THE SYSTEMATIC IDENTIFICATION OF PERFORMANCE STANDARDS

Regional Educational Laboratory
Contract #ED-01-CO-0006
Deliverable #2005-04

Prepared by
John S. Kendall
Susan E. Ryan
Amy T. Richardson

The following individuals contributed to work described in this report:

Jill Williams, Consulting Associate
Amitra Blake-Schwols, Content Analyst

This document has been funded at least in part with federal funds from the U.S. Department of Education under contract number ED-01-CO-0006. The content of this publication does not necessarily reflect the views or policies of the Department of Education nor does mention of trade names, commercial products, or organizations imply endorsements by the U.S. Government.

TABLE OF CONTENTS

PREFACE

This paper describes a practical method for identifying and communicating expectations of student performance in the classroom. The method has been used at McREL for clients who require a systematic approach to setting performance expectations and communicating these to their constituents.

Performance standards are needed for the same reasons that content standards are needed. Until expected levels of student performance are described, the application of content standards will vary from one school and classroom to another. Districts and states can use performance standards to strengthen the connection among standards, instruction, curriculum, and classroom assessment in order to improve student achievement on standards-based assessments.

The method described in this report is intended to help district curriculum directors and teachers establish performance standards that are meaningful for the classroom, and make possible a better alignment between the content of the curriculum, instruction, and assessment.

THE NEED FOR PERFORMANCE STANDARDS

Although many states and education agencies use similar terms [for different types of standards] and use them consistently, others adopt different definitions for the same terms or use different terms to communicate very similar ideas. What is needed in standards work is to make clear the distinctions between content and performance standards and their related concepts.

The particular emphasis in standards-based education differs somewhat among states and organizations (Lauer, Snow, Martin-Glenn, Van Buhler, Stoutemyer & Snow-Renner, 2005). Nonetheless, there are a few characteristics of these systems that appear to be consistent among state departments of education and the National Research Council. What *is* definitive is that standards-based education identifies what students should learn. Standards established for each content area provide the basis for the alignment of curriculum and assessment and so remain a focal point of reform. Even today, more than fifteen years after the first standards document was published, states and districts continually work to improve their standards, seeking to ensure that they reflect the best thinking about curriculum content, but also communicate challenging and appropriate expectations for student performance. This report describes a process that can be used to systematically identify and communicate performance standards. The primary goal is to offer a method that provides well-grounded performance standards in the classroom that can be tied to assessments at the state level, but that also can be used to make day-to-day decisions for grading, and align performance tasks and activities.

Confusion about standards-based education can be a result of a lack of understanding about the various types of standards. Although many states and education agencies use similar terms and use them consistently, others adopt different definitions for the same terms or use different terms to communicate very similar ideas. Thus, what is needed in standards work is to make clear the distinctions between content and performance standards and their related concepts. This paper makes those distinctions by defining a core set of terms: content standard, benchmark, strand or topic, knowledge/skill statement, performance standard, performance expectation, and performance indicator. The definitions presented here have evolved during the development of the McREL standards database, an effort initiated in the early years of the standards movement.[1]

CONTENT STANDARDS

A content standard is a description of what students should know and/or be able to do within a particular discipline. Content standards primarily serve to organize an academic subject domain through a manageable number (from 5-12) of generally stated goals for student learning. These statements help to clarify the broad goals within the discipline and provide a means for readers to navigate the standards document when searching for specific content. For example, a standard might state that the student "understands and applies basic and advanced properties of the concepts of geometry." In addition to "content standard," labels such

[1] The current database is in its 4th online edition (Kendall & Marzano, 2004), and represents the synthesis of 137 documents, including standards developed by national subject-area organizations, standards from states highly regarded for their work in specific subject areas, and other sources, such as the assessment frameworks from the National Assessment for Educational Progress (NAEP). The database includes 4,100 benchmarks that comprise over 23,000 statements of knowledge and skill, organized by 256 standards and nearly 1000 topics. This content appears across 14 major areas, including health, physical education, language arts, mathematics, science, and others.

as “goals,” “expectations,” and “learning results” serve to identify the same or a similar level of subject-matter organization in different states’ standards documents.

Strand/Topic

Oftentimes a content standard is too broad to effectively guide lesson planning, reporting, or record keeping. A strand or topic is the level of organization between the standard level and benchmarks (described below). Under the geometry standard, for example, topics or strands might include Shapes and Figures, Lines and Angles, or Transformations/Motion Geometry.

Among the purposes served by the topic are to (1) provide teachers with an easier way of finding related benchmarks for instruction (2) make clearer the connections that exist among benchmarks within and across the disciplines for more integrated curriculum planning, and (3) provide a convenient level of specificity for providing feedback to students and parents. (For a further discussion of the uses and development of topics, see *Topics: A Roadmap to Standards* [Kendall, 2000]).

Benchmark

A benchmark is a description of a specific knowledge or skill that students should acquire by a particular point in their schooling. The information provided by a benchmark is more specific than that provided by a content standard. For example, a benchmark found within a geometry standard might be, “Students understand basic properties of figures (e.g., two- or three-dimensionality, symmetry, number of faces, type of angle).”

The specificity of content description identified here by the term “benchmark” appears in other documents under various other names, including “indicator,” “learning expectation,” and even “performance standard.” A benchmark should be specific enough that readers are clear about the instruction and learning it should entail, but not so narrow as to prescribe the day-to-day curriculum. For example, the idea that every student should distinguish between a triangle and a square is both too prescriptive for a benchmark, in that it stipulates exactly how each student will demonstrate their understanding, and too narrow, in that it would likely not take very long for most students to master.

