A Guide to Understanding and Developing Performance-Level Descriptors

Marianne Perie, *National Center for the Improvement of Educational Assessment*

There has been much discussion recently about why the percentage of students scoring Proficient or above varies as much as it does on state assessments across the country. However, most of these discussions center on the leniency or rigor of the cut score. Yet, the cut score is developed in a standard-setting process that depends heavily on the definition for each level of performance. Good performance-level descriptors (PLDs) *can be the foundation of an assessment program, driving everything from item development to cut scores to reporting.* PLDs *should be written using a multistep process. First, policymakers determine the number and names of the levels. Next, they develop policy definitions specifying the level of rigor intended by each level, regardless of the grade or subject to which it is applied. Finally, content experts and education leaders should supplement these policy definitions with specific statements related to the content standards for each assessment. This article describes a process for developing* PLDs*, contrasts that with current state practice, and discusses the implication for interpreting the word "proficient," which is the keystone of No Child Left Behind.*

Keywords: performance level descriptors, standard setting, No Child Left Behind

In discussions of cut scores, or the minimum score value required of a performance level, the performance-level descriptors (PLDs) are often absent from the conversation. Yet, they are crucial to determining where the cut scores are set. PLDs describe the level of knowledge and skills required of each performance level. They are receiving increased attention under the *No Child Left Behind* Act of 2001 (NCLB), as every assessment developed under this act must include a minimum of three performance levels, with the focus on proficient. The standards and assessment guidance also indicates that PLDs need to be written prior to and used for standard setting (U.S. Department of Education, 2004). Moreover, many researchers argue that the descriptors should be written early in the test development process and be used in developing test blueprints and item specifications (see Bejar, Braun, & Tannenbaum, 2006).

In fact, the PLDs are of such influence that, in a well-run standard-setting workshop, they determine the rigor of the performance and thus the placement of the cut score. Many of us 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).

Recently, many questions have been raised about why the percentage of students scoring at proficient or above varies as much as it does on state assessments across the country. Some research shows that cut scores appear very lenient in some states and stringent in others, particularly when compared to the cut scores on the National Assessment of Educational Progress (Braun & Qian, 2007; McLaughlin, 2006; McLaughlin & Bandeira de Mello, 2005). However, any analysis of cut scores without consideration of the PLDs used in determining the cut scores provides an incomplete picture. Differences in the percentages of students scoring at or above proficient are due to differences in content standards, tests, and, importantly, the definition of proficient—all of which contribute to the cut score. PLDs are the foundation of standard-setting activities as they provide the explanation of how student achievement differs from one level to the next. Some of us argue that the fact that each state writes its own definition of proficiency is a primary reason for the amount of variability in the percentage of students scoring at or above proficient.

> Saying that all students must be at the proficient level or above by 2014, but leaving the definition of proficient achievement to the states has resulted in so much state-to-state variability in the level of achievement required to meet the proficient standard that "proficient" has become a meaningless designation. (Linn, 2005, p. 14)

This article provides background and examples of PLDs. It describes a method for developing descriptors and compares and contrasts descriptors currently used by states to define Proficient for their NCLB tests. Most importantly, however, this article discusses how these differences can impact the

Marianne Perie is a Senior Associate at the National Center for the Improvement of Educational Assessment, P.O. Box 351, Dover, NH 03821; mperie@nciea.org.

interpretation of the percentage of students meeting proficient under NCLB. The first section provides background context for the use of performance levels under NCLB. The second section describes the best practices for developing PLDs, including considerations in terms of the number and names of descriptions, the importance of developing a generic definition of each level that can be applied across grades and subjects, and the process for writing the full PLDs that links them directly to the test specifications. The third section compares the best practices to the current practices and describes variances in state approaches to this task. Finally, the article ends with a discussion of some of the implications for interpreting the success of NCLB when states have large differences in meaning of "proficient" performance.

Background

A performance standard, also referred to as an achievement standard, consists of three components: the name of the level, a written description of the level, and a minimum cutoff score. Determining the number and names of the levels should be done by policymakers in consideration of how the results will be used and the amount of distinction between levels required. Defining the levels themselves should also be done with careful consideration of the purpose of classifying students into different levels.

Under the *No Child Left Behind* (NCLB) Act, states must formally adopt achievement descriptors for each grade and subject. The federal government requires states to develop a minimum of three levels, with one level designated as the proficient performance expected of all students and at least one level each above and below proficient. While flexibility is given in naming the levels and in the number over three allowed, the peer review guidance specifies that states must develop descriptions of the competencies associated with each level and cut scores to differentiate among the levels. As stated in Section 2 of the guidance, "the State's academic achievement standards must include descriptions of the content-based competencies associated with each level." (U.S. Department of Education, 2004)

While these academic achievement standard descriptions, which will be referred to as PLDs for the remainder of the article, are most commonly associated with setting cut scores, they are also a useful development and reporting tool. The descriptors state in words what the cut scores mean and can help teachers and parents interpret what their students know and can do and, potentially, what they do not know and cannot do.

Ideally, if an assessment program has clearly defined levels and purposes for using those levels, the levels should be designated early in the test development process. It should be clear to those designing and using the assessment what levels of performance policymakers want to report on and how those levels are distinguished from one another in terms of knowledge and skills. That way, items can be written to clearly distinguish among the levels and ensure a more reliable categorization of student abilities. However, because PLDs are most commonly associated with setting cut scores, they often are written just prior to a standard-setting workshop or even as part of a standard-setting workshop, but after items have been developed and administered. Then, panelists must consider those PLDs when determining the minimal level of performance required at each level. The panelist must balance what is intended by policy to be valued by the assessment with which items actually appear in the assessment. It is more likely that these two considerations will be aligned if the PLDs were developed early and taken into consideration during the item writing process. Otherwise, the PLDs have little influence on item writing but have a large impact on the location of the cut score.

