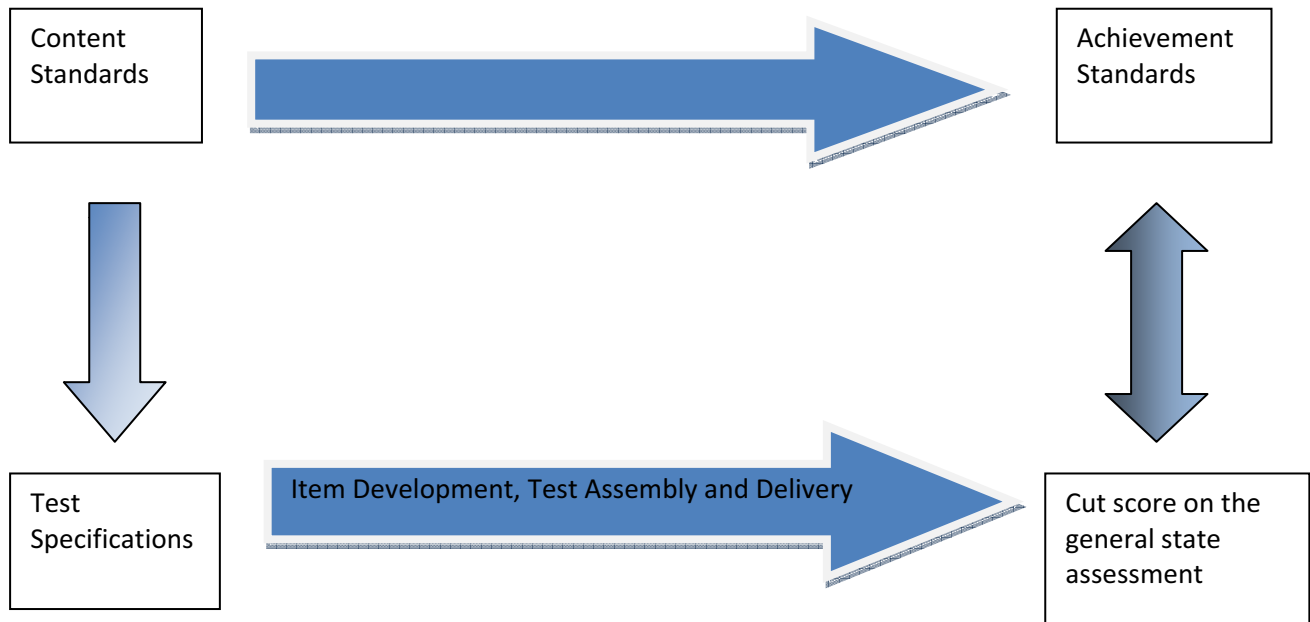


Figure 6-1 implies that there is a parallel relationship between content and achievement standards and test specifications and cut-scores. The content standards define what students need to know. The achievement standards define how well the students know the content standards and determine which students are proficient and which are not.

The test specifications are the translation of the content standards into assessment language (what will be on the test that the students need to know). The cut-score on the assessment determines the minimum score that indicates proficiency.

If test developers are not allowed to modify the content standards, then the test development effort for the AA-MAS should focus on test specifications (and item and test development) for allowable modifications. Depending upon the extent of these modifications, the AA-MAS may require either a new standards-setting process to locate the cut score representing the modified achievement standards, or a validation study to examine the fidelity of the existing cut score given the extent of the modifications to test specifications and items. If

the achievement descriptors for the modified achievement standards are rewritten to reflect a changed definition of proficiency, a new cut score study would also be necessary.

Special Considerations for Test Specifications

As with any assessment, the establishment of the test domain for the AA-MAS is the first consideration. State content standards and achievement standards define the test domain in the assessments being used to meet the requirements of NCLB and in the AA-MAS. The content standards are not to be modified. However, a review of the state content standards would be an appropriate first step. Grade-level content standards could be reviewed to ensure that students have an opportunity to achieve grade-level content. Students must have access to and must have received instruction in the grade-level content.

As discussed earlier, test specifications should include articulation of the content areas, process skills, and the balance between the two. They also include decisions about the item format, test length and technical characteristics of the assessment. Careful consideration of modifications introduced in the test specifications phase of development may produce assessments that more closely model the range of grade-level content appropriate for students eligible for this assessment. This may provide students with a better opportunity to be assessed on the same grade-level content standards as all other students, but with modifications to the expectations for the mastery of the content. Access to these modified assessments based on these changes to test specifications will ideally provide a better estimate of the student's achievement. However, making these modifications is not without compromise. Changes to the test specifications can result in modified assessments that are less comparable, less reliable or even less valid than the original assessment. Table 6-1 provides an overview of modifications that would be viable based on changes to test specifications. The advantages of making these changes and the potential limitations of such changes are also provided.

Table 6-1. Proposed Modifications to Test Specifications

Test Specification	Proposed Modification	Advantages	Possible Limitations
Content	Reduce the number of items per content standard	Maintains alignment with content standards	Comparability of test specifications of the AA-MAS to the general assessment
Process (or Cognitive Level)	Reduce the number of items per process standard Reallocate the process skills to reflect a more appropriate match to student abilities in terms of breadth, depth and complexity	Maintains alignment with content standards Ensures accessibility of test materials Reduces difficulty	May alter proportional representation of tested construct(s)
Content by Process Weighting	Adjust the relative weights of the content and process dimensions	Improves match of content strand to appropriate level of cognitive processes	The interaction between process and content is often difficult to quantify.
Item Formats	Diversify the item formats to maximize inclusion of those that are preferable for content and process coverage	Allows for partial credit to be given for short-answer, extended responses, and other types of open-ended items Allows for fewer items to cover more content and process standards if the appropriate items are written and appropriate adjustments made to scoring rubrics	Comparability of the test specifications of the AA-MAS to the general assessment Comparability of the scores from the AA-MAS to the general assessment New scales may need to be established Additional open-ended items would require additional resources for scoring, additional time for reporting Exposure of items and need for additional forms of the assessment Designing scoring rubric to be aligned with content standards while improving accessibility for students of interest
Test Length	Reduce the reading load of the assessment but maintain the number of items Reduce the overall number of items in the test	Allows students more time per item Decreases speediness impact of the assessment Reduces impact of student fatigue	Field testing of open-ended items on appropriate student population Reduces reliability Reduces precision of the cut-score decisions

<p>Technical Characteristics</p>	<p>Reduce the overall difficulty of the assessment by eliminating the most difficult items proportional to content standards</p> <p>Replace the most difficult items with simpler items covering the same content standards</p> <p>Increase the overall discrimination of the assessment by adding appropriate items</p> <p>Reduce the overall difficulty of the assessment by eliminating higher order process items</p>	<p>Increase proportion of students with IEPs exceeding the cut score</p> <p>Increase information about total score per-item included</p>	<p>Reduces reliability</p> <p>Reduces precision of the cut-score decisions</p> <p>May alter construct representation</p> <p>Increase costs associated with item development</p>
----------------------------------	---	--	---

Whenever modifications such as these are considered, experts who possess knowledge of the student population, can access relevant information, and are familiar with the state content standards are a critical part of the process. These experts need to address the complex interactions of the various approaches. For example, reducing the number of higher-order process items may benefit this particular student population, but may not be consistent with the state content standards.

Any modifications to test specifications must be consistent with the guiding assumptions cited previously in this chapter. That is, AA-MASs must adhere to the state content standards and must cover the same breadth and depth as the general assessment.

Using information from the *New York State Testing Program 2007 Technical Report for Mathematics, Grades 3-8* (NYSED, 2007, December), the example shown in Table 6-2 illustrates a hypothetical modification of content specifications and the relative weights of item format for the Grade 5 Mathematics. The entries in Table 6-2 in regular type are the number of items for each content standard for the general assessment in 2007, whereas the entries in italics are proposed modifications for a potential AA-MAS.

Table 6-2 – Implications Based on Modifications to Test Specifications

Content Standard	Multiple-Choice	Constructed-Response	Points Allocated
Number Sense and Operations	14 (9)	1 (1)	16 (11)
Algebra	3 (2)	1 (1)	6 (4)
Geometry	4 (2)	3 (2)	12 (8)
Measurement	2 (1)	2 (1)	6 (4)
Statistics and Probability	3 (2)	1 (1)	6 (4)
Totals	26 (16)	8 (6)	46 (31)

Note: Entries in regular type are for the general assessment, while the entries in italics and parentheses are for a potential AA-MAS.

The modification to the test specifications preserves (as closely as possible given fixed counts of items and points in the general assessment) the proportion of items aligned to each content standard as well as the proportion of items in each format based on a one-third reduction in the total number of items. It should be emphasized that the one-third reduction in

this example does not represent an arbitrary selection of items to remove from a general assessment, such as the most difficult items, but instead items whose removal does not alter the content balance nor detract from the technical quality of the resulting AA-MAS. The guiding principle is to remain true to the overall specifications while reducing the length (i.e., number of items) of the test.

An additional dimension that may be considered at this stage is process or cognitive level of the items in the AA-MAS. NCLB guidelines require evaluation of cognitive level, and test specifications in many states reflect this aspect of items as well as content strand. Even though cognitive level may not be specified on an item-by-item basis during test assembly, a distribution of items is often identified for three or more levels of a cognitive hierarchy, and attention to these features of items is important in proposed modifications for AA-MAS. Because constructed-response items (and the rubric specifications for high scores on those items) typically define higher levels of a cognitive hierarchy, their proportional representation in the AA-MAS is critical.

Proportional representation of content specifications, cognitive levels, and item formats is intended to preserve certain aspects of test validity to yield comparability. The reduction in total-score points and number of items can have a predictable effect on reliability. In the example, the NYSED math assessment had reported reliability coefficient of .93. The reliability estimate of the modified assessment depicted in the table is .87.

Special Considerations for Item Development

The item development process involves many varied, yet related, considerations. In this context, item development refers to the three major processes of item writing, item reviewing, and field testing. Although many of the processes are similar for both selected-response and constructed-response items, there are also characteristics of these two item types that would

suggest they should be discussed separately. Tables 6-3 and 6-4 present possible modifications, their advantages and limitations, for these two categories of items.

Consistent with needs for the design of the test specifications, item development for the AA-MAS will involve the identification of experts who are familiar with the student population and who are expert in providing appropriate and sufficient access to the general curriculum to prepare students to complete this assessment. Identifying experts to assist in the drafting or revising of items and reviewing these items for a variety of issues related to the student population needs will be critical. It will be critical that the role of these experts remain very central throughout the entire development process. Guidelines for use by the item writers and reviewers should include strategies for adapting items for students eligible for the AA-MAS. Frequent iterations of items should be expected in this process. All newly created items will need to be generated, reviewed, and revised throughout the development process by experts. All modified items should be subjected to the same rigorous review and refinement process. Reviews should take place as early in the process as possible, maximizing the benefits of the reviews prior to field testing.