Knowledge/Skill Statement

A knowledge or skill statement describes in specific terms one of the number of ideas that comprise a benchmark. “Unpacking” is a term commonly used to describe the analytic process of identifying the knowledge and/or skills that make up a benchmark. Like the benchmark, the knowledge or skill statement is written as either declarative or procedural knowledge. As an example, consider the following benchmark, which appears in relation science standard 10 at the 6–8 level in the McREL database: “Knows that an object’s motion can be described and represented graphically according to its position, direction of motion, and speed.”

There are a number of items of information within this benchmark, which can be clearly represented as separate knowledge statements.

1. Knows that an object’s motion can be partly described by its change in position
2. Knows that an object’s motion can be partly described by its direction of motion
3. Knows that an object’s motion can be partly described by its speed
4. Knows that an object’s change in position can be represented graphically
5. Knows that an object’s speed can be represented graphically

The knowledge or skill statement provides a level of detail that can be useful both for lesson planning and assessment. For lesson planning, the statements make clear what knowledge and skills contribute to mastery of the benchmark. For assessment, the knowledge/skill statement provides a level of specificity that may be more appropriate for a single test item and the specificity often required for establishing a clear, shared standard of performance.

PERFORMANCE STANDARD

Performance standards are descriptions of performance that specify "how good is good enough" (National Education Standards and Improvement Council, 1993, p. iii).

Performance standards are established by selecting a level of expectation (in other words, by determining the appropriate level of difficulty for the given content) and can be communicated through a number of ways, including through performance indicators. Ideally, a performance standard is communicated in such a way that it is not only rigorous enough to form the basis for an external assessment, but also detailed enough to make clear the implications for curriculum, instruction, and assessment in the classroom.

Steps to Establishing Performance Standards

1. Develop or adopt a systematic approach to describing levels of difficulty relative to the mastery of knowledge and skill.
2. Become informed about and reach consensus regarding the appropriate performance expectation for specific descriptions of student knowledge and skill.
3. Collect and use a format that clearly communicates this performance to those we will benefit from its use.

Performance Expectation

Performance expectations form the foundation for performance standards. Most educators are familiar with the levels of difficulty described in what is called Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl [eds.], 1956). The performance expectation within that taxonomy is identified by such terms as Comprehension, such as interpreting facts, Application, such as using concepts, and Analysis, such as separating components.

Performance expectations, therefore, describe the expected level of cognitive demand through taxonomies—for example, whether the student should simply identify or recognize features of information, or whether the student can make decisions about the use of that knowledge. A performance expectation should be specific enough so that it is clear how it can be applied to a broad range of curricula and assessments, but not so specific that it narrowly prescribes how teachers should teach, or preclude students from learning in a variety of ways.

Performance Indicators

A performance indicator provides a measure of the level of difficulty of knowledge or skills that students might be expected to demonstrate. For instance, a benchmark might state:

> The student understands the relative size of whole numbers, commonly used fractions, decimals, and percents.

Using Bloom's taxonomy, a related performance *expectation0* might be *Application*. A related performance *indicator* might state: The student orders commonly used fractions, including halves, thirds, fourths, fifths, sixths, and eighths, using concrete materials.

A SYSTEMATIC APPROACH TO ESTABLISHING PERFORMANCE STANDARDS

Any synthesis of content standards and benchmarks, whether compiled by states or by professional organizations, represents a consensus among educators and experts about content in a subject area. The McREL database, described earlier, could be said to be a compendium of agreements about the content of the curriculum across a variety of subjects. The analytic method that produced the current database depended upon the fact that unambiguous inferences about what students should know and be able to do could be drawn from these documents. Sometimes the commonalities among documents were clear and explicit; at other times, the agreement was implied, but could still be reasonably inferred. However, in various documents performance levels, unlike content statements, cannot be synthesized into one level without losing meaning. For example, consider the following two statements; the first is from the *Provisional Item Specifications: 1994 National Assessment of Educational Progress in U.S. History* (Council of Chief State School Officers, 1992), the second from the *National Standards from History: Basic Edition* from the National Center for History in the Schools.

> [Students should]… define the New Deal and explain some ways it attempted to reverse the effects of the Depression (CCSSO, 1992, p. 72)
>
> [Students should] analyze the links between the early New Deal and Progressivism. (NCHS, 1996, p. 118)

From a content analysis of these passages, it seems clear that students are expected to possess a basic understanding of the New Deal. The NAEP (CCSSO, 1992) specification requires that students explain in detail how the New Deal attempted to address the effects of the Depression. The NCHS expectation requires a comparison between the New Deal and Progressivism, a more challenging level of performance. However, NAEP specifications reference Progressivism elsewhere, so clearly both documents require some familiarity with both the New Deal and Progressivism. Although the academic content can be reconciled, the differing expectations for performance relative to that content can not. In the NAEP example above, students are expected to remember information about the Depression, a less demanding task than is expected by the NCHS, which requires an analysis and comparison. (Differences in performance levels varied across the documents; the NAEP document was as often more demanding than the NCHS document.)