In some methodologies, the PLDs are not written until after the cut score is set, typically using an item-mapping or scale-anchoring approach to describe each level. In this scenario, the policy is determined by the items in the form on which the cut scores are set and by the cut scores themselves. Logically, the policy directive should come first, and then the items and cut scores should work to enact this policy. In fact, there is quite a body of literature discussing the necessity of developing descriptors that define the knowledge and skills of an examinee at the borderline of a performance level prior to setting a cut score. Livingston and Zieky (1982) considered the definition of "borderline" knowledge and skills as one of the five essential steps of any standard-setting study. They recommend having judges describe in their own words an examinee on the borderline of acceptable and unacceptable knowledge and skills and encourage consensus on this task. Impara, Giraud, and Plake (2000) conducted a study showing that cut scores are dependent on definition of borderline student. If the borderline student was not clearly defined, the panelists produced a wider range of recommended cut scores. Berk (1996) also discussed the importance of using PLDs in setting cut scores. Specifically, he focused on the need to provide explicit behavioral descriptions of each level, saying "the interpretation of the final cut scores hinge on the clarity of the behavioral definitions" (p. 224).

Operationally, however, states use different methods for developing PLDs and produce them at different points in the assessment process. Some even wait to generate full PLDs until after the cut scores have been set, using items students do and do not perform well on to define the state's policy. This practice seems contrary to the intent of PLDs in criterion-referenced testing, which is to define state policy on the knowledge and skills required of students to be categorized at a particular performance level.

Best Practice for Developing PLDs

Perhaps the best work on developing PLDs was described by those working on the National Assessment for Educational Progress (NAEP) achievement levels as summarized in Loomis and Bourque (2001). The guidance from that work led to the framework of developing PLDs by first specifying the numbers and names of the levels, next drafting policy definitions, and then fleshing out the policy definition with full descriptors for each subject and grade level. This is the approach that appears to be most valid and supported by research.

This approach involves different experts at different points in the PLD-writing process. Ultimately, these descriptors communicate both the policy behind the meanings of labels such as "proficient" as well as the content expectations for each subject and grade assessed. Therefore, both policymakers and content experts need to be involved in developing PLDs. The process described in this section divides the writing into steps that first require the policymakers to name the levels and then state in words the level of rigor intended by each name. Then, the content experts and educators apply

their knowledge of the grade-level content standards[1] to supplement these generic terms with subject-specific explanations appropriate to each grade level assessed.

Determining the Number and Names for the Performance Levels

The first decision is the number of performance levels to use. Ideally, policymakers should choose the fewest performance levels needed to fulfill their purpose. The goals for and use of the test should be considered in determining the number of performance levels needed. In many certification or licensure tests, only two levels are needed: Pass and Fail. NCLB requires state to develop at least three levels, one for proficient, one above and one below. However, a majority of states have four performance levels, allowing them to differentiate between students who are close to Proficient and those who are well below Proficient, often called Basic and Below Basic, respectively, in addition to those who are Proficient or Advanced.

Typically, no more than four levels are needed. Beyond this number, it becomes difficult to describe meaningful differences across more 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. If there is only one cut score (giving two performance levels), a good test developer can focus most of the test's measurement power and test information around that cut score. If there are two cut scores (giving three performance levels) the test developer has to split the available power and information across the two cut scores, and so forth. 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. Finally, the greater the number of performance levels, the greater the work required to produce PLDs and cut scores. The level of effort required can soon become unmanageable.

After determining the number of levels, the next task is to name the performance levels. The terms themselves carry meaning, even without further description; therefore, naming a level is the first step in defining performance. Some typical naming conventions include pass/fail, below basic/basic/proficient/advanced, does not meet standards/partially meets standards/meets standards/exceeds standards. The words chosen express the values of the policymakers and thus should be selected carefully.

Beck (2003) indicated that naming conventions should be developed as the first step in defining performance. He recommends *avoiding* the following types of terms when naming performance levels:

1. Nebulous, unclear, or unreasonable terms or oxymorons (needs improvement, reasonable mastery);
2. Normative terms (average, typical);
3. Moving terms (nearly X, approaching the standard, emerging, progressing) as they apply to all parts of the level, making it more difficult to distinguish borderline performance;
4. Noneducational terms (normal, inadequate, novice/apprentice); and
5. Nonparallel terms (outstanding, pass, warning).

Although there are no clear-cut guidelines on how to develop names for performance levels, Cizek and Bunch (2007) recommend that they be "thoughtfully chosen to relate to the purpose of the assessment, to the construct assessed, and to the intended, supportable inferences arising from the classifications." Zieky, Perie, and Livingston (2008) offer this recommendation: "If you have the option, consider the use of neutral labels, such as Level 1, Level 2, and Level 3, for performance levels. The neutral labels avoid the excess meanings that are often attached to more descriptive labels." The appendix at the end of this article shows the names of performance levels used in state education assessments developed under NCLB.[2]

Writing Policy Definitions for Each Performance Level

Once the number and names of the levels have been selected, they need to be defined. One recommendation is to develop a generic policy definition for each performance level prior to drafting any PLDs. Policy definitions determine how rigorous and challenging the standards will be for the assessments. They are not linked to content but are more general statements that assert a policymaker's position on the desired level of performance or rigor intended at each level. For instance, proficient can mean "mastery of grade-level subject material," "solid performance on academic content," or "partial success on challenging content." These definitions are consistent across all grades and subjects and help ensure a similar level of rigor is implied by the performance level for each assessment.

Policy definitions are particularly useful in the context of an assessment program with multiple assessments. First, they facilitate the articulation of performance levels across grades by ensuring the same level of rigor at each level across each grade. Second, they allow a reader to interpret proficient in a similar manner regardless of the subject assessed. Consider, for example, a parent reviewing a student end-of-year assessment report. Although multiple subjects will be represented, the level of proficient should have a similar meaning in terms of the depth and breadth of understanding of a content area.

Writing an initial policy definition is the model used by NAEP. For instance, NAEP defines *proficient* as follows:

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

This policy definition applies to all three grade levels across every subject assessed. Further detail is added as a subsequent step to provide specific descriptions of the performance required to show proficiency for each test individually that relates directly to that assessment's content.

A policy definition needs to be written for each performance level, not including the lowest level. For instance, if a state is interested in setting just one cut score, at the proficient level, then a policy definition should be written for the proficient category. It is not necessary to write another definition for below proficient.

NAEP has policy definitions for Advanced, Proficient, and Basic but not for below Basic. The definition for Proficient was provided above. Compare that to the definitions for Basic and Advanced, which are written

using Proficient as the focal performance level:

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

The key to writing strong policy definitions is to use a similar set of words that are memorable and that distinguish clearly among the performance levels. Policymakers should begin drafting the policy definitions by making a statement that is directly linked to their instructional program and goals. The definitions should clearly state the degree of knowledge and skills expected of students at each performance level. They should be concise, approximately 1–2 sentences, and clear. Because it is the backbone of all further writing, policymakers should carefully consider the wording and be sure each definition communicates the intended goals and clearly distinguishes one level from the next.