Research on Item Modifications. Since the first draft regulations for the AA-MAS were issued by USED, ideas for item and test modification have appeared in white papers and plans submitted by states for peer review. Some of these ideas are included in this discussion. Much like the early years of work on testing accommodations for students with special needs, students with disabilities, and English language learners, ideas and innovations grow out of administrative imperatives and policy considerations for inclusion of all students in assessment and accountability programs. Empirical studies of the effects of accommodations on comparability of test score interpretations tend to lag behind the innovations themselves.

Although some research on adapting learning and assessment tasks for students with mild disabilities has been completed (AIR, 2000; Bergeson, Wise, Gill & Barlett, 2001) that would provide support for these suggestions, it would be misleading to assert that the

suggestions offered here for modifications of existing items for accessibility in the AA-MAS context have a strong research base or have been shown empirically to justify the spirit and intentions of the law with regard to comparability of test-based inferences or fidelity to the accountability provisions of NCLB. Rather they should be understood as rational approaches to the challenges of the AA-MAS that should be validated as any other aspect of an accountability system should be validated.

Empirical studies need to be conducted in order for test developers to provide the information necessary for appropriate interpretation. For example, studies that demonstrate a consistency between scores on a child's AA-MAS with other types of information about the child (IEP team evaluations, classroom performance) should be conducted. Studies that examine the relationship between the AA-MAS and other measures of the same constructs that are not necessarily used for state accountability purposes (performance on formative assessments, performance on diagnostic assessments) should also be planned. Research should also be planned to examine the internal structure for the AA-MAS as compared to the general assessment. Results from factor analysis for the AA-MAS could be compared to factor analysis results on the general assessment.

Table 6-3. Proposed Modifications to Item Development Process for Multiple-Choice Items

Process	Proposed Modification	Advantages	Possible Limitations
Item Writing	<p>Prepare reading passages; and related items with as much scaffolding as possible</p> <p>Control stimulus complexity to allow for the minimum level of complexity while remaining aligned with the content standards</p> <p>Write items with effective distractors for the AA-MAS population</p> <p>Use figures, pictures and graphs to aid students in understanding the items</p> <p>Remove irrelevant language from items that may distract students</p>	<p>Increase accessibility of items</p> <p>Maximize students' ability to demonstrate what they know</p> <p>Reduce effects due to distractibility and fatigue</p> <p>Remove construct irrelevant variance due to visual acuity, linguistic complexity</p>	<p>Increased development time</p> <p>Increased development budget</p> <p>Comparability issues</p> <p>Alignment issues</p> <p>Increase chance of correct response by guessing</p>
Item Reviews	<p>Review items for the possible revision of distractors that are attracting a very limited number of students</p> <p>Review items for the possible revision of distractors that are misleading to students</p> <p>Review items for irrelevant language</p> <p>Review figures, pictures and graphs for appropriate contributions and relevance</p> <p>Conduct cognitive interviews, cognitive labs and think-alouds</p>	<p>Ensure distractors are contributing to student information</p> <p>Ensure relevance to classroom experiences and consistency with everyday learning supports</p> <p>Graphics aid understanding</p>	<p>Increased development time</p> <p>Increased development budget</p> <p>Comparability issues</p>
Field Testing	<p>Field test items on student populations that are representative of students eligible for the AA-MAS to investigate the appropriateness and feasibility of the modifications.</p> <p>Field test all new or revised items on the appropriate sample of students</p> <p>Field test parallel variations of items (i.e.,</p>	<p>Provides preliminary analysis of processing levels</p> <p>Provide relevant statistics for use in forms assembly</p> <p>Provide statistics on both the AA-MAS and general assessment students to identify different response patterns</p>	<p>Increase development budget</p> <p>Limited access to appropriate students</p>

	<p>with various levels of scaffolding, with various levels of instruction, with various levels of language complexity) to identify those working most appropriately</p> <p>Beta test items on small samples of students. Conduct think-alouds with students to identify characteristics that are benefiting students and could be duplicated in future item.</p> <p>Generate DIF statistics from the field test to help make item selection</p> <p>Generate item analysis statistics from the field test to help evaluate an item's ability to discriminate for the students of interest</p>		
--	--	--	--

Table 6-4. Proposed Modifications to Item Development Process for Constructed-Response Items

Process	Proposed Modification	Advantages	Possible Limitations
Item Writing	<p>Develop items that lend themselves to scaffolding, allowing students the opportunity to work through controlled sections of the items</p> <p>Use figures, pictures and graphs to aid students in understanding the items</p> <p>Articulate the scoring criteria when the item is unusually drafted</p>	<p>Increase accessibility of items</p> <p>Maximize student's opportunity to fully demonstrate what they know</p>	<p>Generalizability of results may be reduced</p> <p>Reliability of the assessment may be reduced</p>
Item Reviews	<p>Review scoring criteria for content, fairness and universal design considerations</p>	<p>Increase appropriate difficulty of items</p>	
Field Testing	<p>Conduct cognitive interviews, cognitive labs and think-alouds</p> <p>Field test items on student populations that are representative of students eligible for the AA-MAS to investigate the appropriateness and feasibility of the modifications</p> <p>Beta test items on small samples of students. Conduct think-alouds with students to identify characteristics that are benefiting students and could be duplicated in future items.</p>	<p>Provide preliminary analysis of processing levels</p> <p>Solicitation of responses from representative students to be used establish the scoring criteria.</p>	<p>Costs of studies</p> <p>Exposure of items</p> <p>Security of items</p>
Scoring	<p>Allow for partial credit of responses (based on scaffold items structure)</p> <p>Evaluate item responses separately for both content and process skills</p> <p>Apply different types or scoring rubrics to the same item responses</p> <p>Generate distributional statistics for all constructed – response items for the AA-MAS and general assessment students</p>	<p>Provide relevant statistics for use in forms assembly</p> <p>Maximize the unique information available from constructed-response items</p>	<p>Generalizability of results may be reduced</p> <p>Reliability of the assessment may be reduced</p> <p>Comparability to general assessment may be weaker</p>

Modifications in Item Development. Consistent with the previous section on *Research on Item Modifications*, the effect of item modifications should be empirically studied. The methods used to modify the items should be thoroughly described as part of the validation process. Empirical and logical evidence should be also be provided. Table 6-5 illustrates the application of item modifications to several sample items. Modifications such as those recommended for items 2 and 3 employ the principles of universal design. Such principles are most appropriately included in the standard development procedures for all new item development. When the selected approach is the modification of an existing assessment, universal design principles are critical to inclusion in the modification process.

Table 6-5. Examples of Modifications in Development

Original Item	Revised Item	Description of Modification																		
<p>Anna baked 6 of 10 cupcakes for her classmates. Which number sentence describes how many more cupcakes Anna has to bake?</p> <p>A $10 + 6 = \square$ B $10 - 6 = \square$ C $10 \times 6 = \square$ D $10 \div 6 = \square$</p>	<p>Anna needs to bake 10 cupcakes. She has baked 6. Which number sentence describes how many more cupcakes Anna needs to bake?</p> <p>A $10 + 6 = \square$ B $10 - 6 = \square$ C $10 \times 6 = \square$ D $10 \div 6 = \square$</p>	<p>Simpler sentence structure Use of additional space between distractors Use of bold text to highlight question</p>																		
<p>Recycling brought to Green River Recycling Plant last month:</p> <table border="1" data-bbox="813 1360 1101 2043"> <tbody> <tr> <td>Week 1:</td> <td>1,178 pounds</td> </tr> <tr> <td>Week 2:</td> <td>1,065 pounds</td> </tr> <tr> <td>Week 3:</td> <td>1,879 pounds</td> </tr> <tr> <td>Week 4:</td> <td>1,997 pounds</td> </tr> </tbody> </table> <p>The closest estimate of the total recycling taken to Green River Recycling Plant was _____.</p> <p>A 4,000 pounds B 6,000 pounds C 8,000 pounds D 10,000 pounds</p>	Week 1:	1,178 pounds	Week 2:	1,065 pounds	Week 3:	1,879 pounds	Week 4:	1,997 pounds	<p>Green River Recycling</p> <table border="1" data-bbox="813 821 987 1178"> <thead> <tr> <th>Week</th> <th>Pounds</th> </tr> </thead> <tbody> <tr> <td>1:</td> <td>1,178</td> </tr> <tr> <td>2:</td> <td>1,065</td> </tr> <tr> <td>3:</td> <td>1,879</td> </tr> <tr> <td>4:</td> <td>1,997</td> </tr> </tbody> </table> <p>What is the best estimate of the total recycling taken to Green River Recycling last month?</p> <p>A 4,000 pounds B 6,000 pounds C 8,000 pounds D 10,000 pounds</p>	Week	Pounds	1:	1,178	2:	1,065	3:	1,879	4:	1,997	<p>Table with title, clear headings and reduced verbiage Alignment of numerals Less text in title Question format changed from incomplete sentence</p>
Week 1:	1,178 pounds																			
Week 2:	1,065 pounds																			
Week 3:	1,879 pounds																			
Week 4:	1,997 pounds																			
Week	Pounds																			
1:	1,178																			
2:	1,065																			
3:	1,879																			
4:	1,997																			

<p>Sarah and her family went to the grocery store. At the store Sarah and her brother Kyle went up and down the aisles looking for their favorite snacks. They each bought 2 snacks. One snack cost \$2. How much did the children pay for the snacks altogether?</p> <p>A \$4 B \$8 C \$12 D \$24</p>	<p>Sarah and her brother bought 2 snacks each at the grocery store. One snack cost \$2. How much did the children pay for the snacks altogether?</p> <p>A \$4 B \$8 C \$12 D \$24</p>	<p>Reduce demands on working memory Use of additional space between distractors Alignment of numerals</p>
<p>How do the authors portray Luis in the second paragraph?</p> <p>A As an eager student with many interests B As a popular boy with many friends C As someone who preferred performing to schoolwork D As someone who had trouble deciding what he wanted to do</p> <p>According to the passage, what is a dory?</p> <p>A A wild bird B A large pail C A small boat D A body of water</p>	<p>How is Luis described in paragraph 2?</p> <p>A A student with many interests. B A boy with many friends. C Someone who didn't like schoolwork. D Someone who couldn't decide what he liked.</p> <p>In the line marked with ✓, what is a dory?</p> <p>A A wild bird B A large pail C A small boat D A body of water</p>	<p>Simpler sentence structure Reduce irrelevant detail Use of additional space between distractors</p>
		<p>Visual aid introduced to help focus student on appropriate place in the reading passage Example of the use of a support Potential for scaffolding</p>

Special Considerations for Forms Assembly

Table 6-6 presents similar modifications for the forms assembly and administration of the AA-MAS. As with the previous sections, advantages and limitations exist for every type of modification. After items have been developed, field tested, revised, and deemed eligible for inclusion on an operational form, the test developer will select the operational items from the pool of field-tested items, using all available data (item-level statistics for difficulty, discrimination, DIFF, IRT parameter estimate). In general, test developers will be more successful in assembling forms if an item pool exists that allows for some degrees of freedom in the selection of items for inclusion on the operational form. Although building such a pool would require additional time and resources from the state, the benefit of such efforts would be realized in the assembly process.