Thus, differing performance levels cannot be accurately represented in a single description. This is the principal reason that the McREL online database provides only a description of declarative knowledge (introduced by the generic verbs 'know' and 'understand') or skill (introduced by the skill itself, such as 'adds' or 'writes'), rather than any expressions that suggest a level of expected performance, such as 'recognize', 'compare', or 'evaluate'. Nonetheless, the defining element of standards-based education is clarity and agreement not only about what students should know and be able to do, but also about *how well* they should master this knowledge and skill. There are several key steps to successfully developing performance standards, each of which is addressed in the following sections:

- Adopt or adapt a systematic method (a taxonomy) for assigning performance expectations
- Determine and assign performance expectations
- Communicate performance expectations through indicators, rubrics, and other means

The first step concerns the need for a systematic method for determining and assigning performance expectations in a way that does not impede good teaching and learning. The proposed method, described below, uses what is commonly called a taxonomy of educational objectives to organize performance expectations, usually in terms of a hierarchy of difficulty. The next essential step for the establishment of performance standards is determining what level of expectation is appropriate for given knowledge and skill at a grade level. This step requires a consensus that reflects the best thinking among educators and others about the developmental capacity of students and the level of achievement required by society. Last, and as important, is the communication of these expectations within the education system. The two most common approaches to communication are the use of performance indicators and rubrics.

ADOPT OR ADAPT A SYSTEMATIC METHOD (A TAXONOMY) FOR ASSIGNING PERFORMANCE EXPECTATIONS

A number of the standards documents adopt a system for identifying distinct levels of difficulty, commonly known as a taxonomy of objectives. However, we found no evidence among the 137 documents analyzed for the Compendium, except for assessment frameworks from the National Assessment of Education Progress (NAEP), that avowed any strict application of an approach for describing levels of difficulty.

Bloom's Taxonomy

The most common source for the terms used in standards documents is Bloom's taxonomy, alluded to earlier. This taxonomy was established in the attempt to articulate a hierarchical structure by which learning objectives could be ordered by degree of difficulty. Commonly, at least in many standards documents, the taxonomy is used by assigning verbs that indicate the various levels of the system — such as retrieval, analysis, and evaluation — to indicate the level of difficulty. Although still in use, Bloom's taxonomy cannot reflect the research in cognitive science developed in the forty years since its publication. Most problematic, the hierarchical structure has been found to be flawed logically, and not supported empirically (for a discussion, see Marzano, 2001, p. 1–9). Bloom's taxonomy was revised in 2001. One of the editors of the original taxonomy, D.R. Krathwohl, is among its editors. The revised taxonomy (Anderson, et al., 2001), re-forms the hierarchy into a two-dimensional framework, or table. A number of additional refinements are made, and the strict hierarchy of the original model is loosened somewhat.

Several recent efforts have helped codify levels of difficulty for learning expectations. For example, in a model designed to judge the alignment between expectations and assessments (discussed in Webb, 1997), the criterion "depth of knowledge" is used to communicate the relative level of difficulty. Andrew Porter (2002) has created five descriptors by which student expectations can be characterized in terms of their "cognitive demand": memorize, perform procedures, communicate understanding, solve non-routine problems, and conjecture/generalize/prove.

An Alternative Approach

Of most interest for the purposes of developing performance standards for classroom use, however, is a taxonomy of educational objectives designed by Marzano (2001). This taxonomy is consistent with recent research in cognitive science about the relative difficulty of mental tasks. Marzano notes that, with the complexity of a mental process or skill — such as performing long division — the more familiar one is with a process, the more quickly one executes it and the easier it becomes. Thus, mental processes and skills should not be ordered hierarchically in terms of their complexity. They can, however, be ordered in terms of levels of control — that is, some mental processes exercise control over other processes. For example, deciding to use the division process to solve a problem means the student is engaged in the higher-order process of problem solving. Processes can also be ordered in terms of the conscious awareness that is required to execute them. For example, the lowest level, the retrieval process, can be monitored for accuracy by the higher mental process of metacognition. In this organization, the process of comprehension requires slightly more conscious thought than the process of retrieval, and the process of analysis, and of utilization, even more conscious thought. Thus, the hierarchically ordered levels of difficulty, which do not depend upon the complexity of a task for their ordering, provide a useful means for analyzing and describing levels of student performance. There is some evidence that the taxonomy can be used to distinguish levels of difficulty. For example, for a set of released grade 4 math items (NAEP, 2003), approximately 19% of the variation in item difficulty was related to the taxonomic level of the items (see the Appendix).

Particularly useful in the taxonomy is the distinction maintained between declarative and procedural knowledge. Levels of difficulty are described not only in their relationship to each other, but also with respect to how they relate differently to information and skill. A summary of the taxonomy is presented in Table 1.

The New Taxonomy consists of six levels:

1. Retrieval
2. Comprehension
3. Analysis
4. Utilization
5. Goal setting and monitoring
6. Self-system thinking

For the purposes of this discussion of performance standards, the first four levels are examined. Although the highest levels, goal setting and monitoring and self system thinking, have a significant impact on student performance, they are not yet common enough in standards documents to warrant treatment in this overview of standard-setting.

Table 1. Summary of Taxonomy Levels

LEVEL 1: RETRIEVAL	
Recall	The student can identify or recognize features of information, but does not necessarily understand the structure of knowledge or is able to differentiate critical from non-critical components.
Execution	The student can perform a procedure without significant error but does not necessarily understand how and why the procedure works.
LEVEL 2: COMPREHENSION	
Synthesis	The student can identify the basic structure of knowledge and the critical as opposed to non-critical characteristics of that structure.
Representation	The student can construct an accurate symbolic representation of knowledge, differentiating critical from non-critical elements.