For example, Pennsylvania distinguishes among four levels of performance using the following definitions:

Advanced: Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in Pennsylvania's academic standards;
Proficient: Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in Pennsylvania's academic standards;
Basic: Marginal academic performance, work approaching, but not yet reaching, satisfactory performance, indicating partial understanding and limited display of the skills included in Pennsylvania's academic standards; and
Below Basic: Inadequate academic performance that indicates little understanding and minimal display of the skills included in Pennsylvania's academic standards.

Notice that the definitions are short, yet clearly distinguish among the different levels and relate directly to the content standards. However, there is room for interpretation in the use of words such as "partial," "satisfactory," and "adequate." Those terms would need further clarification during the development of the full PLDs.

Alabama uses even shorter descriptions that specify performance relative to the content standards but does not provide any indication as to how one determines whether the student has met or exceeded the standards:

Level IV-—Exceeds academic content standards;
Level III-—Meets academic content standards;
Level II—Partially meets academic content standards; and
Level I—Does not meet academic content standards.

In contrast, Arizona uses more specific descriptions of the level of performance expected of students at each level:

Exceeds the Standard: This level denotes demonstration of superior academic performance evidenced by achievement substantially beyond the expected goal of all students.
Meets the Standard: This level denotes demonstration of solid academic performance on challenging subject matter reflected by the content standards. This includes knowledge of subject matter, application of such knowledge to real-world situations, and content-relevant analytical skills. Attainment of at least this level is the expectation for all Arizona students.
Approaches the Standard: This level denotes understanding of the knowledge and application of the skills that are fundamental for proficiency in the standards.
Falls Far Below the Standard: This level denotes sufficient evidence that the prerequisite knowledge and skills needed to approach the standard have not been met. Students who perform at this level have serious gaps in knowledge in skills related to Arizona's academic standards.

States differ not only in the way they define proficient, but in the method they use to develop their definitions. Approaches to writing the policy definitions can vary in terms of timing, people involved, degree of collaboration among stakeholders, and reviews conducted.

Starting with differences in the timing, some policymakers write policy definitions early on as a first step in developing an assessment program. Others draft policy definitions after test development but prior to writing full PLDs, as a separate activity, while some write policy definitions during a PLD-writing workshop as the first step of the descriptor. And some state testing programs choose not to use policy definitions at all. In terms of the people involved, policy definitions can be drafted in one of three ways:

1. Policymakers draft the policy definitions alone.
2. Policymakers work with a small group of content experts and assessment leaders to draft the policy definitions as a stand-alone activity.
3. Content experts draft policy definitions as the first step of a full PLD workshop to be approved later by policymakers.

Policymakers can work alone, or in conjunction with content experts, educators, or other stakeholders. Reviews can be public or private and may be quite extensive with numerous iterations or may constitute as little as a rubber stamp from a state board of education. There is no one correct way to establish the policy definitions. The state/district should decide who needs to approve these definitions and whether or not they want content experts to weigh in on this first step, and then choose the appropriate option. If either option 1 or 2 is chosen, then the policy definitions need to be written *and approved* at least 1 day prior to a PLD-writing workshop. If option 3 is chosen, then allow approximately 1 to 2 hours at the beginning of the workshop to write the policy definitions. Because policy definitions often need to be approved by a state superintendent or board of education, sufficient time should be allocated for that review and subsequent edits.

Again, only one definition is needed per performance level, regardless of the number of grades or subjects assessed. Also, the policy definition typically is used only to define the levels associated with a threshold cut score. Remember, although NAEP has four performance levels, policy definitions have only been created for the top three levels, as Below Basic is defined as performance falling below the minimum performance required to achieve Basic.

Developing Full PLDs

After the policy definitions have been completed and adopted, content descriptions are added to develop full

PLDs. PLDs express the knowledge and skills required to achieve each level of performance for a specific assessment and are linked directly to the content standards for that assessment. They should be developed prior to setting cut scores and used to inform the cut score setting process. In addition, they can be used to provide parents, teachers, and other stakeholders with more information on what students at each level know and are able to do and what they need to know and be able to do to reach the next level. And, as mentioned previously, they can drive item development if developed early in the process.

Ideally, policymakers will convene a small group of content experts in a PLD-writing workshop. To develop PLDs, content experts start with the policy definitions and expand these definitions in terms of specific knowledge, skills, and abilities required at each level for each subject for each grade. PLDs should be built from test content, either in the form of content standards, test specifications or blueprints, or item specifications, depending on when in the process they are being written and what is available at that point. The test items can also be used as supplemental information to help develop the descriptors. Care should be taken, however, to ensure that the descriptors are not written to address a specific item. Rather, they should list the knowledge and skills required to answer correctly that item and others like it. It is important to keep in mind that test items are periodically replaced, in some states after each administration. Therefore, we do not want descriptors that are specific only to the test form that was operational when the descriptors were written.

For example, the following descriptors were writing for NAEP fourth-grade mathematics:

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

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

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

Oftentimes, the language in a descriptor relies on models of cognitive processing, such as those defined in Bloom's taxonomy. That is, a lower level of performance may include words such as "identify" or "describe" while a higher level of performance may include words such as "analyze" or "evaluate." There is much research on the progression learning and instruction that focuses on the type and quality of knowledge (cf. Jong & Ferguson-Hessler, 1996). For example, consider the differences between concrete and abstract knowledge or among declarative, conceptual, and procedural knowledge. Facilitating a discussion on the hierarchy of cognitive learning with those on the PLD-writing committee may help them distinguish among levels while still addressing similar content. Following this approach would most likely result in descriptors that reflect a similar breadth of content but different depths of knowledge and understanding.

The articulation of these descriptions across grade levels must be considered when writing PLDs. The PLDs should be aligned across grades when several consecutive grade levels are included in an assessment system (as required by NCLB and some state accountability systems). For example, the PLD for proficient performance in reading in grade 5 should require more knowledge and skill than is required to be proficient in grade 4, but less knowledge and skill than is required to be proficient in grade 6.