Test developers should complete a match-to-specifications report based on the final assembled form. This process ensures the alignment of the modified assessment to the content standards, the test design and specifications and the guidelines for item selection. This process also provides documentation of the overall characteristics of the form and how these characteristics compare to the target test specifications. Comparisons of distributions of item difficulties and discriminations from the field test statistics to the target technical distributions should be made. Estimates of reliability for the assembled form and estimates of the standard error of measurement should also be included in the match-to-specifications report. This is critical information to the review and approval of the assembled form. This information provides one last opportunity for the test developer to make changes to the composition of the assessment before an operational administration.

Table 6-6. Proposed Modifications to Test Assembly Process

Process	Proposed Modification	Advantages	Possible Limitations
Selecting items for inclusion on the operational form	<p>Assemble forms from the least difficult to the most difficult item</p> <p>Assemble items to reduce the number of items within any one test section</p> <p>Assemble items to minimize changing from one content standard to another (for example, within a math test, group the geometry items together, then group the measurement items together)</p>	Increased accessibility for students	Comparability to general assessment
Test layout	<p>Maximize white space in the test booklet</p> <p>Follow principles of universal design</p> <p>Limit the number of items per page or screen presentation</p>	Eliminate distractions for students	Comparability to general assessment
Test administration	<p>Minimize the number of items presented in any separately timed section of the assessment (for example, if a 44-item math test could be divided into two 22-item sections, assemble and administer in the shorter blocks)</p> <p>Minimize the transferring of information from a test booklet to an answer document by offering online delivery, consumable test booklets or other mechanisms for capturing the student responses</p>	<p>Reduces fatigue</p> <p>Reduces examinee error</p>	May differ from general assessment administration format

The greater the change in development, selection, presentation and administration of items change, the less likely it will be that states can “link” the performance on the AA-MAS to performance on the general assessment. However, one strategy, the reduction of the number of items proportional to the test specifications for the general assessment offers the possibility of relating or linking the reduced version of the assessment to the full length of the assessment. In such instances, the modified assessment may be structured to maintain the intuitive understanding of the standard score scale used for the general assessment. This approach may offer some utility with respect to the interpretability of the results. Table 6-7 offers a quick summary of the impact on comparability and the scale for three tiers of change.

Table 6-7. Impact of AA-MAS Strategy on Comparability and Scale

Possible Strategy	Comparability to General Assessment	Scale Considerations
Develop new assessment	None	New standard setting New scale necessary
Modify items (i.e., shift in item formats, number of distractors, scaffolding)	Limited	New standard setting New scale possible
Reduce number of items (proportional to content standards)	Linked	Retention of scale Validation of cut scores with standard setting study

Special Considerations for Evaluating Statistical Characteristics of AA-MAS Items

A standard activity in any test development context is the statistical evaluation of items for an assessment. In the AA-MAS context this might happen in various places in the item and test development workflow as different types of item statistics become available. Preliminary data might exist, for example, from small samples in which item modifications are pilot-tested for accessibility and feasibility of administration. Field test data on larger samples may be reviewed after formal content and sensitivity reviews take place and prior to test form assembly. In test development processes using item-response theory, item- and person-fit statistics on larger samples may be needed. A significant challenge in the AA-MAS context, however, is the fact

that item statistics may not be readily available at ideal times from ideal samples of students from the population.

A reasonable approach to developing an empirical basis for item selection and modification is to examine conventional item analysis statistics for items in the general assessment. Because states have been administering their general assessment to all students except the one-percent with the most severe disabilities since 2006, item-level data for students who might be deemed eligible for the AA-MAS assessment presumably exist. Statistical characteristics of items in this target population may provide some insights for item selection and modification.

Item Analyses for Contrasting Groups. Conventional item analyses for dichotomous, multiple-choice items produce observed percent correct or p-values to measure item difficulty, correlations between items, and total scores to measure item discrimination, as well as more detailed indicators of item functioning such as the percent of examinees choosing each multiple-choice option, and correlations between option choice and total score as measures of distractor discrimination. Also informative for the latter concept is the percent of high- and low-scoring examinees choosing each distractor. In addition, many state assessments are likely to have similar item statistics based on item-response theory. Test developers use indicators of item difficulty to assemble test forms appropriately matched to the achievement level of the examinee population and indicators of discrimination to ensure some degree of homogeneity in the selected items. Both item characteristics influence the reliability and internal validity of the assembled test form.

Of interest to the present discussion is the extent to which item statistics might provide insight into the performance of items in the target population for the AA-MAS. The specification of that population means that p-values are likely to be, by definition, smaller in the AA-MAS population than in the full examinee population as that population consists of students not likely to be proficient on the general assessment. One might also expect item-total correlations to be

smaller due to the restricted range of total scores in the AA-MAS group. A full array of item analysis statistics can provide test developers with some guidance on the relative performance of items in the two-percent population. For example, items with marked differences in difficulty and markedly low discrimination for the AA-MAS group could be argued to contribute to low scores without contributing to observed score variance for the students of interest. Such item statistics combined with poorly performing distractors could support the elimination of these types of items from the two-percent assessment.

Table 6-8 provides a concrete example of distributions of p-values and item-total correlations on grade 5 state math and reading assessments in an examinee group identified as potentially eligible for the AA-MAS in a given state and the general student population. The AA-MAS group consisted of students who had IEPs and who were deemed not proficient in two consecutive years of the general assessment. The results in the table are based on the second of those years.

Table 6-8. Mean (SD) Difficulty and Discrimination for Items in an AA-MAS and General Assessment

Population	Difficulty	Reading		Math	
		Discrimination	Difficulty	Discrimination	Difficulty
General	.67 (.14)	.57 (.12)	.68 (.13)	.57 (.13)	.68 (.13)
AA-MAS	.37 (.15)	.27 (.09)	.35 (.10)	.30 (.10)	.35 (.10)

As can be seen from the entries in Table 6-8, the mean difference between item difficulty in the two student populations is substantial and translates into effect sizes of 2.5 and 2.1 for math and reading, respectively. Standardized mean differences of this magnitude are extremely rare in typical comparisons of subgroups in educational testing and suggest that whatever modifications of the general assessment are introduced for the AA-MAS population, their impact on performance must indeed be great if the AA-MAS form is expected to alter AYP results. Proposed modifications of items during AA-MAS test development must attend to characteristics of items in such a way that the modification effort will have a measurable impact on test results

in the context of accountability. Specific modifications based on item statistics will be considered below.

The distributions of p-values and item-total correlations are shown in the stem-and-leaf plots in Figure 6-2. Each statistic is expressed on a scale from 0 to 1. The plots show the tenths digit in bold and the hundredths digit in regular type. Hundredths digits to the right in italics are for the general student population, and those to the left are for students eligible for the AA-MAS. As can be seen from the figure, the distributions of p-values and item-total correlations are generally symmetrical in both the general population and the group identified for the AA-MAS. The p-values for math in the AA-MAS sample are somewhat positively skewed. The distinctive feature of these distributions is the small degree of overlap. This is of particular interest in the case of the item discrimination indices. Ideally test developers would like the distribution of discrimination indices to be similar. However, range restriction on total score is likely to systematically lower item-total correlations in the AA-MAS population. The dramatic separation of the distributions of these correlations suggests there may be additional reasons for low discrimination. Understanding why would be an important part of AA-MAS test development if modification of items from the general assessment is the selected approach. The item-total correlations suggest that even within the AA-MAS population, items in these sets, irrespective of content, possess idiosyncratic characteristics that reduce their overall correlation with total scores. If these characteristics can be isolated by content or statistical analyses of item keys and distractors, for example, then perhaps target modifications at the item level could at once make items less difficult and increase their internal consistency and correlations with total test scores. Some illustrations of these ideas are presented below.

Figure 6-2. Stem-and-Leaf Plots of Item Difficulty (proportion correct) and Discrimination (item-total correlation)

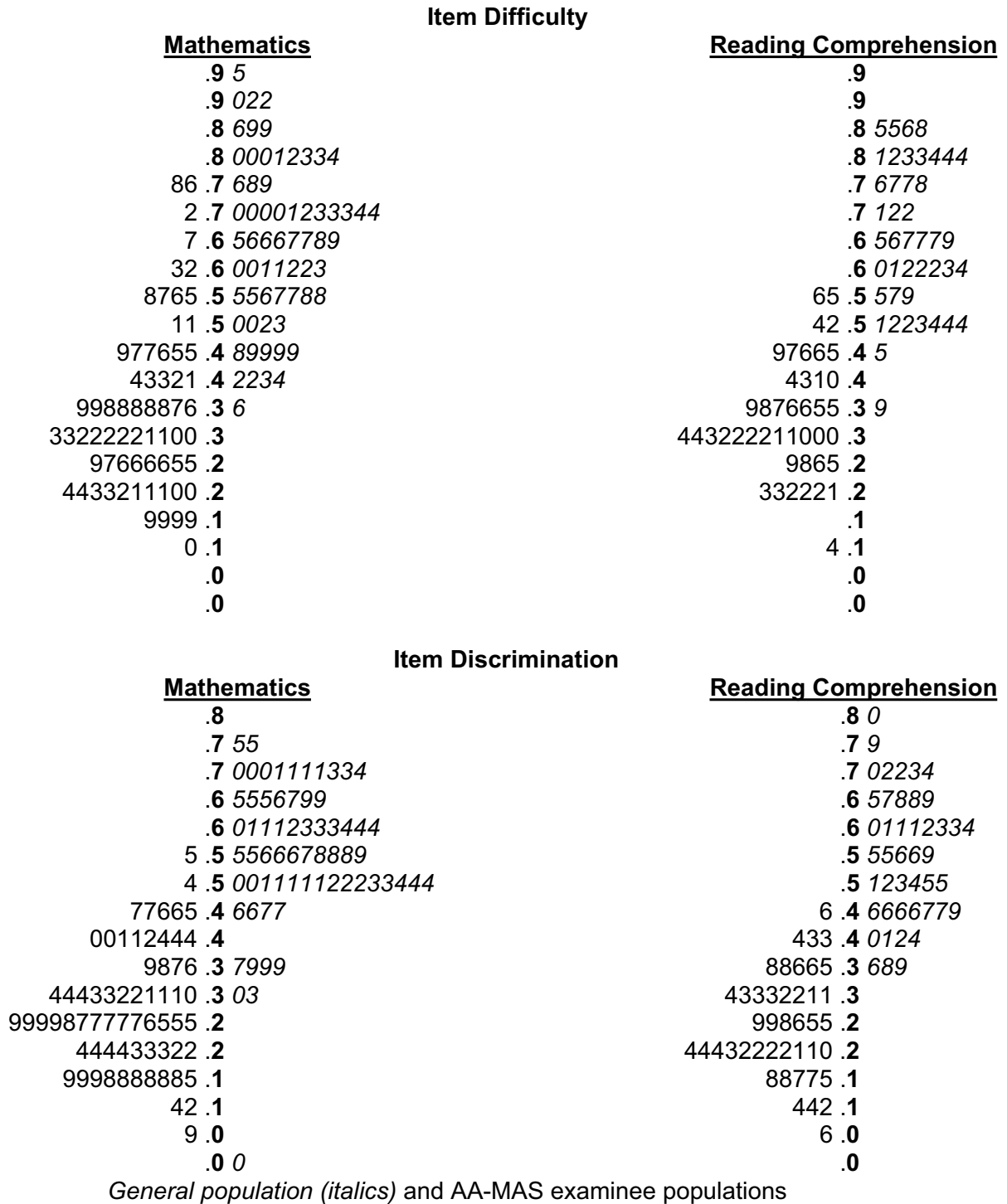


Table 6-9 gives an example of a distractor analysis from a statewide mathematics assessment given to nearly 40,000 students in grade 5. The item measures a student's ability to compute total length of six objects and to convert inches to feet. Data marked AA-MAS are from a group identified previously as eligible for a modified assessment. This item has reasonable statistical properties in the general population (difficulty = .42, discrimination = .53). In that population, 35% are drawn to the combination of numbers, 5 and 4, that reflect correct calculation of total length but no conversion from inches to feet (54 inches, thus 5 feet 4 inches). A simple distractor analysis shows the nature of the error most common in the general student population.