LEVEL 3: ANALYSIS	
Matching	The student can identify important similarities and differences in knowledge or skill.
Classifying	The student can identify superordinate and subordinate categories related to knowledge or skill.
Error Analysis	The student can identify errors in the presentation or use of knowledge.
Generalizing	The student can construct new generalizations or principles based on knowledge.
Specifying	The student can identify specific applications or logical consequences of knowledge.
LEVEL 4: UTILIZATION	
Decision Making	The student can use the knowledge to make decisions or the student is able to make decisions about the use of the knowledge.
Problem Solving	The student can use the knowledge to solve problems or to solve problems about the knowledge.
Experimental Inquiry	The student can use the knowledge to generate and test hypotheses or to generate and test hypotheses about the knowledge.
Investigation	The student can use the knowledge to conduct investigations or to conduct investigations about the knowledge.
LEVEL 5: GOAL SETTING AND MONITORING	
Goal Setting	The student can set a plan for goals relative to the knowledge.
Process Monitoring	The student can monitor the execution of the knowledge.
Monitoring Clarity	The student can determine the extent to which he or she has clarity about the knowledge.
Monitoring Accuracy	The student can determine the extent to which he or she is accurate about the knowledge.
LEVEL 6: SELF	
Examining Importance	The student can identify how important the knowledge is to him or her and the reasoning underlying this perception.
Examining Efficacy	The student can identify beliefs about his or her ability to improve competence or understanding relative to knowledge and the reasoning underlying this perception.
Examining Emotional Response	The student can identify emotional responses to knowledge and the reasons for these responses.
Examining Motivation	The student can identify his or her level of motivation to improve competence or understanding relative to knowledge and the reasons for this level of motivation.

Adapted from: Marzano, Robert (2001). *Designing A New Taxonomy of Educational Objectives*

Most, if not all, standards in the language arts anticipate that at some point in their schooling students should understand genres in literature. How well they should understand genres is a way of asking about the level of difficulty students are expected to master relative to the content. If, for the sake of an example, we determined that students should understand genres well enough that they are able to contrast characteristics among several of them, we find that according to the description provided by Marzano's taxonomy, the expectations are for Analysis: Matching (see Table 2). It seems likely that at an earlier grade of schooling, students would simply be expected to know the characteristics of one or more genres. Providing information about a given genre would be characterized at a lower level of the taxonomy, namely, Level 1, Retrieval: Recall, as appears in Table 2. The general description for declarative knowledge at each level makes clear how the two levels are distinct in terms of their difficulty, yet related in the type of knowledge they address.

Table 2. Sample Distinction between Taxonomic Levels for Declarative Knowledge

	GENERAL DESCRIPTION	APPLICATION TO CONTENT
Analysis: Matching	Identifies similarities and differences in details, principles, or generalizations	Identifies the similarities and differences among legends, fables, and folktales as genres
Retrieval: Recall	Can produce information related to details, principles, and generalizations	Describes the major characteristics of the fable as a genre

An analogous structure is available for considering expectations in regard to procedural knowledge. For example, the ability to read and understand a variety of texts is a basic skill. In the taxonomy, it appears at Level 1, Retrieval: Execution. At a later stage of schooling, students should be cognizant of the various strategies they use in reading — for example, how the strategies for locating information differ from those of reading comprehension. Considering the description for Procedural Knowledge as shown in Table 3, understanding such differences between the two processes would appear at Level 3, Analysis: Matching.

Table 3. Sample Distinction between Taxonomic Levels for Procedural Knowledge

LEVEL	GENERAL DESCRIPTION	APPLICATION TO CONTENT
Analysis: Matching	Can identify how mental skills or procedures are similar or different	Identifies how strategies of reading for locating information differ from strategies of reading for comprehension
Retrieval: Execution	Can perform the skill without significant error	Uses reading skills and strategies to understand a variety of texts

In the examples just discussed, the benchmarks themselves were at different levels of the taxonomy. The taxonomy describes all levels of expected performance, so it is possible to identify various levels of performance even for the same benchmark. An example is provided in Table 4, in which a benchmark on congruent or similar shapes is assigned various taxonomic levels along with performance indicators that reflect each level.

The advantage of this approach then, is that it codifies and makes systematic a way of talking about expectations. Any number of performance expectations could be developed that conform to the level of difficulty described for a given taxonomic level. Such an approach provides a mechanism for greater clarity at all levels of the school system. At the classroom level, the addition of information about performance expectations clarifies the targets for instruction, makes clearer whether the curriculum serves the objective, and can guide the development of classroom assessment. At the district level and state level, there is improved communication about expectations for a given benchmark, and some assurance of common agreement about expectations as to what might be assessed.

Table 4. Different Taxonomic Levels for the Same Benchmark

DECLARATIVE BENCHMARK (UPPER ELEMENTARY)	TAXONOMIC LEVEL	PERFORMANCE INDICATOR
Understands that shapes can be congruent or similar	**Level 1: Retrieval** Recall: The student can identify or recognize features of information.	The student identifies or draws congruent or similar shapes.
	Level 2: Comprehension Representation: The student can accurately represent the component parts of knowledge in non-linguistic or symbolic form.	The student models geometric figures that are congruent or similar, or some combination of these properties. The student describes motions needed to show congruence. For example, the student states that to make shape A and shape B congruent, shape A must be turned 90 degrees, flipped vertically, and rotated 180 degrees.
	Level 4: Knowledge Utilization Problem-solving: The student can use the knowledge to solve problems or can solve problems about the knowledge.	The student solves problems by showing relationships between and among figures, such as similarity or congruency.

DETERMINING AND ASSIGNING TAXONOMIC LEVELS

Once an approach has been adopted for defining levels of expected performance, it is also necessary to choose the level of content specificity (e.g., standard, benchmark, topic) — for which performance standards will be established. Once that is determined, the task is to determine consensus about the appropriate taxonomic level for each content element.