Organizing and facilitating these PLD writing workshops needs to be done with care to ensure a quality product. Mills and Jaeger (1998) produced the first published set of steps for producing test-based descriptions of performance categories. However, their focus was on writing PLDs specific to a test form rather than to a bank of items. Instead, the modified steps proposed below use test specifications and incorporate the requirements for full PLDs under NCLB.

Components of a PLD-Writing Workshop

Prior to convening a committee, policymakers need to determine who will write the policy definitions that determine the level of rigor they want for each performance level. The remainder of this section assumes that the policy definitions were drafted by a committee composed of policymakers and perhaps educators and other stakeholders. It also assumes that the policymakers associated with the assessment, be it a state department of education or governing board, have approved the policy definitions prior to convening the committee.

Forming a committee of educators, content experts, and other stakeholders to write the PLDs is the next task. However, the committee need not be large, as long as the members are familiar with the content standards, student learning, and the purpose of the assessment system. Five to eight people per subject and grade span will suffice. Those with an understanding of the policy context should work alongside those with an understanding of teaching and learning the subject-matter content to write the PLDs. The content experts will bring their knowledge of the content standards and the test

specifications to bear on the descriptions, and the policymakers will provide input on the level of rigor and type of language that will be politically acceptable. Other stakeholders, such as representatives of local businesses, higher education, or the community can also provide valuable feedback on the interpretability of the PLDs. Multiple panels are ideal, but may not be feasible.

The first step in the process is to have the committee review the content standards, the specific content strands that serve as the basis for organizing the test specifications, and the test specifications. This step of reviewing the content standards is crucial. Panelists need to be familiar with the breadth of the content specifications because the breadth needs to be reflected in their descriptions. Provide the committee with multiple sources of data regarding the content that is to be assessed. If not all of the content standards are assessed, this should be made clear to the panelists. Also, if some indicators are rotated across test administrations, that information also needs to be made clear to the panelists to help them understand the appropriate level of specificity for their descriptors. The specifications will provide the necessary detail on the emphasis of the test.

If the PLDs are written after the test has been developed and there are questions about how a specification could be translated to an item, sample test forms may be provided but should be used only as an example of the types of questions that may be asked. Care needs to be taken not to write descriptors so specific that they only apply to one form of the test. In addition, if the test has already been created, the panel should be trained on how the test was structured and scored. The committee members must understand how the content is assessed, but should be cautioned against writing descriptors to a particular form of a test. PLDs should not go beyond the content of a test, but they should not address every item, either.

Once the committee is familiar with the content, the training should focus on the process. Each committee should begin by reviewing the policy definitions to ensure a common understanding that should form the basis for the full descriptors. At this point, policymakers can provide helpful insights as to the political context behind the definitions and explain the level of rigor desired.

The next step is to ask the committee to draft statements that reflect the desired rigor and are specific to the content of the test. Recall that the policy definitions are written at a level that can be applied to any test in a testing program (e.g., reading and math, grades 3–8). But at this point, the description needs to be specific to one test (e.g., grade 3 reading). The committee should be encouraged to focus more on the assessment blueprint and test specifications and not to write a PLD specific to a test form. Often it is helpful to begin this part of the process by encouraging the committee members to brainstorm ideas by content strand and record their thoughts simply as bullet points. Time can be taken later to connect the bullets into complete sentences. Some states prefer to leave their descriptors in bullet format.

One suggestion by Mills and Jaeger (1998) was to start by specifying the knowledge and skills of candidates in a content strand at a middle performance category and build on this knowledge and these skills to arrive at the Advanced level, or remove or lower the knowledge and skills to get to the descriptions at the lower levels. This exercise can be conducted for each content strand or for the test as a whole. In an NCLB context, this would mean first writing the definition for Proficient, and then focusing on Advanced and Basic. It also makes sense to begin with the level that has the most stakes attached to it, such as Proficient under NCLB.

Depending on the number of committees and committee members used, consensus may be developed as part of the process, rather than as a separate step. However, it is important that the content experts and policymakers agree on the components of each of the descriptors. The descriptors may be edited further for sentence structure, but the essence of the levels should be set by the end of the workshop. Similarly, if the panel has been split, so that one group works on PLDs for grades 3–5 and the other works on PLDs for grades 6–8, for example, time should also be allotted for the two panels to work together to ensure that the PLDs articulate well across all grades 3 through 8.

Final Descriptor

Once the PLDs have been written and used to set cut scores, they can be fleshed out even further through the use of exemplar items. That is, when the cut scores are known, psychometricians can identify items that students at one performance level are likely to answer correctly and that students in the lower level are not likely to answer correctly. Different criteria have been used to identify exemplar items, such as the p-value of an item for students scoring proficient must be at least .65 and must be at least .35 higher than the p-value of the same item for students in the lower level. An item's location on a difficulty scale based on item response theory (IRT) also can be used to identify items that fall in the middle of a performance level. Describing common characteristics of these items or including 2–3 of these items with each PLD can add a richness and depth to the final PLD and may provide valuable interpretive information for teachers, students, and parents. These descriptions can be updated periodically throughout the testing program as new items are released.

Current Practice

Although every state now uses PLDs in their assessment program, states differ tremendously in their development process. Under NCLB, all states must use a minimum of three performance levels, but some use as many as five. The naming conventions also differ across states. More importantly, however, states vary in their use of policy definitions, the timing of the development of PLDs, and the level of rigor implied by each descriptor. The following sections detail some of these differences.

Determining the Number and Names for the Performance Levels

Approximately 10 states have decided to use the minimum number of performance levels required by NCLB, with the lowest level simply being below proficient. However, a majority of states (29) have four performance levels, allowing them to differentiate between students who are close to proficient and those who are well below proficient, often called basic and below basic, respectively. The remaining 13 states[3] define five levels of performance, with some adding another level below proficient and others adding a level above proficiency. For instance, California uses five levels with an

additional level below proficient: Far Below Basic, Below Basic, Basic, Proficient, and Advanced, while Delaware uses five levels with an additional level above Proficient: Well Below the Standard, Below the Standard, Meets the Standard, Exceeds the Standard, Distinguished (where Meets the Standard is used to report proficiency). The number of levels above or below a key level (such as Proficient) can often influence the degree of rigor associated with that key level. For instance, if there are two levels above Proficient, both panelists writing the descriptors and those setting the cut scores may feel influenced to leave sufficient "space" above Proficient to define the higher levels.