Table 6-9. Illustration of Distractor Analysis

<i>A brick is 9 inches long. If 6 bricks are lined up, one after the other, in a row, how long is the row of bricks?</i>			
	Options	General Performance	AA-MAS Performance
A	1 foot 3 inches	9%	17%
B*	4 feet 6 inches	42%	19%
C	5 feet 4 inches	35%	25%
D	6 feet 9 inches	14%	39%
Item Statistics:			
Sample Size		37,223	2,432
Difficulty		.42	.19
Discrimination		.53	.12

The students in the AA-MAS eligible group were drawn to option C as well. Those students demonstrated a similar misunderstanding. However, option D (6 feet 9 inches) was the most frequent (39%) response in the AA-MAS group. This distractor simply repeats the specific numbers used in the item stem and indicates no calculation and no conversion of units. The summary statistics (difficulty = .19, discrimination = .12) indicate this item to be providing very little information about total test scores for the AA-MAS population. Based on such information, the test developer could choose to replace distractor D with a different option, eliminating the repetition of specific numbers used in the stem, or reduce the number of distractors by eliminating D.

This concrete example highlights the complexities involved in modification strategies such as elimination of distractors. Eliminating the most popular distractor in the total group would do little to change the behaviors of this item in the AA-MAS group. Moreover, such a strategy would eliminate the distractor that carries a meaningful message in an error analysis. Distractor elimination is clearly going to alter the construct interpretation of item performance and perhaps do so without any gain in relative item difficulty and impact. This concrete example is designed to illustrate that eliminating distractors can have untold effects on the meaning of resulting test scores. Cases such as this might be better addressed by the elimination of entire items that can be arguably shown to contribute little to total scores in the AA-MAS population.

Differential Item Functioning. An apparently straightforward analysis for identification of items for modification in the AA-MAS context would be Differential Item Functioning (DIF). DIF methods are designed to detect items with different item characteristic curves in two populations, in other words systematic differences in the item performance of examinees in groups matched on general achievement in the domain. As discussed by Abedi (Chapter 8, this volume), DIF methods are routinely used as part of the test development process to screen items for psychometric appropriateness with respect to background characteristics of examinees such as gender, ethnicity, SES, native language, and disability status. DIF methods have the potential to provide insight into facets of item design that may unknowingly create a relative advantage or disadvantage to examinees that is unrelated to the construct measured by the test.

In the AA-MAS, DIF methods might be thought to offer insight into differential performance at the item level. As noted by Abedi, however, when DIF methods are used in assessment context with multiple focal groups of interest (e.g. students with disabilities, linguistic minorities, ethnic minorities, etc.) it can become difficult to find consistency in the flagging of items. Moreover, the statistical limitations of DIF methods have necessitated the development of judgmental criteria for evaluating the magnitude of DIF (e.g. its expected

influence on total scores, Zieky, 1993) to supplement statistical criteria used in testing null hypotheses of no DIF. In particular, DIF methods tend to perform poorly when groups differ markedly in overall test performance and when the variable used for matching (typically total test score) does not allow adequate matching throughout its range (add DIF references). Given the effect sizes presented previously as well as the distributions of p-values in a prospective AA-MAS population relative to the general population, DIF methods are likely to prove difficult to apply in the test development process for the AA-MAS. Large mean differences between groups and sparseness of scores in the upper ranges of total score distributions are likely to produce spurious DIF in the AA-MAS context (Holland and Thayer, 1988; Camilli, 2006).

Validating the AA-MAS

Validity remains the most fundamental consideration in developing and interpreting any assessment. Although this chapter has been devoted to the development of a sound AA-MAS, it is the use and interpretation of these scores that must be validated. There are numerous sources of evidence that might be used to evaluate a proposed interpretation. One critical type of validity evidence is based on the test content. As defined in the Standards (AERA, APA, & NCME, 1999), content refers to the themes, wording and format of the items, tasks or questions on a test, as well as the guidelines for procedures regarding development, administration, and scoring. This chapter has attempted to discuss issues about differences in meaning or interpretations of test scores for an AA-MAS when compared to a general assessment. Of particular concern was the extent to which construct-irrelevant components could be eliminated to avoid disadvantaging students eligible for an AA-MAS, to create an assessment that provides students an opportunity to demonstrate what they know and can do. Consistent with Marion (Chapter 9, this volume), content-related evidence requires evaluating the interaction of both content and process required of the test items and documenting that the interaction is what is expected.

The responsibility for validating the AA-MAS is shared between test developers, users and education policymakers. An important aspect of test validation in this regard is the documentation of the test and item development processes used for the AA-MAS, the specific steps followed in the test development workflow, the types of item modifications chosen, the expert analysis of the cognitive demands of the items, the impact measured through think-alouds, cognitive interviews, and field testing, studies examining differences in performance on items, and the changes to test specifications and distributions of items across formats, content strands and cognitive levels. Education policymakers are likely to weigh in on general parameters of AA-MAS development such as the representation of subject matter included, the process of setting standards, and the budget allocated for test development, delivery, and reporting. More specific aspects of validation are a joint responsibility of test developers and users. In statewide assessment, SEAs are both developers and users, but SEAs typically work with one or more contractors who carry out the activities associated with item and test development. Validation of the AA-MAS may be a responsibility of an SEA but an action step that is incorporated into the RFP process and the deliverables specified during contract negotiations. If the goal is to develop the foundation for a validity argument in support of proficiency-related inferences based on the AA-MAS (Kane, 2006), then the outline of the argument needs to be formulated in the joint work of an SEA and its contractors.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- American Institute of Research (AIR) (2000). *Effects of item scaffolding on student responses: A cognitive laboratory study*. Washington, DC: AIR for Institutes for Research for the National Assessment Governing Board in support of contract # RJ97153001.
- Assessment and Accountability Comprehensive Center (2007, November). *Assessments based on modified achievement standards: Critical considerations and implications for implementation*. San Francisco: WestEd.
- Bergeson, T., Wise, B. J., Gill, D. H., & Bartlett, K. M. (2001). *Adaptations are essential: A resource guide for adapting learning and assessment tasks for students with mild disabilities*. Olympia, WA: Special Education Section of the Office of the Superintendent of Public Instruction. Available at:
<http://www.k8accesscenter.org/accessinaction/documents/EARLYwritingADAPTATIONS.pdf>.
- Bhola, D. S., Impara, J. D., & Buckendahl, W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport, CT: American Council on Education/Praeger.
- Code of Fair Testing Practices in Education* (2004). Washington, DC: Joint Committee on Testing Practices.
- Filbin, J. (2008) *Lessons from the initial peer review of alternate assessments based on modified achievement standards*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Brown (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota: National Center on Educational Outcomes.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 119-158). Washington, DC: American Council on Education.

- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Education Researcher*, 20(8), 15-21.
- Marion, S. (2007). *A technical design and documentation workbook for assessments based on modified achievement standards working draft*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15-29.
- New York State Department of Education (2007, December). *New York State Testing Program 2007: Mathematics, Grades 3-8, Technical Report*. Albany, NY: Author.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report 566). Los Angeles: UCLA Center for Research on Evaluation, Standards and Student Testing.
- Sato, E., Rabinowitz, S., Worth, P., Gallagher, C., Lagunoff, R., & Crane, E. (2007, September). *Evaluation of the technical evidence of assessments for special student populations* (Assessment and Accountability Comprehensive Center Report). San Francisco: WestEd.
- Schmeiser, C.B. & Welch, C.J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: American Council on Education/Praeger.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. (2002). *Universal design applied to large-scale assessments* (NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- U.S. Department of Education (2007a, April 9) *Final Rule 34 CFR Parts 200 and 300: Title I — Improving the Academic Achievement of the Disadvantaged: Individuals with Disabilities Education Act (IDEA)*. Federal Register. 72(67), Washington DC: Author. Available at: <http://www.ed.gov/admins/lead/account/saa.html#regulations>.
- U.S. Department of Education (2007b, July 20), *Modified Academic Achievement Standards: Non-regulatory Guidance*. Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education. Available at: <http://www.ed.gov/admins/lead/account/saa.html#regulations>.
- U.S. Department of Education (2007c, December 21). *Standards and Assessment Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001*. Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education. Available at: <http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf>
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. (Research Monograph No. 18). Madison, WI: National Institute for Science Education.

Welch, C. (2006). Item and prompt development in performance testing. In S.M. Downing and T.M Haladyna (Eds.), *Handbook of test development* (pp. 303 – 327). Hillsdale, NJ: Erlbaum.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

CHAPTER 7

DEVELOPING MODIFIED ACHIEVEMENT LEVEL DESCRIPTORS AND SETTING CUT SCORES

Marianne Perie

In developing an alternate assessment based on modified achievement standards, test developers can improve the design process by paying close attention to those modified achievement standards. By collaborating with policymakers, they can together work through the issues of what proficiency means for this group of students, what type of modified achievement standards are appropriate, and how best to measure them. In fact, defining the achievement standard is the one area where it is most important for policymakers to work directly with test developers to ensure coherence so that what is intended as a state standard and goal is actualized through the assessment and interpretive materials. As discussed previously, the USED regulation requires that the modified achievement standards:

- Be aligned with a state’s academic content standards for the grade in which the student is enrolled.
- Be challenging for eligible students, but may be less difficult than grade-level academic achievement standards.
- Be developed by grade level, not grade span.
- Include at least three achievement levels.
- Be developed through a documented and validated standard-setting process that includes broad stakeholder input.

The only other guidance that policymakers are given from the federal government in defining modified achievement standards is that they are expected to represent a “less difficult expectation of grade-level content standards.” Inherent in this guidance is a tension between ensuring the tests measure the same breadth and depth while being less difficult. This chapter will examine how the achievement standards work within that tension to provide a less difficult

performance target. It will analyze the different dimensions of modified achievement standards, describe the various components, and provide suggestions for drafting descriptors and setting cut scores.