Levels of Specificity

It is possible to establish performance standards for students at a number of levels of specificity. A performance expectation can be assigned at the level of a standard, or for a set of benchmarks — as for example, organized by a topic — at the benchmark level, or for each of the knowledge or skill statements that a benchmark comprises. Table 5 provides examples of what the performance *indicators* might look like if performance *expectations* were to be established at each of these levels.

Table 5. Sample Performance Indicators by Levels of Content Specificity

CONTENT DESCRIPTION	PERFORMANCE INDICATOR
Standard	
Understands and interprets a variety of literary texts	Reads and evaluates a variety of literary texts for their use of literary style and technique, management of story elements, and use of genre characteristics
Topic	
Literary genres	Understands a variety of literary passages in terms of the genre and its defining characteristics, and how literary works differ in genre and culture
Benchmark	
Knows the defining characteristics of a variety of literary forms and genres, such as myths, fairy tales, poems and folk tales	Distinguishes the similarities and differences between myths and other genres, such as fairy tales, legends, fables, and folk tales
Knowledge/Skill	
Knows the defining characteristics of myths	Identifies the essential characteristics that distinguish myth

Clearly, the level of specificity at which a performance expectation is established has a significant impact on the scope and detail of the performance indicator. As to be expected, the performance description at the standard level is broad and covers a considerable amount of material. At the other end of the spectrum, the performance indicator for the knowledge/skill statement is quite specific and detailed. The use of one content level to establish expectations does not preclude the use of any other, however, and performance expectations at different levels can suit differing purposes. For example, the detail at the knowledge/skill level should make clear what instruction and curriculum would best help students to meet the performance indicator described. As a practical matter, the alignment of curriculum, instruction, and assessment is most straightforward at this level. However, for monitoring student strengths and areas for improvement and for general record-keeping, the topic or benchmark level may be more useful. There are fewer items to track, yet there is enough in common among them, that the focal point of instruction is obvious. The topic or, more commonly, the standard level is most useful for reporting out to parents and stakeholders.

Determining the Consensus on Performance Expectation

The specificity at the knowledge/skill level of content description, in addition to providing clarity for alignment of curriculum and classroom assessment, also makes it easier to compare expectations across documents. As with content standards, performance expectations should reflect the consensus among educators as to what students should be able to do at various points in their schooling. This is especially important if states and districts wish to be clear that their standards for students are equal to the best in the country or the world. Fortunately, the use of the taxonomy also simplifies this comparison task. For example, analysts at McREL, when asked to provide recommendations as to level of difficulty for a set of benchmarks, review the source documents — national standards in the subject area, released assessment items from national and international assessments, as well as highly regarded state standards — to identify the level of performance most commonly associated with the knowledge or skill. It is often the case that it is relatively easy to identify commonly held expectations in terms of the taxonomy, as in Table 6.

Table 6. Commonly Held Expectations among Standards Documents

Benchmark: Knows that a variable is a letter or symbol that stands for one or more numbers	
Information from Source Documents	**Taxonomic Level**
Represent the idea of a variable as an unknown quantity using a letter or a symbol (Grades 3–5, NCTM, p. 158)* Use concrete objects and symbols to model the concepts of variables, expressions, equations, and inequalities (Grade 3, CBE, p. 189)* Use concrete objects and combinations of symbols and numbers to create expressions that model mathematical situations (Grade 5, CBE, p. 189)* Use rules and variables to describe patterns, functions, and other relationships (Grade 5, CBE, p. 189)* Use letters and symbols to represent an unknown quantity in a simple mathematical expression (Grade 4, NAEP, p. 32)* Use letters, boxes, or other symbols to stand for a number, measured quantity, or object in a simple situation with concrete materials, i.e., demonstrate understanding and use of a beginning concept of a variable (Elementary, New Standards, p. 62)*	Level 2: Comprehension: Representation

*CBE: *Standards for Excellence in Education* (1998); NAEP: *Mathematics framework for the 2005 National Assessment of Educational Progress* (September 2004); NCTM: *Principles and Standards for School Mathematics* (2000); New Standards: *Performance Standards: English Language Arts, Mathematics, Science, Applied Learning, Volume 1, Elementary School* (1997).

From Table 6, one can deduce that the expectation most commonly associated with variables in the 3–5 grade band is that students should be able to represent their conceptions of the variable. In a case like the one just described, the assignment of the benchmark to Comprehension: Representation is relatively straightforward.

Sometimes, however, the expectations will vary significantly by document. An example is provided in Table 7. The content concerns the basic measures of perimeter, area, volume, capacity, mass, angle, and circumference. In this case, performance expectations relative to the content are found across all levels — retrieval, analysis, and knowledge utilization. In such a case, those who are seeking to establish performance expectations may determine what level of understanding they agree is most suitable for the content at a given grade. An alternative, which recognizes the hierarchical nature of the taxonomy and thus the information it organizes, is to assign each more difficult level to a correspondingly higher grade. For example, educators may decide that using direct methods of measurement are appropriate at the earlier grades, but problem solving and using other methods, or solving problems using these methods, are appropriate at later grades.