Even states that have the same numbers of levels often name the levels using very different terms. As discussed earlier, the terms themselves carry meaning, even without further description. Looking solely at the level used to report proficiency under NCLB, we find that 28 states use the term Proficient both in their state reporting system and in meeting federal AYP requirements. The other 24 use a different term internally and map it to the federal requirement for Proficient. Other terms include:

- Meets standards
- Achieve the standard
- Met expectations
- Mastery
- Pass
- Satisfactory
- Intermediate
- Level III

Although eight states use the same naming convention as NAEP—Below Basic, Basic, Proficient, and Advanced—there are many alternatives to a four-level naming system. Several states replace the term Below Basic with terms such as Novice and Limited. Others use a very different naming convention, such as Oklahoma, which labels its performance levels Unsatisfactory, Limited Knowledge, Satisfactory, and Advanced. Others take the approach of labeling the levels relative to the academic standards. For instance, Maine uses the labels: Does Not Meet the Standard, Partially Meets the Standard, Meets the Standard, Exceeds the Standard.

Approximately 12 states follow a similar pattern of reporting proficiency in terms of how close the student is to meeting/achieving the content standards. A full list of performance levels used in each state can be found in the appendix at the end of this article.

Writing Policy Definitions for Each Performance Level

As described earlier, the policy definition sets the stage for how difficult it will be to reach each level. For instance, the NAEP definition of proficient uses the phrase "competency over challenging subject matter" that implies a high level of performance.

The Webster's definition of proficient is "performing in a given art, skill, or branch of learning with correctness and facility (adj.); an expert (n.)." Thus, the initial meaning of this word implied an extremely high level of performance. However, given the different definitions currently in use to describe "proficient" in K–12 education across all the states, the word has lost meaning in a national context.

In 2003, Michael Beck collected definitions for proficient and found 15 different ways that states describe the proficient level:

- Satisfactory achievement.
- Adequate understanding of the on-grade content.
- Solid understanding of challenging subject matter.
- Competency indicating preparation for the next grade level.
- Ability to apply on-grade standards capably.
- Acceptable command of grade-level content and processes.
- Ability to apply concepts and processes effectively.
- Solid academic performance . . . competency with challenging subject matter.
- Solid academic performance . . . prepared for the next grade.
- Mastery of grade-level standards.
- High level of achievement . . . ability to solve complex problems.

Obviously, a state that uses the phrase "satisfactory achievement" to describe proficient performance has a lower expectation for their students than a state that uses the word "mastery." Thus, we would expect the cut scores to be lower in the first state than in the second state, meaning more students would reach proficient in the first state given relatively equal achievement. Using words like "challenging" or "complex" also will result in higher cut scores.

A more recent search of state web sites found that fewer than half of the states have published policy definitions online. These definitions are sometimes called general descriptors or benchmarks in the state reports. In some states, it is these more generic descriptions that appear on score reports. Table 1 provides some examples of how states define proficient for the purposes of reporting AYP.

Other states have descriptions somewhere between policy definitions and full descriptors, creating intermediate descriptions for each subject that are not differentiated across grade levels. For example, Colorado uses the following definitions for reading and math:

> *A student scoring at the Proficient Level routinely utilizes a variety of reading strategies to comprehend and interpret grade-level appropriate text. Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the Colorado Model Content Standards for reading.*

> *A student scoring at the Proficient Level solves practical real-world problems and demonstrates understanding of the skills, concepts and procedures contained in the six Colorado Model Content Standards for mathematics at grade level. Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the Colorado Model Content Standards.*

The appendix at the end of this article notes those states that have developed some version of policy definitions.

Developing Full PLDs

Final PLDs can be in a number of formats. Both the style of the final descriptor and the manner in which the descriptors differentiate among the levels differs across states.

Style

Some testing programs prefer to write their PLDs as summary paragraphs, often one paragraph per level. Others list the key points of their PLDs in bullets. Within those two primary formats, some states write overarching descriptions for each level within each test, while others divide it further and provide a description for each content strand, benchmark, or indicator.

Table 1. Selected State Policy Definitions of Proficient

State	Definition
Arizona	This level denotes demonstration of solid academic performance on challenging subject matter reflected by the content standards. This includes knowledge of subject matter, application of such knowledge to real-world situations, and content-relevant analytical skills.
Florida	Performance at this level indicates that the student has partial success with the challenging content of the Sunshine State Standards but performance is inconsistent. A Level-3 student answers many of the questions correctly but is generally less successful with questions that are most challenging.
Hawaii	Assessment results indicate that the student has demonstrated the knowledge and skills required to meet the content standards for this grade. The student is ready to work on higher levels of this content area.
Idaho	Student demonstrates thorough knowledge and mastery of skills that allows him/her to function independently on all major concepts and skills at his/her educational level.
Kansas	Mastery of core skills is apparent. Knowledge and skills can be applied in most contexts. Ability to apply learned rules to most situations is evident. Adequate command of difficult or challenging content and applications is competently demonstrated. There is evidence of solid performance.
Minnesota	A score at or above Level 3 represents state expectations for achievement of all students. Students who score at Level 3 are working successfully on grade-level material. This level corresponds to a "proficient" level of achievement for NCLB.
Montana	"Proficient level" means solid academic performance for each benchmark, reaching levels of demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
New England Consortium	Students performing at this level demonstrate minor gaps in the prerequisite knowledge and skills needed to participate and perform successfully in instructional activities aligned with the GLE at the current grade level. It is likely that any gaps in prerequisite knowledge and skills demonstrated by these students can be addressed during the course of typical classroom instruction.
Pennsylvania	Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in Pennsylvania's academic standards.
Puerto Rico	The proficient level means that the student recognizes and utilizes the major part of the concepts of the grade level and satisfactorily applies the basic skills of the grade level.
South Carolina	A student who performs at the proficient level on the PACT has met expectations for student performance based on the curriculum standards approved by the State Board of Education. The student is well prepared for work at the next grade. The proficient level represents the long-term goal for student performance in South Carolina.
Texas	This category represents satisfactory academic achievement. Students in this category performed at a level that was at or somewhat above the state passing standard. Students demonstrated a sufficient understanding of the knowledge and skills measured at this grade.
Utah	A student scoring at this level is proficient on the measured standards and objectives of the core curriculum in this subject. The student's performance indicates sufficient understanding and application of key curriculum concepts.
West Virginia	Student demonstrates knowledge, comprehension, and application of skills, which meet the standard.
Wisconsin	Demonstrates competency in the academic knowledge and skills tested.
Wyoming	Students who perform at the proficient level use concepts and skills to solve problems using appropriate strategies and to communicate meaning as required by the standards.