Defining Achievement Standards

An achievement standard defines a level of performance and includes both a minimum cut score and a written description that distinguishes the level of performance from other defined levels. It consists of four components: number of levels, names of levels, a descriptor for each level, and a cut score for each level.

Numbers

According to the regulations, the number of modified achievement levels to be defined includes a minimum of three: one distinguishing proficient performance, one above, and one below. However, some states may want to add one or two more levels to meet the goals of their assessment. The majority of states have four performance levels for the general assessment. It may be desirable to mimic the structure of the general assessment by including the same number of achievement levels for the modified assessment so that report cards and interpretive material can be standardized as much as possible. Or, as is the case for New York, it may be necessary to include the same number of levels to more easily incorporate them into a formula for a performance index calculation. However, it is also an option to consider these levels extensions of the lowest grade-level achievement standard, in which case a parallel number would not be necessary. In this case, a state would need to consider the number of levels necessary to convey the message intended by these modified achievement standards.

Caution is urged in developing more than four levels. As noted in Perie (2008), it can be difficult to describe meaningful differences across more than four levels. In addition, any particular test has a fixed amount of measurement power that depends primarily on the number and quality of the questions in the test. "The more cut scores there are in any given test, the

less measurement power the test developer can devote to each cut score, and the less information there is around each cut score.” (ibid, page 17). Given the nature of the AA-MAS and the typical rationales for developing one (i.e., school accountability), it seems unlikely that more than four levels would be needed.

Names

Beck (2003) indicated that naming conventions should be developed as the first step in defining performance. With modified achievement levels, the first question in naming them is whether the names should be the same or different from the level used in the general assessment. While it may be tempting to assign the levels the same name, state policymakers could also consider using different names to avoid confusion and simply designate one name to be the equivalent of “proficient” for purposes of AYP. In fact, some states have received feedback from their peer reviewers advising them to select different names for their modified achievement levels.

Policymakers can also consider how these modified standards relate to grade-level standards and portray that in the name. That is, if these modified achievement standards are truly downward extensions of the grade-level achievement standards, the names should reflect their relationship with the general assessment. For instance, some state policymakers have considered naming the levels relative to the general assessment, such as “not ready for the general assessment,” “almost ready for the general assessment,” and “ready for the general assessment.” Or, the same idea could be used to talk about achievement relative to grade-level standards (e.g., “near grade-level proficiency”). Some state policymakers have also chosen not to call any modified achievement level “advanced” as they believe student performance needs to be measured against grade-level standards before it can be called “advanced.” It is important to keep in mind that the names of the modified achievement levels often express the values of the policymakers or the intent of the assessment.

Descriptors

Achievement level descriptors put into words how good is good enough. That is, they qualitatively describe the performance expected of a student at the “proficient” level or the “basic” level. They must be aligned with the state academic content standards and describe breadth and depth of the standards appropriate to the assessment, so that they represent knowledge and skills that are evaluated by the assessment. Although the breadth and depth of the assessment must be parallel to the general assessment, these modified descriptors may be written to a “less difficult level.” That is, while the assessment must measure similar levels of depth of knowledge, perhaps competency of a lower depth of knowledge is all that is needed to be proficient using the modified achievement standards.

Ideally, the descriptors will be written so that they clearly differentiate among levels and progress logically across levels. That is, to improve articulation across levels, write the “proficient” descriptor to be appropriately more rigorous than the “basic” descriptor. In addition, considering the entire assessment program will help ensure that the descriptors also progress logically across grade levels (e.g., the descriptor for grade 5 “proficient” is sufficiently more challenging than the descriptor for grade 4 “proficient.”) It is important to take great care in writing the descriptors as they drive not only the standard setting process, but also the reporting, score interpretation, and potentially the item-writing process. In fact, many in the field claim that the descriptors are instrumental to the validity and defensibility of the standard-setting process (cf., Cizek & Bunch, 2007; Hambleton, 2001). More detail will be provided about this step later in this chapter.

Cut Scores

The fourth component of achievement standards is the cut score. Cut scores define the number of points necessary to reach each performance level. They are typically set after the assessment has been field tested so that statistics are available to inform the process. Then

recommended cut scores come from a committee using any of a number of possible methodologies to determine the best cutoff points. The regulations require the use of a documented and validated methodology, but the choice of methods is left up to the test developers and policymakers. Ideally, a broad range of stakeholders would be involved in the process, typically including both special educators and content experts. It is important to fully document the process, including a rationale for selecting a particular methodology and the process for selecting the committee. More detail on the methods, procedures, and documentation will be provided later in this chapter.

Defining Proficiency

The biggest issue that state policymakers will wrestle with is what proficiency means for these students. That is, we need to determine what we mean by “modified” achievement standards. Defining the levels is an important step in standard setting. Berk (1996) discussed the importance of providing explicit behavioral descriptions of each level, saying “the interpretation of the final cut scores hinge on the clarity of the behavioral definitions” (p. 224). Previous chapters discussed issues related to the interaction of cognition, instruction, and assessment and provided some insights into providing this clarity. Understanding cognition and improving instruction can have large implications for determining what proficient means on a given assessment. And it is here that policymakers will wrestle with making a test of similar breadth and depth “less difficult.”

Taking information from the earlier chapters on how students learn the content, and ways in which the content increases in difficulty, provides some insights into writing meaningful descriptors. If there was one learning progression that all students followed, the task of writing achievement level descriptors would be greatly simplified as we could simply identify points on the learning continuum that represent “basic,” “proficient,” or “advanced” achievement. However, as discussed previously (see Pellegrino, Chapter 4, this volume; Pugalee &

Rickelman, Chapter 5, this volume), there is little to no agreement on standard learning progressions for any population, let alone a population of students with disabilities.

With this population that may include all disability types and different learning progressions, it will be vitally important to clearly define the population and understand why the students are not achieving at grade level before we can describe proficient performance for them. By considering the grain sizes (depth and breadth) of learning targets along a continuum (Gong, 2007), instructional scaffolding that best supports how they learn, and an appropriate level of cognitive challenge for their grade level, we can better understand achievement of these students as compared to students without disabilities. These differences will greatly influence the writing of PLDs.

For example, Pellegrino (Chapter 4, this volume) discusses the possibility that low achievers may have a similar set of knowledge and skills as high achievers but may not have cognitively organized that information as efficiently so they are not able to access it as readily. One solution is to design a test that reduces the burden on working memory or that includes supports to help students better organize information or more easily determine the best strategy to solve a problem. This type of theory would need to be captured both in the test design and in the definition of proficiency. As another example, Pugalee and Rickelman (Chapter 5, this volume) discuss ways of modifying the domain targets systematically within each depth of knowledge level. This approach could again be explored in both the test design and the descriptors. Most importantly, there should be a guiding philosophy about the model of learning for students with disabilities who are low achievers and thus eligible for this assessment. That guiding philosophy should drive the definition of proficiency and the test design simultaneously.

As discussed in Perie, Hess, & Gong (2008), it is usually important to consider the definition of proficiency for the 2% assessments long before standard setting, as it could drive the design of the assessment. That is, we can work to develop items that measure the features that policymakers have determined are important to distinguish proficient performance from

performance below that level. However, as discussed by Welch and Dunbar (Chapter 6, this volume), it is also possible to modify the general assessment using statistical information gathered from an administration of the general assessment to the target population. If, as they suggest, a test developer takes the option of creating an AA-MAS by simply eliminating the most difficult items proportional to the content standards, the cut score could be mapped from the general assessment to the AA-MAS. Then, the descriptor would be modified after the fact—focusing on the general knowledge and skills measured by the items that appear to map to each achievement level.

One issue that several states are considering is whether the AA-MAS is at the lower end of some continuum that includes the general assessment or whether it is a completely separate test that measures the same content standards but to a less rigorous extent. For instance, policymakers need to decide whether they see the AA-MAS as a stepping stone for students to move towards grade-level achievement standards, or whether they believe that a student's disability will require a different type of assessment. One implication for this decision is the definition of proficiency. Should proficiency be defined in terms of how ready a student is to be assessed on grade level assessments or should it be defined simply as proficient on this separate assessment with no explicit or implicit link to performance on the general assessment?

Another, similar, consideration is how this AA-MAS fits between the AA-AAS and the general, grade-level assessment. Most states appear to be developing an AA-MAS that is closer in design to the general assessment than to the AA-AAS. But, how should the achievement standards compare? One possibility is to consider proficiency as being just below proficiency on the general assessment—that is, somewhere between “basic” and “proficient” performance on the general assessment. Another possibility is to simply shift down the levels one step, so that “proficient” performance on the AA-MAS will be similar in nature to “basic” performance on the general assessment. This approach is one way to keep the breadth and depth similar across the two assessment types but making the AA-MAS “less difficult” by requiring less knowledge and

fewer skills to reach proficiency. This type of relationship among the assessments would have implications for the intended comparability of the assessments. (Please refer to Abedi, Chapter 8, this volume for more details.) It also has implications for the development of the modified achievement level descriptors. In this case, the committees would start with the grade-level descriptors for both basic and proficient, and try to write a modified proficient descriptor that falls in between the two, or perhaps closer to the basic level.

State policymakers' beliefs and values also come into play as they consider whether students who would take this assessment are capable of learning grade-level materials to the grade-level standard. One possible theory is that these students can learn grade-level material as well as their nondisabled peers, but they take longer to master each unit and thus do not complete the curriculum by the end of the year. Following this theory would lead to a description of proficiency that is similar to grade-level proficiency for material learned earlier in the year, but requires less of students on material learned later in the school year. However, this approach could be difficult to defend as it may violate the mandate that the breadth must remain equivalent across the two assessments and only the difficulty may be modified. The breadth described by the modified proficient descriptor should not be narrower than the breadth of the grade-level proficient descriptor.

Another theory is that these students can learn grade-level material as well as their nondisabled peers, but they require specific supports to do so. That is, the ultimate goal for reaching proficiency may be the same, but it includes conditions. For example, the proficiency standard may include clauses that describe the scaffolds available on the test, such as segmenting text, providing strategies, supplying definitions, etc. Then the descriptor could indicate that the student measured against modified achievement standard has similar knowledge as the proficient student measured against grade-level achievement standards, but he/she may require more supports (e.g., less vocabulary load in the test item, use of graphic organizers to organize information before solving a problem) to demonstrate that knowledge.

Regardless of which theory drives the process, it is important to articulate that theory and clearly state the inferences policymakers and educators wish to draw from the AA-MAS. Fitting this context with the design decisions made and the definition of proficiency is central to forming a coherent validity argument, which will be discussed more fully in Marion (Chapter 9, this volume).

Applying Theories of Learning to Modified Achievement Level Descriptors

If state policymakers start with the perspective that the modified achievement level descriptors are closer in nature to the grade-level achievement standards than to the alternate achievement standards, then one strategy for drafting the descriptors is to start with the grade-level descriptors and modify them appropriately.² These modifications can take several forms, depending on the theory one is following, as described in the previous paragraph.