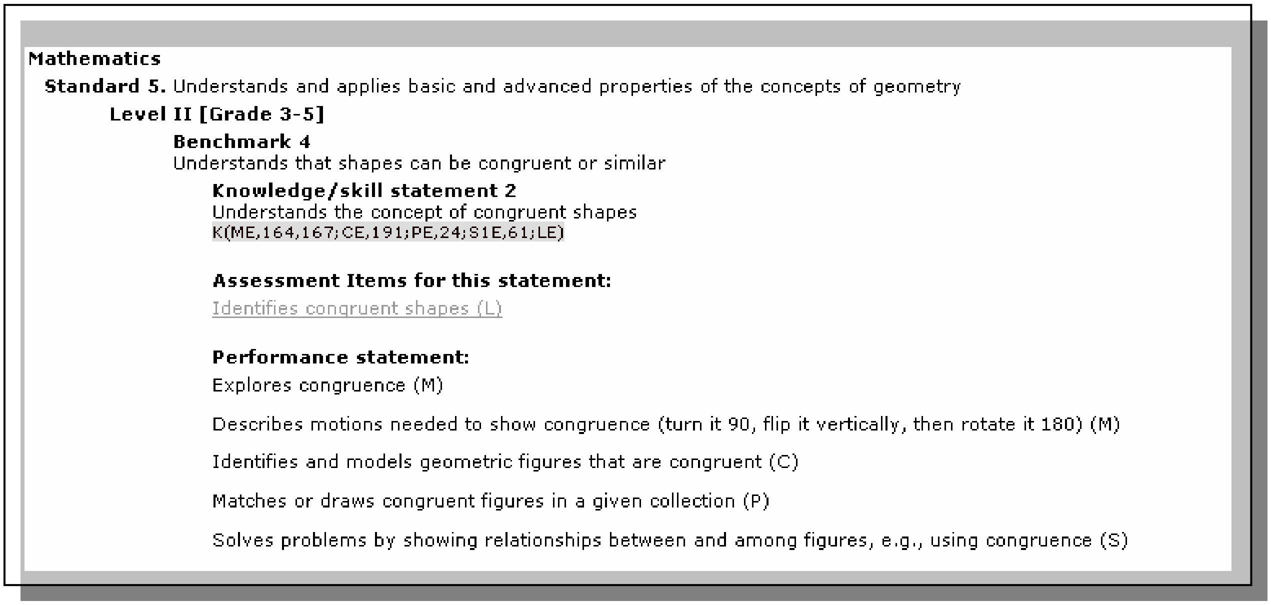
Clearly, researching all the documents such as those identified in Tables 5 and 6 could be a prohibitively labor-intensive task for educators who wish to make informed decisions about the expectations they hold for students. Fortunately, such information is available online. By late 2005, visitors to McREL's online standards database will find information of this type provided for mathematics and language arts from kindergarten through grade 8. This research identifies for each knowledge or skill statement in the database the performance expectation expressed in key documents. Exhibit 1 provides an example of what visitors to the site will find. Expressions that indicate performance levels have been paraphrased or excerpted directly from standards documents, which are identified by code in the parentheses following each statement. Using this information, educators can determine what level of expectation they should consider when reviewing content, whether they are using Marzano's or some other taxonomy.

Table 7. Differing Expectations among Standards Documents

Benchmark: Understands the basic measures of perimeter, area, volume, capacity, mass, angle, and circumference.	
Information from Source Documents	**Taxonomic level**
Select appropriate type of unit for measuring area, weight, volume, and size of angle (Grades 3–5, NCTM, p. 170) Find the perimeter and area of rectangles with direct methods, including using concrete objects as tools (Grade 3, CBE, p. 187) Determine area of polygon using shape to measure (Grades 3–4, IEA, TIMSS item #U-1) Estimate the size of an object with respect to a given measurement attribute, such as length or perimeter (Elementary, New Standards, p. 61)	Level 1: Retrieval: Execution
Demonstrate conservation of area (CBE, Grade 3, p. 187)	Level 2: Comprehension: Representation
Distinguish between area and perimeter, finding both using a variety of methods (CBE, Grade 5, p. 191) Compare objects with respect to a given attribute, such as area or volume (Elementary, New Standards, p. 61)	Level 3: Analysis: Matching
Given the perimeter, determine length of side (Grade 4, NAEP item #10) Solve problems involving the perimeter and area of plane figures (Grade 4, NAEP, p. 20)	Level 4: Knowledge Utilization: Problem Solving

*CBE: *Standards for Excellence in Education* (1998)*;* NAEP: *Mathematics framework for the 2005 National Assessment of Educational Progress* (September 2004); NCTM: *Principles and Standards for School Mathematics* (2000); New Standards: *Performance Standards: English Language Arts, Mathematics, Science, Applied Learning, Volume 1, Elementary School* (1997); IEA: *TIMSS mathematics items: Released Set for Population 1 (Third and Fourth Grade)*(1998).

In this section, we have reviewed the various levels of specificity at which performance expectations can be established, noting that the most specific level appears to be best suited both for aligning curriculum and assessment and for comparing expectations across documents. We have also provided examples from standards documents to show how performance expectations can be inferred and used, in combination with a taxonomy, to ensure that expectations established for students reflect the best in current thinking.



Mathematics

Standard 5. Understands and applies basic and advanced properties of the concepts of geometry

Level II [Grade 3-5]

Benchmark 4
Understands that shapes can be congruent or similar

Knowledge/skill statement 2
Understands the concept of congruent shapes
K(ME,164,167;CE,191;PE,24;S1E,61;LE)

Assessment Items for this statement:
Identifies congruent shapes (L)

Performance statement:

Explores congruence (M)

Describes motions needed to show congruence (turn it 90, flip it vertically, then rotate it 180) (M)

Identifies and models geometric figures that are congruent (C)

Matches or draws congruent figures in a given collection (P)

Solves problems by showing relationships between and among figures, e.g., using congruence (S)

Exhibit 1. Screen shot of McREL's Resource for Determining Performance Expectations

COMMUNICATING ESTABLISHED EXPECTATIONS

Once established, performance expectations must be clearly communicated in order to create a performance standard. Communicating performance expectations can range from identifying the performance expectation according to its taxonomic level, to providing a sample performance assessment, to providing a rubric at the topic or more specific benchmark level. Ideally, a performance standard would include the foregoing plus examples of student work that exemplify performances that meet and do not meet the expectations. Providing a fully elaborated performance standard for each benchmark would be prohibitively expensive for most schools, districts, and even states. A few examples of common approaches — identifying taxonomic levels, providing sample performance tasks, and providing benchmark-based rubrics — are described here.