As examples of some of these methods, the following shows a partial example of North Dakota's grade 4 mathematics PLDs, which are written to each specific benchmark within each content strand:

Benchmark 4.1.1 Construct and interpret meanings through real-world experiences.

Level 4 The student accurately constructs, interprets, and extends number meanings through real-world experiences.
Example: Given two makes/models of an automobile, the student can draw conclusions as to why prices may vary between the two, such as options or mileage.

Level 3 The student accurately constructs and interprets number meanings using such strategies as grouping, ordering, one-to-one correspondence, and use of fractions, decimals, and percents to represent real-world experiences.
Example: The student can solve this problem: Which is the better deal and why: 3 pencils for $1.00, 6 pencils for $1.75, or 9 pencils for $2.70?

Level 2 The student constructs and interprets number meanings through real-world experiences with some inaccuracies.

Level 1 The student attempts but is unable to construct and interpret

number meanings through real-world experiences.

Benchmark 4.1.2 Understand the characteristics and properties of our numeration system.

Level 4 The student accurately and consistently applies the characteristics and properties of our numeration system and understands the relationship among numerical concepts.

Example: Given a simple fraction, the student can convert it to equivalent decimals or percents (e.g., 1/4 = 25% = 0.25)

Level 3 The student accurately uses the characteristics and properties of our numeration system, such as place value, grouping, ordering, base 10, fractions, decimals, percents, standard numbers, expanded numbers, ordinal numbers, cardinal numbers, and odd and even numbers.

Example: The student can properly identify the place values in a four-digit number.

Level 2 The student uses the characteristics and properties of our numeration system with some inaccuracies.

Level 1 The student inconsistently uses the characteristics and properties of our numeration system.

Contrast that approach to Georgia's where 1–2 paragraphs are used to describe each level of performance in grade 4 mathematics:

Does Not Meet: The student's overall performance in mathematics is below the standard set for students in fourth grade.

Students performing at this level are inconsistent in the application of their mathematical skills. They demonstrate limited evidence of mathematical conceptual understanding and procedural knowledge. Their computation skills are not fully developed. They may use basic graphing skills. Students performing at this level recognize properties of shapes and names of solid figures. They recognize some numerical relationships. They may have difficulty transferring learning from guided experience to new problems.

Meets: The student's overall performance in mathematics meets the standard set for students in fourth grade.

Students performing at this level generally apply mathematical skills appropriately. They demonstrate evidence of mathematical conceptual understanding and procedural knowledge. Their computation skills are usually accurate. They have some ability to analyze and interpret data from graphs. They recognize geometric relationships of shapes. They can represent pictures or arrays as number sentences. Students performing at this level show evidence of problem-solving ability.

Exceeds: The student's overall performance in mathematics exceeds the standard set for students in fourth grade.

Students performing at this level are consistent in the application of their mathematical skills. They demonstrate strong mathematical conceptual understanding and procedural knowledge. Their computation skills are accurate. They can use data to predict the probability of events. They apply their understanding of geometric relationships by sorting and classifying objects. They are able to determine an element when given a relation or rule. Students performing at this level are able to solve multi-step problems.

In determining which approach to follow a state should consider carefully how the descriptors will be used and by whom in determining the best approach to writing their PLDs. Writing descriptors to each indicator implies a conjunctive rule to determining performance, although most tests use compensatory scoring rules. That is, although the PLD may imply a certain level of performance on each indicator, within the test itself a high score on one indicator can offset a low score on another. If a test heavily weights one indicator or content strand over another, that should be reflected in the PLD. For example, if numbers and operations make up 50% of the mathematics assessment in grade 3 and 15% of the assessment in grade 8, those differing weights ought to be reflected in the descriptors.

There is no real benefit to writing the descriptors in bullet or paragraph form, but the policymakers should consider the reports in which the descriptors will appear. Is the intent to include all levels on a student report card or only the level in which that particular student falls? Will PLDs only appear on a report card or in supporting documentation? Given the answer to those questions, which format will be the most readable to parents, teachers, and other stakeholders?

Differentiating Between Levels

There are several approaches to differentiating performance between adjacent levels. Some states define different content that will be mastered at each level, while other states indicate that the same content will be covered across all levels but to a different degree. As discussed earlier, models of cognitive processing, such as Bloom's taxonomy, may be helpful to those drafting PLDs, as this approach allows panelists to address similar content but to varying degrees of understanding across levels. Another approach is to consider whether students performing at a higher level might actually be learning different content or acquiring different skills from those performing at a lower level. That is, consider whether advancing through the performance levels means the student learns the same material at a different depth or has actually been exposed to more or different material.

As one example, New Jersey provides qualitative differences in the level of reading and writing at proficient and advanced proficient for their grade 4 Language Arts Literacy assessment:

Proficient: The student performing at the proficient level demonstrates abilities to work with, analyze, and critique text. As a proficient reader, the student recognizes the central idea, supporting details, purpose, and organization of text. The student demonstrates the ability to comprehend text literally, to make inferences, and to express understanding of the text in written responses.

As a proficient writer, the student establishes a central focus, generally organizes and connects ideas, and includes some supporting details. The student demonstrates some variety in sentence structure and word choice and uses basic conventions of print.

Advanced Proficient: As a reader, the fourth-grade student performing at the advanced level of proficiency consistently demonstrates the qualities outlined for proficient performance. In addition, the advanced proficient reader makes connections and synthesizes details of the text in order to generate new ideas.

As an advanced proficient writer, the student establishes and develops a central focus, organizes and connects ideas, and elaborates on supporting details coherently. The student varies sentence structures, chooses words effectively, and uses conventions of print.

Puerto Rico, on the other hand, describes a similar set of knowledge and skills between proficient and advanced but differentiates between the levels in the degree of consistency shown in the skills for their grade 4 test in English. This "adverb approach" is simpler to write but provides less interpretative information to the end user, such as the parent or teacher:

Proficient: *Reading Comprehension Skills*

- Usually identify the main idea in the passage.
- Usually identify salient details in a short reading passage.
- Usually identify the sequence of events in the passage.
- Usually identify cause and effect relationships.
- Usually determine character traits.