The first question that needs to be answered is whether the knowledge and skills required for proficiency within the modified achievement standard are the same as with the grade-level achievement standard but with more supports and scaffolding, or the knowledge and skills are actually different. If those drafting the descriptors believe the first description is true, that the standards are the same but students require appropriate supports, then, they can modify the grade-level descriptor for “proficient” accordingly. For example, a grade-level standard may state “student is able to read a fictional text and identify key elements of the story” while a modified standard may state “when the text is chunked meaningfully, the student is able to read a fictional text and identify key elements of the story.” Other examples of adding scaffolds to the descriptors include: “A proficient student can comprehend the main message within segmented grade-level text. With suggested reading strategies or graphic organizers, students are able to generate and/or answer inferential questions.” These statements only differ

² Note that while it is also possible to start with the alternate achievement level descriptors and modify them to make them more difficult, this approach may be more challenging as many alternate achievement standards do not cover all content standards and are often based on extended content standards rather than grade-level content standards.

from grade-level descriptors through the addition of the scaffolds. Note that it is important to ensure that these scaffolds are included in the test design if they are included in the descriptor. Furthermore, these scaffolds will only be helpful on the test to the extent that they have been used during instruction. (See Pugalee & Rickelman, Chapter 5, this volume for more information on scaffolding.)

Other strategies for modifying descriptors apply if those designing the test believe that the knowledge and skills required of these students should be different. First of all, under the current federal regulations “different” can only mean less difficult. There are several ways to make grade-level achievement standards less difficult. One option is to focus on the cognitive complexity of the requirement and reduce it appropriately. For instance, a grade-level descriptor at grade 8 may state that a student can “evaluate algebraic expressions” while the modified descriptor could require the student to “identify algebraic expressions.” Likewise, if the grade-level descriptor says a student can solve “two-step problems”, a possible modification is to require students to solve “one-step problems.” For English Language Arts, we can reduce the complexity either by reducing the depth of knowledge required (e.g., move from analyze to describe) or qualify broader statements of knowledge. For instance, if the grade-level standard requires students to identify various parts of speech, including “nouns, verbs, pronouns, adjectives, adverbs, conjunctions, and interjections,” one could modify that standard by reducing the number of parts of speech required by removing the requirement of identifying conjunctions and interjections. Both standards require students to identify broader parts of speech, but the modified standard reduces the difficulty by only requiring students to identify simpler parts of speech. These modifications to the descriptors make the achievement standards less difficult to reach by reducing cognitive complexity, which complies with federal regulations as long as the depth and breadth of the assessment itself remains similar to the general assessment.

In practice, those drafting the modified achievement level descriptors could choose to adopt more than one of these strategies. That is, they could choose to reduce the depth of

knowledge required for proficiency on some of the skills, add scaffolds to the statements about other skills, and provide specific examples to others indicating that the student is required to perform a narrower range of the tasks than what is required in the grade-level standards, as long as that narrower range still matches the content standards and indicators. Specifics on writing achievement level descriptors will be discussed in the next section.

Procedures for Drafting Modified Achievement Level Descriptors

Regardless of the type of assessment, it is usually preferable to start considering achievement level descriptors early in the test development process. In the case of the AA-MAS (and all assessments developed under NCLB), the most important distinction is between achieving “proficient” and not, so a strong understanding of proficiency is needed. By considering this early, test developers can start an iterative process of using the descriptors to help design the test and then refining the descriptors as needed to match the final test blueprint.³ When the descriptors are used to drive the test design, test developers can ensure that the test blueprint supports the desired judgments and that the items themselves provide opportunities for students to show what they know and can do relative to the achievement standards. Consideration can be given to distinguishing items that would likely be answered correctly by students who met the definition for proficiency and incorrectly by those who did not.

Our recommended approach for drafting achievement level descriptors is to involve a committee of people who know the content and the students. However, they will always need some direction from the policymakers regarding the intent of the assessment program. In the case of modified achievement level descriptors, the committee will typically include both special education teachers as well as content specialists for each subject area (e.g., reading and mathematics). Content specialists could be subject area teachers, curriculum supervisors, or

³ If the test developer is shortening the test by eliminating the most difficult items proportional to the content standards, then the descriptors will be considered after the test is administered. This process will be discussed further in this section.

members of the general public with a specialty in that subject (e.g., a mathematician). Approximately 5–8 participants are needed per subject area, but if descriptors are being developed for multiple grade levels, consider inviting more participants and splitting them into teams. That is, a group of eight participants can write the descriptors for grade 4, and then they can separate into two groups of four with one group working on grade 3 and the other on grade 5.

The direction required from the policymakers will include the assumptions made about the population, including who the students are and what the barriers are to their ability to achieve grade-level proficiency on the general assessments. The committee members will also need to understand the theory behind the revisions and enhancements made to the assessment as well as see examples of those revisions and enhancements. They also need to understand the type of modifications that will NOT be permitted, such as providing below grade-level passages on a reading assessment. In addition, if any data analyses have been done such as those suggested in Quenemoen (Chapter 2, this volume), the committee could be informed by concrete information about what was learned from these analyses, including specific examples of items this population seemed to perform well on and those they did not.

Once the committee members have sufficient background, the real work drafting the descriptors begins. The majority of those developing modified achievement level descriptors are starting from the grade-level descriptors and editing them rather than starting new descriptors from scratch. Regardless of the approach taken, Perie, Hess, and Gong (2008) recommend that the committee discuss several issues, including:

- Interactions of process and content (e.g., is this a routine application of skill or transfer of known skill to a new context?)
- How students move both across performance levels and across grade levels.

- Are the knowledge and skills required of Proficient on the modified achievement standard the same as on the general assessment, but some scaffolding is needed, or are the knowledge and skills different?
- If they are different, is the content different or the processes? (e.g., both can make inferences at the Proficient level but the grade-level achievement standards require that the inferences are made in a more complex context than the modified achievement standards, or grade-level achievement standards require students to make inferences, while modified achievement standards require students to only draw basic conclusions from concepts presented directly)
- How do you see students moving across grades? For example, how does Proficient in one grade compare to Proficient in the next?
- Transition from this assessment to the general assessment – how are they linked?
 - Should the proficient level of the modified achievement standards be an indicator of readiness for achievement on grade-level standards?
 - Should the state adopt a policy regarding the modified achievement standards, such as students who score at the advanced level on the modified achievement standards must take the general grade-level test the following year?

Given their answers to these questions and the theories regarding appropriate revisions to the test design, the committees can then draft descriptors. Recall from the previous section that some of the modifications could include: (1) reducing the cognitive complexity of the required skill, (2) decreasing the number of elements required, or (3) adding appropriate supports and scaffolds to the description of the knowledge and skills required. The following is an example of a fifth-grade reading descriptor for a general assessment and the modified version that includes all three types of modifications.

Grade-Level Descriptor

Proficient students comprehend the message within grade-level text. Using supporting details, they are able to analyze information from the text and summarize main ideas. Before, during, and after reading, students generate and/or answer questions at the literal, inferential, interpretive, and critical levels. Students interpret and use organizational patterns in text, (e.g., description, cause/effect, compare/contrast, fact/opinion) in order to gain meaning. They use informational text features (e.g., index, maps, graphs, headings) to locate information and aid in comprehension. Students are able to identify and analyze elements of narrative text (e.g., characters, setting, and plot). Additionally, Level II students can identify author's purpose and recognize how author's perspective influences the text.

Modified Descriptor

Proficient students comprehend the message within segmented grade-level text. Students will be able to identify the main idea and retell information from the passage with supports (e.g. a web, 5 W's chart, T chart), when appropriate. During and after reading, students are able to generate and/or answer questions at a literal level. Students identify and use organizational patterns in text (e.g., sequence, compare/contrast, fact/opinion) in order to gain meaning. They use informational text features (e.g., index, maps, graphs, charts) to locate information and aid in comprehension. When given supports (e.g., story maps, character web, illustrations), students are able to identify basic elements of narrative text (characters, setting, beginning/middle/end). Additionally, Level II students identify author's purpose when given the definitions.

Having similar structure between the grade-level and modified descriptors helps teachers, administrators, and parents see the difference between grade-level proficiency and modified proficiency, providing useful information in what it takes to move a student from the alternate assessment based on modified achievement standards to the general assessment based on grade-level achievement standards.

Earlier, a different approach to modifying descriptors was introduced, following from the suggestion by Welch and Dunbar (Chapter 6, this volume) that an AA-MAS could be designed by eliminating the most difficult item proportional to the content standards. This approach would result in an AA-MAS that was similar to the general assessment in scope but shorter. The two assessments could be statistically linked together since there are common items across both populations. Then, the cut score could be mapped directly onto the AA-MAS. As discussed by Welch and Dunbar (*ibid*), a standards validation will need to be conducted to ensure the cut scores divide student performance meaningfully into the achievement levels. Once the cut scores have been validated, the grade-level descriptors can be modified by taking into consideration the items that map to each achievement level. Different rules have been used to identify items within each level, usually focusing on the likelihood that a student within that level would answer the item correctly compared to the likelihood of a student below that level answering the item correctly. Items that are distinct between these two groups are identified as mapping to that level. Then, content experts can summarize the types of knowledge and skills represented by those items and use those summaries to write descriptors. This approach focuses solely on the item specifications, as scaffolds have not been used in this test design. However, caution must be taken to avoid writing descriptors that are too specific to one test form. In addition, there would still need to be a guiding philosophy driving this approach, including defining proficiency. The philosophy should relate to our understanding of how reducing difficulty in this manner addresses some of the concerns about the cognitive processing of low achievers discussed by Pellegrino (Chapter 4, this volume).

Regardless of what approach to writing modified descriptors is taken, the articulation across grades should be considered. Often when committees are working on drafting modified achievement level descriptors, they are split into smaller groups to work on specific grade levels. If this occurs, it will be important to spend time at the end of the workshop examining the descriptors across all grades. Articulation will be improved if the committee members are asked to consider whether they can see a clear progression across levels and how well these descriptors translate to instruction.

Once the modified achievement level descriptors have been drafted, they will need to be finalized by the state department of education and then approved by the state policymaker (typically a board of education). When the state department of education is reviewing the draft descriptors, they typically consider them as a whole, analyzing the consistency in rigor across grades and subjects, the natural progression of difficulty from one grade to the next, and the alignment between the descriptors and the test blueprints.

Setting Cut Scores

At first glance, it appears that any standard setting method that a state uses for its general assessment would work for the modified assessment, particularly since most states appear to be starting with their general assessment and applying various types of modifications. However, there are additional considerations that come into play when selecting an appropriate method for setting cut scores.

Keeping in mind that there may be some state policymakers who choose to develop a brand-new assessment or to modify their AA-GLAS, we will start with the scenario that a state has modified the general assessment. Almost all state general assessments are comprised primarily of multiple-choice items with some states choosing to include some open-ended items as well. With these types of tests, a test-based approach to standard setting is typically used. Test-based approaches are those where the judgments are made about the test itself—usually

about individual items—rather than about the students or their actual performance. Another way to think about the type of methods is based on the type of judgment required. According to Zieky, Perie, & Livingston (2008), there are four types of standard setting judgments: (1) judgments of test questions, (2) judgments of profiles of scores, (3) judgments of people or products, and (4) judgments of groups of people. Examples of judgments of test questions include methods such as Angoff or Bookmark. Methods involving judgment of profiles or scores include Dominant Profile or the Performance Profile Method. Methods that require judgments of people or products include Contrasting Groups and Body of Work. Methods that involve judgments of groups of people are rarely used in the educational context.