Identifying the Taxonomic Level

McREL assisted the department of education in American Samoa in developing benchmarks and with the identification of levels of difficulty. Analysts provided recommendations for performance expectations, using the techniques described earlier. That is, they researched the expectations as expressed in highly-regarded reference documents, as well as checked them against the client's currently held expectations. An example of the resulting work is provided in Table 8. The organization of standards in American Samoa calls for a topic (which American Samoa happens to call a benchmark) for a band of grades — in the sample table, grades one through four — and specific expectations expressed for each grade. Appended to each grade-specific expectation in square brackets is a common language term or phrase that identifies the level of Marzano's taxonomy selected as appropriate for that content at that grade. For example, in 2nd grade, students are expected to be able to describe the relationships between categories that distinguish different types of data. This level of expectation is equivalent to the taxonomy level Analysis: Matching. The abbreviated reference to the taxonomy for this skill is indicated in the bracketed phrase "compares facts." This common language term is used to reduce the amount of technical language that teachers would have to understand in order to use the standards document. In 4th grade, students are expected to use tables and charts to the make predictions. The bracketed term, "prediction" formally indicates that the level of difficulty selected in the taxonomy is Analysis: Specifying. In the taxonomy, Marzano says about this level:

> The analysis skill of specifying involves making and defending predictions about what might happen or what will necessarily happen in a given situation. (p. 81).

The translation of these terms — from Analysis: Matching to "compare facts" and from Analysis: Specifying to "prediction" — helped make the performance level more immediately apparent, with no loss in precision. Because the equivalences between terms used and the taxonomic table were established at the outset and provided in a reference table, there was no ambiguity about which term indicated which level in the taxonomy.

Table 8. Communicating the Taxonomic Level in American Samoa

	BY THE END OF			
	1ST GRADE	***2ND GRADE***	***3RD GRADE:***	***4TH GRADE:***
Benchmark 4.2: *Collects, organizes, and reads data in charts, graphs, and tables.*	**The student:** 1. Knows ways to sort, represent, and compare objects using concrete materials. [Describes steps of a skill] 2. Uses simple tables, charts, and graphs to represent and compare objects. [Represents data]	**The student:** 1. Sorts data into categories and describes their relationships. [Compares facts] 2. Uses appropriate sampling techniques for gathering data. [Describes steps of a skill] 3. Organizes data into charts using tally marks. [Represents data]	**The student:** Collects and organizes simple data using pictographs, tables, charts, and bar graphs. [Represents data]	**The student:** Uses tables, charts, and graphs to make predictions and draw conclusions about data. [Predicts]

A different approach was used for communicating performance expectations for standards in the state of Hawaii. This approach has some affinity with the use of Bloom's original taxonomy, in that verbs alone are used to indicate levels of expectation. However, while verb lists from Bloom's provide only generic guidance. In this approach, the choice of verb phrases is more informative, because their selection depends upon whether the performance is related to declarative or procedural knowledge. For example, students might be asked to describe the relationship between ideas (declarative knowledge) or execute the steps in a procedure (procedural knowledge).The verbs are organized according to Marzano's structure, outlined in Table 1. The sample performance assessment (Table 9) provides an example of what that taxonomic level might mean.

The use of the introductory verb "recognize" identifies this content as appropriate for the Retrieval: Recall level of the taxonomy. The Sample Performance Assessment supports this, in that it outlines a task that would test one aspect of the benchmark — whether or not the student recognizes the orientation of words on a page.

Table 9. Communicating the Taxonomic Level in the Hawaii Standards*

<table>
<tr><th colspan="2">BENCHMARK</th><th colspan="2">SAMPLE PERFORMANCE ASSESSMENT</th></tr>
<tr><td colspan="2">[Students] Recognize that spoken words correspond to printed words, how letters and words are oriented on the page, and that words are read from left-to-right across the page situation with a number sentence</td><td colspan="2">The student:
Follows text from left to right and from top to bottom of a page as it is being read aloud; locates the front cover, title, and back cover of a book; and demonstrates knowledge of a book's orientation by holding and opening the book correctly.</td></tr>
<tr><th colspan="4">RUBRIC</th></tr>
<tr><th>ADVANCED</th><th>PROFICIENT</th><th>PARTIALLY PROFICIENT</th><th>NOVICE</th></tr>
<tr><td>Consistently recognize that spoken words correspond to printed words, how letters and words are oriented on the page, and that words are read from left-to-right across the page</td><td>Usually recognize that spoken words correspond to printed words, how letters and words are oriented on the page, and that words are read from left-to-right across the page</td><td>Sometimes recognize that spoken words correspond to printed words, how letters and words are oriented on the page, and that words are read from left-to-right across the page</td><td>Rarely recognize that spoken words correspond to printed words, how letters and words are oriented on the page, and that words are read from left-to-right across the page</td></tr>
</table>

*Excerpted and adapted from Hawaii Department of Education's *Content Standards for Mathematics K–12* (August 2005)