Proficient: *Writing Skills*

- Usually apply the rules of grammar correctly.
- Usually apply the rules of punctuation correctly.
- Demonstrate some understanding of the more advanced grammatical structures.

Advanced: *Reading Comprehension Skills*

- Consistently identify the main idea in a reading passage, even if it is not explicitly stated.
- Consistently identify details in a reading passage.
- Easily recall the sequence of events in a reading passage.
- Consistently identify cause and effect relationships.
- Consistently determine character traits.

Advanced: *Writing Skills*

- Accurately apply the rules of grammar on a consistent basis.
- Accurately apply the rules of punctuation on a consistent basis.
- Demonstrate understanding most advanced grammatical structures.

Again, there is no one right way to differentiate across levels, but there are implications based on the approach. Content experts would be the most appropriate people to determine whether a higher level of performance is identified more with students knowing different skills or knowing the same skills to a different degree. Making a policy statement on this issue would also send a message as to the importance of teaching all skills to a certain level compared to teaching certain skills to a more advanced level before moving on to new skills. It also would have implications for test design.

Final Descriptor

Finally, detail is sometimes added to the PLDs used for standard setting to assist with reporting. That is, once the cut scores have been set, policymakers can choose to include exemplar items at each level to help with the interpretation of the level. For instance, the New England Common Assessment Program developed a policy definition first, and then developed subject-specific descriptors to be used in standard setting. After the cut scores were set, full descriptions by grade level were developed incorporating information about items falling within each performance level, focusing particularly on items that discriminated well across performance levels.

However, some states that do not use a formal PLD-development process wait until this final stage to write a PLD at all. Oftentimes, with the Bookmark standard-setting process, the PLDs are not written until after cut scores have been set. The items falling between each bookmark are summarized and used to describe each level. In fact, this is not a true descriptor-writing process reflective of the content required, but an item-mapping approach summarizing examinee performance on test items. Thus, this process is very dependent on the test form used in standard setting and on the relative difficulty of the items in that particular testing year. Changes in items from year to year, changes in scaling techniques, or changes in the relative difficulty of the items due to curriculum or motivation effects could strongly influence the descriptors. Additionally, as mentioned earlier, writing descriptions after setting the cut score is a little like the cart driving the ox as the definitions should drive the placement of the cut score and not the other way around.

Discussion

When developed and used correctly, PLDs convey a wealth of information about a state's goals for its students. They represent policymakers' intentions about the amount of knowledge and skills required of students in each subject and grade. They provide an indication of how knowledge is assumed to be attained across levels and grades and provide a blueprint for demonstrating growth. Well-written PLDs disseminated in a timely manner can impact not only decisions about test development and cut scores, but also can inform teachers, parents, and students of the knowledge and skills intended to be learned in a year.

Rather than a last-minute task done prior to standard setting, or worse, as a last step of standard setting, developing PLDs should be an activity undertaken early in the assessment development process, so that all interested parties are clear on the types of information expected from the assessments.

If, as argued here, PLDs drive the placement of the cut scores, then it's no wonder that the percentage of students performing at or above proficient varies so much across the states. Currently, proficient has so many definitions under states' adaptations of NCLB that it has virtually lost its meaning in any national discussion of results on state assessments. Although there are many complexities in developing PLDs that can lead to wide variances among states, many of these differences result from the nature of our national accountability system. NCLB allows each state to define both their content and achievement standards. Therefore grade 7 mathematics, for example, could include different content in different states as well as require different levels of performance for proficiency. In addition to focusing on methods for placing the cut scores for all state assessments on the same scale, researchers should also be examining the relationship between these relative cut scores and each state's definition of proficient.

If it is important to be able to compare the proficient level across states on the state assessments, federal policymakers might consider developing a policy definition for proficient that all states can adopt. If all states started from the same generic policy definition for proficient and then wrote their PLDs by supplementing this policy

definition with details from their own state content standards, the level of rigor would be much more similar across states. Of course, we would still caution against making strong comparative statements across states as the content standards, test specifications, item difficulties, and standard-setting methodology would still vary considerably across states. However, starting with the same intended level of rigor for the term proficient would at least remove one area of variability from the equation.

This position does not advocate adopting the NAEP policy definition for proficient, as that definition implies a level of excellence not necessarily achievable by 100% of the students by 2014. However, a definition that includes phrases most commonly found in state policy definitions, such as "adequate demonstration of knowledge and skills described in the state content standards" and applied to all states may help standardize the interpretation of the rigor implied by the word "proficient." At the very least, states should review their current PLDs thoughtfully, compare them to the current assessment content, and determine whether they are facilitating appropriate interpretations of the scale scores.

Acknowledgments

The author would like to thank numerous colleagues for reviewing and commenting on early drafts of this article, including Michael Zieky, Michael Beck, Hillary Michaels, and Scott Marion. Further, the three anonymous reviewers of this article provided useful and detailed comments that greatly improved the final product.

Notes

[1]Generally, every state develops statements of the knowledge and skills that students are expected to learn for each subject and at each grade level called content standards or academic standards.

[2]All of the information regarding state performance levels was downloaded from state web sites in 2006.

[3]For the purpose of this article, we consider 52 "states" as both the District of Columbia and Puerto Rico are included in the analysis.

Appendix: A Summary of State Performance Levels

State	Number of Performance Levels	Names of Levels	Cutoff Level Used to Report Proficiency Under NCLB	Uses General Policy Definitions Across All Grades/Subjects
Alabama	4	Level I—Does Not Meet Academic Content Standards Level II—Partially Meets Academic Content Standards Level III-–Meets Academic Content Standards Level IV—Exceeds Academic Content Standards	Level III	
Alaska	4	Far Below Proficient Below Proficient Proficient Advanced	Proficient	
Arizona	4	Falls Far Below the Standard Approaches the Standard Meets the Standard Exceeds the Standard	Meets the Standard	√
Arkansas	4	Below Basic Basic Proficient Advanced	Proficient	
California	5	Far Below Basic Below Basic Basic Proficient Advanced	Proficient	
Colorado	3	Basic Proficient Advanced	Proficient	
Connecticut	5	Below Basic Basic Proficient Goal Advanced	Proficient	