While any of the three prominent types of judgments could apply to an AA-MAS, the methods most appropriate for a test that is primarily comprised of multiple-choice items with a few (or no) open-ended items include judgments of test items. This section will focus on the two most common test-based methods—Angoff and Bookmark—and then discuss the feasibility of using methods based on judgments of profiles, people, or products.

Test-Based Approach

Test-based approaches typically require standard-setting committees to make judgments about test items. The two most commonly used methods for K–12 educational assessments are the modified Angoff method and item mapping, typically the Bookmark method. The applications of these two methods to set cut scores on the AA-MAS will be discussed in this section.

Modified Angoff. The modified Angoff method (Angoff, 1971) is probably the most widely used and best researched standard-setting method. In it, participants are asked to state the probability that a borderline test taker (e.g., someone who is just barely proficient) would answer each test item correctly. Summing the probabilities across all test items provides the test score for a borderline test taker, which becomes the cut score for that achievement level. Typically, for a

multiple-choice test with four response options, we recommend that panelists limit their judgments of probability to a range of 0.25 to 0.95. The reasoning is that even if the student has minimal ability to answer the item correctly, he will have a 25% probability of answering it correctly by chance (1-in-4). We limit the upper end to show that we never expect perfection from a student. The only exception that panelists are given is if they think that one distractor will be so appealing to a student with minimal knowledge that he is likely to be drawn to that distractor to the point that he has a less than 25% chance of answering the item correctly, then they can provide a rating below 0.25.

Now consider an AA-MAS where the revisions have included reducing the answer options from four to three. In this situation, the student has a 1-in-3 (33.3%) probability of answering the item correctly by chance, further restricting the range of possible judgments to 0.35–0.95. This adjustment will almost certainly result in a higher cut score, which may not be desirable.

Another option for states wanting to stick to a modified Angoff approach is to use another modification of the Angoff method—the yes/no method (Impara & Plake, 1997). In this option, the judgment would be a simpler yes/no that the borderline test taker either would or would not answer this item correctly. There have been some concerns raised that the yes/no method rounds judgments too inaccurately (c.f. Reckase, 2006; Zieky, et al, 2008). For instance, a panelist who feels that a borderline test taker who has a 25% chance of answering the item correctly would record a 0. He would also likely record a 0 for an item he thought the borderline test taker had a 45% probability of answering correctly and another 0 for an item he thought a borderline test taker had a 40%% chance of answering correctly, resulting in a cut score of 0 out of 3, whereas the traditional Angoff would calculate a cut score of 1 out of 3. Thus, it would be reasonable to consider adding in a guessing factor.

For example, if on a 50-item test a group of panelists agrees that the borderline Proficient student would answer 23 items correctly, then the unadjusted raw cut score would be 23 out of 50 points. However, to adjust for guessing, we could then assume that of the remaining 27 items that

the student does not have the ability to answer correctly, they would answer 1/3 of them correctly by guessing (assuming 3-option answer choices). Therefore, they would answer 23 items correctly through their ability and 9 items correctly by chance, making the adjusted cut score 32 points out of 50.⁴ This raw score cut can then be transformed to a scale score cut if desired.

Note that no change would be needed for applying an Angoff methodology to an open-ended item on an AA-MAS. The method most commonly used in K-12 assessments for the open-ended items is the mean estimate method, where the panelists estimate the mean (or average) score a roomful of 100 borderline test takers would achieve. Those averages are then added to the probabilities for the multiple-choice items (which are, in fact, averages of 0/1 scores) or to the sum of 0s and 1s. Modification should not affect a panelist's ability to make this type of judgment and no adjustment for guessing would be needed.

Item Mapping. Item Mapping approaches include Item Descriptor Matching (Ferrara, Perie, & Johnson, 2008) and the more commonly used Bookmark method (Mitzel et al., 2001). The Bookmark method was developed to be used with tests that are scored using Item Response Theory (IRT). It is now one of the most widely used cut score-setting methods for state K–12 assessments. To use this method as it was designed, the state will need a test that was calibrated using IRT and be able to order the items from easiest to most difficult based on the calibrations. The panelist uses an “Ordered Item Booklet” that displays the questions in order of difficulty from easy to hard and is asked to place a bookmark at the spot that separates the test items into two groups—a group of easier items that the borderline test taker would probably answer correctly (with a response probability of .67, meaning a chance of at least 2 out of 3 or .67), and a group of harder items that the borderline test taker would probably not answer correctly (i.e.,

⁴ This adjustment could result in a cut score higher than the panelist intended if they are not confident in their judgment of the 1s. They should be instructed to record a 1 only if they feel the borderline test taker would have a strong probability of answering this item correctly. Another option would be to substitute the 1s and 0s with probabilities before summing the judgments to calculate a cut score. For instance, the 0s could be transformed to 0.33 and the 1s could be transformed to 0.95.

the test taker would have a probability of less than .67 of answering correctly). The bookmark placement is then translated to an ability level of a student who has at least a .67 probability of answer the items before the bookmark correctly and a less than .67 probability of answering correctly the items after the bookmark. That ability level (or theta value) can be translated to a scale score and mapped back to a raw score.

A concern with using this (or any item-mapping) method on an AA-MAS is in the item ordering. Typically, an ordered-item booklet reflects a large population of students with a wide degree of variance in their abilities. While there may be some “distance” in the associated theta values at extreme ends of the booklet, the majority of items are close enough together that it is a fairly simple transformation to map a bookmark placement to an ability score. However, some states have experienced difficulties with an ordered item booklet of a AA-MAS, where there was not as much variation among test takers resulting in some clumping of item difficulties and areas with large gaps in ability scores between the clumps.

For instance, let's suppose that in a traditional Bookmark item map, items 10–16 have associated theta values of 1.02, 1.04, 1.05, 1.05, 1.07, 1.08, and 1.10. Although there are different methods for selecting the actual cut point (theta value of the item that is bookmarked, the theta value of the item before it, or the mean of those two values), it is relatively straightforward to determine the cut score value for a bookmark that is placed at any of those items. But what if the items had theta values of 1.02, 1.02, 1.03, 1.42, 1.42, 1.43, and 1.67? If the bookmark is placed on item 13 (the fourth value in the string) indicating that the 13th item is the first one that a borderline test taker would not have a 0.67 probability of answering correctly, what should the cut score be? Given the three methods usually used to determine the cut score, this one cut score could be assigned a value of 1.42, 1.03, or 1.225. These are fairly disparate numbers and could result in very different scale score or raw score cuts.

Therefore, before choosing to use an item-mapping approach, it is important to consider the size and variance of the population taking the AA-MAS. That is, be sure that there are enough students taking the test and enough variance in that population of students for the items to both scale well and order sensibly. Theoretically, it may be more feasible for a state the size of Texas to use an item-mapping approach to set cut scores on the AA-MAS than a state the size of Delaware.

An alternative for states who are worried that their samples are too small or too homogeneous is to vary a traditional item-mapping approach using classical measurement theory rather than IRT. Actually, the process described here is similar to a yes/no Angoff except that the items are ordered by difficulty, as in a traditional item-mapping approach.

The approach involves ordering the items and placing them into an ordered item booklet, as in the Bookmark approach; however, p-values rather than IRT difficulties are used to determine the order. Then, ask the panelists to start with the easiest item and simply ask “would a borderline Proficient student be able to answer this item correctly?” If the answer is yes, then they move to the next item. When they reach an item that they answer “no” to, that is where they place their bookmark. As with all Bookmark procedures, we recommend that the panelists continue a little further into the booklet to ensure that the bookmarked item is truly the beginning of the more difficult items and not an anomaly. Then, rather than transforming the bookmark to a difficulty estimate, simply count the number of items before the bookmark and use that number as the initial raw score cut. For instance, in a 50-item booklet, if the panelist places their bookmark on item 22, then the initial cut score would be set at 22 out of 50 raw score points. Again, it is worth adjusting this cut score for guessing. If this booklet contained only multiple-choice items with 4-option answers, then a borderline test taker would have a 1-in-4 chance of

answering the remaining 28 items correctly by guessing. So, we would add 7 raw score points to the cut score for a final cut score of 29 out of 50 points.⁵

Other Standard-Setting Approaches

As mentioned earlier, there are other standard-setting approaches that may be worth considering, particularly if the test design includes more than multiple-choice items. At least one state is developing an AA-MAS that involves collecting student evidence on each content standard assessed. The result will look more like a portfolio assessment than a traditional paper-and-pencil assessment. So, it is important to consider other standard-setting methods for these alternate approaches. Three methods we will discuss here are the Body of Work (Kingston, Kahl, Sweeney, & Bay, 2001), Analytic Judgment (Plake & Hambleton, 2001), Dominant Profile (Plake, Hambleton, & Jaeger, 1997), and Contrasting Group (Livingston & Zieky, 1982) methods.

Body of Work. The Body of Work method falls under the category of judgments of people and products and requires some type of evidence for the panelists to consider. Zieky, et al. (2008) lists this as a type of Contrasting Groups approach that focuses on categorizing student work rather than the students themselves. The method is designed for tests with performance tasks or tasks that yield observable products of a student's work, such as essays or recorded speech or science experiments. This is a popular method for the portfolios often used for the AA-AAS. It would also be suitable for a design that requires students to submit evidence of achievement for each assessed content standard. The method does not work well for tests that

⁵ Note that if the booklet contained open-ended items, they could not be answered correctly by chance and would not be figured into the adjustment. For instance, if 8 of the 28 remaining "items" in the booklet represented various point values for open-ended items, we would simply calculate the probability of guessing correctly on the 20 multiple-choice items, adding 5 points to the initial raw score cut.

include large numbers of multiple-choice questions, but it will work if there are a few multiple-choice questions with the performance tasks.

The panelists are asked to review a full body of evidence (meaning the responses to all test questions) and make a single judgment about the entire set of responses, matching the knowledge and skills exhibited in the responses to the knowledge and skills required to be in an achievement level. The cut score between two performance levels is chosen by finding the point on the score scale that best distinguishes between the sets of evidence placed in each of the achievement levels.

Analytic Judgment. The Analytic Judgment method is also a method where judgments are made on products; however, judgments are made on responses to individual items (or groups of related items) rather than on the product as a whole. It was designed to be used with tests made of several essay or performance tasks. The method will work for tests that include some multiple-choice items with the performance tasks as long as the items can be grouped into meaningful content clusters.

The Analytic Judgment method begins by asking panelists to review samples of test takers' work. As described in Zieky, et al. (2008), it is similar to the Body of Work method, but there are two distinct differences:

1. Panelists make judgments on test takers' responses to individual items or to clusters of related items rather than to the entire body of evidence at once, and
2. In addition to classifying a response into an achievement level, panelists further classify the responses at each performance level into *low*, *middle*, and *high* categories. For example, a response is not simply classified as Proficient. It is, in addition, classified as low Proficient, middle Proficient, or high Proficient.