Continued

Appendix: Continued

State	Number of Performance Levels	Names of Levels	Cutoff Level Used to Report Proficiency Under NCLB	Uses General Policy Definitions Across All Grades/Subjects
Delaware	5	Well Below the Standard Meets the Standard Exceeds the Standard Distinguished	Meets the Standard	√
District of Columbia	4	Below Basic Basic Proficient Advanced	Proficient	
Florida	5	Level 1 Level 2 Level 3 Level 4 Level 5	Level 3	√
Georgia	3	Does Not Meet the Standard Meets the Standard Exceeds the Standard	Meets the Standard	
Hawaii	4	Level 1. Well Below Proficiency Level 2. Approaches Proficiency Level 3. Meets Proficiency Level 4. Exceeds Proficiency	Level 3	√
Idaho	4	Level 2. Approaches Proficiency Below Basic Basic Proficient Advanced	Proficient	√
Illinois	4	Academic Warning Below Standards Meets Standards Exceeds Standards	Meets Standard	
Indiana	3	Did Not Pass Pass Pass +	Pass	
Iowa	3	Low Intermediate High	Intermediate	
Kansas	5	Unsatisfactory Basic Proficient Advanced Exemplary	Proficient	√
Kentucky	4	Novice Apprentice Proficient Distinguished	Proficient	
Louisiana	5	Unsatisfactory Approaching Basic Basic Mastery Advanced	Basic	
Maine	4	Does Not Meet the Standard Partially Meets the Standard Meets the Standard Exceeds the Standard	Meets the Standard	
Maryland	3	Basic Proficient Advanced	Proficient	

Continued

Appendix: Continued

State	Number of Performance Levels	Names of Levels	Cutoff Level Used to Report Proficiency Under NCLB	Uses General Policy Definitions Across All Grades/Subjects
Massachusetts	4	Failing (HS)/Warning (Elementary and Middle Grades) Needs Improvement Proficient Advanced	Proficient	
Michigan	4	Level 1—Exceeded Expectations Level 2—Met Expectations Level 3—Basic Level 4—Apprentice	Level 2—Met Expectations	
Minnesota	5	Level 1 Level 2 Level 3 Level 4 Level 5	Level 3	√
Mississippi	4	Minimal Basic Proficient Advanced	Proficient	
Missouri	5	Step One Progressing Nearing Proficient Proficient Advanced	Proficient	
Montana	4	Novice Nearing Proficient Proficient Advanced	Proficient	√
Nebraska	4	Basic Progressing Proficient Advanced	Proficient	
Nevada	4	Developing/Emergent Approaches Standard Meets Standard Exceeds Standard	Meets Standard	
New Hampshire	4	Novice Basic Proficient Advanced	Proficient	
New Jersey	3	Partially Proficient Proficient Advanced Proficient	Proficient	
New Mexico	4	Beginning Step Nearing Proficiency Proficient Advanced	Proficient	
New York	4	Basic Basic Proficiency Proficiency Advanced	Proficiency	√
North Carolina	4	Level I Level II Level III Level IV	Level III	
North Dakota	4	Novice Partially Proficient Proficient Advanced	Proficient	

Continued

Appendix: Continued

State	Number of Performance Levels	Names of Levels	Cutoff Level Used to Report Proficiency Under NCLB	Uses General Policy Definitions Across All Grades/Subjects
Ohio	5	Below Basic Basic Proficient Accelerated Advanced	Proficient	√
Oklahoma	4	Unsatisfactory Limited Knowledge Satisfactory Advanced	Satisfactory	
Oregon	5	Very Low Low Nearly Meets Standard Meets Standard Exceeds Standard	Meets Standard	
Pennsylvania	4	Below Basic Basic Proficient Advanced	Proficient	√
Puerto Rico	3	Basic Proficient Advanced	Proficient	√
Rhode Island	6	No Score (0) Little Evidence of Achievement (1) Below the Standard (2) Nearly Achieved Standard (3) Achieved the Standard (4) Achieved Standard with Honors (5)	Achieved the Standard	
South Carolina	4	Below Basic Basic Proficient Advanced	Proficient	√
South Dakota	4	Below Basic Basic Proficient Advanced	Proficient	
Tennessee	3	Below Proficient Proficient Advanced	Proficient	
Texas	3	Did Not Meet the Standard Met the Standard Commended Performance	Met the Standard	√
Utah	4	Level 1: Minimal Level 2: Partial Level 3: Sufficient Level 4: Substantial	Level 3	√
Vermont	5	Little Evidence of Achievement Below the Standard Nearly Achieves the Standard Achieves the Standard Achieves the Standard with Honors	Achieves the Standard	
Virginia	3	Fails/Does Not Meet the Standards Pass/Proficient Pass/Advanced	Pass/Proficient	
Washington	4	Level 1. Below Basic Level 2. Basic Level 3. Proficient Level 4. Advanced	Level 3. Proficient	

Continued

Appendix: Continued

State	Number of Performance Levels	Names of Levels	Cutoff Level Used to Report Proficiency Under NCLB	Uses General Policy Definitions Across All Grades/Subjects
West Virginia	5	Novice Partial Mastery Mastery Above Mastery Distinguished	Mastery	√
Wisconsin	4	Minimal Basic Proficient Advanced	Proficient	√
Wyoming	4	Below Basic Basic Proficient Advanced	Proficient	√

References

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.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. (2006). A prospective approach to standard setting. Paper presented in *Assessing and modeling development in school: Intellectual growth and standard setting,* October 19–20, University of Maryland , College Park.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, *9*(3), 215–235.

Braun, H. I., & Qian, J. (2007). *Mapping state performance standards onto the NAEP scale*. Princeton, NJ: Educational Testing Service.

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

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., Giraud, G., & Plake, B. (2000, April). *The influence of providing target group descriptors when setting a passing score*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*, 105–113.

Linn, R. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Educational Policy Analysis Archives*, *13*(33). Retrieved January 15, 2007 from http://epaa.asu.edu/epaa/v13n33.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

McLaughlin, D. (2006). *The state assessment database and NAEP*. Presentation at the CCSSO Large Scale Assessment Conference, San Francisco, CA.

McLaughlin, D., & Bandeira de Mello, V. (2005). *How to compare NAEP and state assessment results*. Training session presented at the CCSSO Large Scale Assessment Conference, San Antonio, TX.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I* (pp. 73–85). Washington, DC: Council of Chief State School Officers.

U.S. Department of Education, Office of Elementary and Secondary Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

Zieky, M., Perie, M., & Livingston, S. (2008). *Cut scores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.