The result is a cut score for each item or group of related items; this cut score is the score that most clearly distinguishes between the best responses in the lower achievement level and the worst responses in the higher achievement level (e.g., between responses classified as

high Basic and low Proficient.) Those are the responses that are close to the borderline of each achievement level. The cut scores for all items or all groups of items are summed to get the cut score for the total test.

Dominant Profile. The Dominant Profile is a method based on profiles of scores and typically results in a conjunctive cut score. That is, the test is divided into meaningful parts that measure different knowledge and skills, and a cut score is determined for each part separately. Thus the outcome is not a single cut score but a set of rules. Those rules can specify separate cut scores for each content strand, or there can be a single cut score for the total score with a minimum score on certain components.

The panelists' task is to become familiar with the test, how it is scored, and with the meanings of the different strands/components. They then work together to specify rules for determining which combinations of scores represent acceptable performance and which do not. The rules can combine information from the scores of different components in various ways, as in the following example:

A mathematics test is divided into 5 strands with 20 points per strand. The panelists determine the follow set of rules to be used before classifying a student as Proficient:

- No score below 10 on any component
- At least one score of 15 or higher
- A total score of at least 60 points

Contrasting Groups. Finally, what if a test developer is in a position where cut scores need to be set, but the data are not yet available, and there is no rubric or student work to analyze? The original contrasting groups method involves judgments about test takers (Zieky & Livingston, 1982). The judgments can be made prior to the test administration and then compared to the actual scores received to calculate the cut score. The method involves identifying teachers familiar with the target population and then training them on the meaning of the achievement level descriptors, paying particular attention to differentiating between high

performance on the lower level and low performance on the higher level. This training does not have to be done in person (videotapes work well), as the method typically works best when there are large numbers of teachers involved (at least 100 per cut score). Once the teachers have been trained, the test developer asks them to place each of their students who will be taking the AA-MAS into one of the achievement levels based on their experience with those students. Once the students have taken the test, they are assigned a total score (either a raw score or a scale score will work for this method). Then, the distribution of scores across assigned achievement levels can be examined to determine the best cut score for each level. For instance, for each cut score, the percentage of students scoring at the higher of the two levels can be plotted against the score. That is, for the basic/proficient cut, plot a graph with the range in total cut scores along the x-axis, and the percentage of students at each of those levels categorized as Proficient by their teachers on the y-axis. Then, choose the cut score for proficient based on the percentages. Zieky, et al. (2008) recommends that “one reasonable choice for a cutscore would be the score at which 50 percent of the test takers are [categorized as] Proficient because that would represent the borderline of the Proficient performance level (page 78).” Another procedure is to plot the distributions of scores for two adjacent levels (e.g., basic and proficient) and set the cut score at the point at which two distributions overlap.

Because this method is based on the judgments of teachers about students they know, it is a reasonable way to match students to achievement levels, but it also introduces some bias. Teachers may factor other considerations into their judgments, such as effort and likability, when the judgment should truly be about the student’s knowledge, skills, and ability. This method is often used to check a cut score set through a method based on judgments of test items. This check can be done a couple of years after the initial standard-setting workshop once teachers have become very familiar with both the test and the meaning of the achievement levels.

Linking Tests through Cut Scores

A final option for consideration is linking the AA-MAS to the general assessment through the cut scores. Although this idea will be discussed more thoroughly in Abedi (Chapter 9, this volume), it is worth introducing here. Some state policymakers have suggested linking the “advanced” or “proficient” level of the modified achievement levels to the “basic” level of the general grade-level achievement levels. One option would be to link the assessments statistically with common items taken by both populations (as described by Welch & Dunbar, Chapter 6, this volume), but another option is to link the assessments judgmentally.

A judgmental linking is where a standard setting method is applied to make the “advanced” level of one test equivalent to the “basic” level of another. There are several ways to do this, but the best is to use many of the same panelists in both standard settings. Start by having the panelists become thoroughly familiar with the “basic” level of the general assessment, both by reviewing the grade-level achievement level descriptor and by examining exemplar items and/or student work at that level. Then, the modified achievement level descriptor for advanced (or proficient) would need to be matched to the grade-level achievement level descriptor for basic. Preferably, the descriptors would be exactly the same, with only slight modifications to allow for the use of the scaffolds that may have been built into the assessment. The judgmental task most commonly used is an item mapping approach where the panelists would work through an ordered item booklet to find the cut score that would allow for the same interpretation of knowledge and skills across the two assessments.

Final Considerations

Although the greatest challenges for developing modified achievement standards lie in defining proficiency for this population and applying an appropriate standard-setting methodology to set a cut score, we would be remiss if we did not discuss the importance of documentation and validity studies. Proper documentation is important for any testing program

and mandated by the peer review guidance. Likewise, we should always be thinking of the validity of the interpretations made using the achievement standards, and peer review requires plans for validating the assessment and the inferences made from the results.

Documentation

It is important to document both the process of developing modified achievement level descriptors and the standard-setting procedures. Two professional *Standards* (AERA, APA, & NCME, 1999) directly address the importance of documenting the rationale, procedures, and results:

- Document the PLDs, selection of panelists, training provided, ratings, and variance measures (Standard 1.7)
- Document the rationale and procedures for the methodology used (Standard 4.19)

As discussed in Perie (2007), there are eight important components that need to be documented regarding the standard-setting process:

1. Achievement level descriptors
2. Panelists
3. Rationale
4. Training
5. Procedures
6. Ratings and variance
7. Any adjustments and adoption of cut scores
8. Validity evaluation

Most of these are fairly straightforward and discussed in several texts on standard setting. Here, we will highlight only two areas that may have particular sensitivities for modified achievement standards and have been discussed within this chapter.

Achievement Level Descriptors. Because of the challenges associated with describing proficiency using a modified achievement standard, it is vital that the test developer both describe and justify the process used, including the selection of participants who may have drafted the descriptors, the directions given to them, the data or information used to inform the process, and the number and type of reviews conducted before the descriptors were formally adopted. Providing a theory of who the students and what the barriers are to their achieving grade-level standards will aid the understanding of how the descriptors were developed.

Rationale. It will be important to document the rationale for selecting the standard setting method used to set the cut scores. If the revisions or enhancements made to the assessments (e.g., the reduction of a response option) or the characteristics of the population (e.g., small variance in performance) affected the choice of available methods, this could be explained in writing to better help a reader understand the purpose and logic. Explaining the rationale behind any selection of a process helps inform the validity argument as discussed in the next section and in Marion (Chapter 8, this volume). Finally, if any modifications to the traditional application of the standard-setting method—such as those described in this chapter—were made, these need to be documented as well along with the rationale for these modifications.

Validation

Validity is a large topic that will be covered more completely in Marion (Chapter 8, this volume), but it is worth touching on the various types of evidence that can be collected during the standard-setting process here. Collecting the information discussed in the documentation section can provide evidence of internal validity of the achievement standards. Providing a rationale for the methods used, ensuring an appropriate panel composition, comparing the results to other external sources can all provide validity evidence to the argument that the achievement standards were set appropriately. Then, thought needs to be given to how the

interpretation and use of the achievement standards contribute to the consequential validity of the assessment.

In examining the validity of the use of the achievement standards, it is important to ask a series of questions about the basic components of those standards. Conducting a series of studies over the first several years of the assessment can provide information to answer these questions on issues regarding appropriateness of the modified achievement level descriptors and the accuracy of the cut scores. For example, questions may include:

- Was the standard-setting procedure internally valid?
- Do the cut scores divide students reasonably in terms of achievement?
- How well does the test classify students compared to their achievement in the classroom?
- Do the effects of the achievement standards match what was intended?
- Have the modified achievement level descriptors had an impact on instruction?
- Have there been any negative consequences of using these achievement standards?

Some of these questions can be answered through the standard-setting process itself. It will be important to show that the panelists were qualified and representative of all possible panelists. Evaluation forms can be used to show that the panelists understood the process and were confident in the results. If feasible, working with two separate panels during the standard-setting process will also provide a measure of consistency in cut score recommendations and provide evidence of validity. To argue for the reasonableness of the cut scores, the test developer can compare the percentage of students categorized into each achievement level by the AA-MAS to the percentages in the equivalent categories by the general assessment and the AA-AAS. If all tests are intended to be developed to the same rigor for their specific populations, then one would expect the impact data to be distributed similarly across all assessments.

Other questions can be answered through teacher surveys and focus groups as well as classroom observations. Conducting a contrasting groups study a year after the cut scores are

set can also provide useful interpretative information. Once the test developers are confident that teachers know and understand the modified achievement level descriptors, they could ask the teachers to classify their students into one of the four achievement levels prior to the assessment. Then the classifications determined by the assessment could be compared to the teacher classifications to see if the teachers would generally assign students into higher or lower categories or if the two sources of data provide similar classifications.

As a final thought, it is important to keep in mind that the process of setting achievement standards does not end with the cut score study or even with State Board approval of the descriptors and cut scores. Instead, consider designing a mechanism within an assessment program to continually monitor the effectiveness and appropriateness of the achievement level descriptors and the usefulness of the categories as defined by the cut scores. Particularly for this population where we expect instruction to continually improve and move closer to grade-level instruction, it is important to frequently monitor the efficacy of the modified achievement level standards.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational measurement* (second edition, pp. 508–600). Washington, DC: American Council on Education.
- Beck, M. (2003, April). Standard setting: If it is science, it's sociology and linguistics, not psychometrics. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, 9 (3), 215–235.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Ferrara, S., Perie, M., Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure. *Journal of Applied Testing Technology*, 9(1).
- Gong, B. (2007). Learning Progressions: Sources and Implications for Assessment. Presentation at the CCSSO Large-Scale Assessment Conference, Nashville, TN, June 2007.
- Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34(4), 353–366.
- Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Livingston, S., & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Perie, M., (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice*, 27(4) pp. 15-29.
- Perie, M. (2007). *Setting alternate achievement standards*. Lexington, KY: University of Kentucky, Human Development Institute, National Alternate Assessment Center. Available online at: <http://www.naacpartners.org/products/whitePapers/18020.pdf>

- Perie, M., Hess, K., & Gong, B. (2008). *Writing Performance Level Descriptors: Applying Lessons Learned from the General Assessment to the 1% and 2% Assessments*. Dover, NH: National Center for the Improvement of Educational Assessment. Available at www.nciea.org.
- Plake, B. S., & Hambleton, R. K. (2001). The Analytic Judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Erlbaum.
- 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.
- Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25 (2), 4–18.
- U.S. Department of Education. (2007). *Final Rule 34 CFR Parts 200 and 300: Title I—Improving the academic achievement of the disadvantaged; Individuals with disabilities education act (IDEA)*. Federal Register. 72(67), Washington DC: Author. Available at: <http://cehd.umn.edu/NCEO/2percentReg/Federal-RegApril9TwoPercent.pdf>
- Zieky, M., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.