Using Rubrics to Communicate Performance Levels

A rubric is commonly used to provide a set of performance levels (see Table 9). Rubrics, originally developed to score performance tasks such as written essays, have recently been adopted for more general use. The rubrics in Table 9 help to clarify what teachers might expect not only for the proficient level, but levels above and below proficiency. This particular rubric was written to address the taxonomic level of Recall: Recognition, and was constructed with the assumption that the common issue is how consistently the student would be able to employ the skill. Note that the special value of this approach is that the taxonomic level was selected first to determine what level of proficiency was expected, and then the rubric was developed, working from the proficient level to the levels on either side. This is a much more specific rubric than used, for example, in the National Assessment for Educational Progress Achievement Levels (see Figure 1). The NAEP levels are much more broadly defined, and expressly cover not only different topics in one rubric (such as understanding of fractions) but both procedural and declarative knowledge (such as understanding decimals and how to use a four-function calculator). Such a rubric is useful to communicate broadly about a related set of abilities, but it has limited use for the classroom.

Rubrics that address declarative and procedural knowledge separately are for that reason more specific than the rubric in Figure 1. Tables 10 and 11 are examples of generic rubrics for declarative and procedural knowledge, respectively.

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

Figure 1. NAEP 4th grade Achievement Level Description for Reading

Table 10. Generic Rubric for Procedural Knowledge

Advanced	Executes the process or skill effortlessly, efficiently, or with no errors
Proficient	Executes the process or skill with some effort, or acceptable efficiency, or with no significant errors
Below Proficient	Executes the process or skill with significant effort, or with some inefficiency, or with a few significant errors
Novice	Executes only some of the process or skill and with significant effort, or little efficiency, or with many significant errors

Table 11. Generic Rubric for Declarative Knowledge

Advanced	Has a comprehensive understanding of the information or concept and its applicability in a variety of contexts
Proficient	Understands the most significant aspects of the information or concept and can apply it within a narrow but appropriate context
Below Proficient	Understands some but not all significant aspects of the information and applies in both inappropriate and appropriate contexts
Novice	Understands few significant aspects of the information or concept and applies it within inappropriate contexts

Although rubrics at this generic level may be useful in providing an overview of relevant skill or knowledge, they are still less useful in providing the kind of direction needed at the classroom level. Yet they can prove a useful template when deciding how to develop scoring rubrics for a specific task.

In work with states and school districts, we have found that rubrics tend to fall into one of a limited number of categories. Depending upon the focus of the benchmark and the performance expectation, a rubric might focus on consistency, difficulty, error, level of detail, significance of detail, quality, speed and fluency, or variety. Each of these categories is discussed separately below. Table 12 provides suggested terms and phrases that may help in constructing rubrics that are consistent within categories and across score points.

Table 12. Suggested Rubric Categories with Qualifying Terms and Phrases

CATEGORY	ADVANCED PROFICIENT	PROFICIENT	PARTIALLY PROFICIENT	NOVICE
Consistency	Always	Frequently	Occasionally	Rarely
	Consistently	Usually	Sometimes	Seldom
	Nearly always	Often	Inconsistently	Occasionally
Difficulty[2]	Utilize	Analyze	Comprehend	Recall
	Analyze	Comprehend	Recall	Recognize
Error	Flawlessly	Minor errors	Some significant errors	Errors prohibit performance or understanding
	Correct computations	Computations essentially correct	Minor computational errors	Serious computational errors
	With accuracy	With no significant errors	With a few significant and/or many minor errors	With many significant errors
	No error	Few errors	Consistent errors	Numerous errors
	Correct with support	Correct	Partially correct	Incorrect
Level of Detail	[Describe] in great detail	[Describe] in detail	[Describe] in some detail	[Describe] in minimal detail
	Compete and detailed understanding	Complete understanding, but not in great detail	Incomplete detail and /or some misconceptions	Understanding so incomplete that student cannot be said to understand
	[Describe] in comprehensive detail	[Describe] in adequate detail	[Describe] in moderate detail	[Describe] in minimal detail
	With thorough support	With support	With partial support	With very little support
	[Describe] with well-supported detail	[Describe] with supported detail	[Describe] with weakly associated details	[Describe] with unrelated details

[2] This structure anticipates a revision of Marzano's Taxonomy that will distinguish recall from recognition. See the Appendix.

CATEGORY	ADVANCED PROFICIENT	PROFICIENT	PARTIALLY PROFICIENT	NOVICE
Quality	Cogent	Logical	Logic hard to follow	No apparent logic
	Highly effective	Effective	Limited	Ineffective
	Highly Original	Creative	Obvious/ Typical	Imitative
	Innovative	Appropriate	Trivial	Inappropriate
	Highly effective	Efficient	Limited	Unrealistic
	Engaging	Relevant	Somewhat relevant	Irrelevant
	Sophisticated	Appropriate	Simplistic	Ambiguous
	Precise	Clear	Clichéd / Generic	Vague/ Unclear
	Strong	Effective	Marginal	Weak
	Uses novel information	Uses appropriate information	Uses information not related to problem	Uses wrong information
	[Demonstrates] new insight	[Demonstrates] a complete understanding	[Demonstrates] some misconceptions	[Demonstrates] significant confusion
	[Demonstrates] extended understanding or ability	[Demonstrates] a complete understanding or ability	[Demonstrates] an some understanding or ability	[Demonstrates] minimal understanding or ability
Significance of Detail	Includes all important characteristics	Includes the most important but not all the characteristics	Includes some trivial and important characteristics	Includes mostly trivial characteristics
	[Identifies] main pattern and all minor patterns in information	[Identifies] main pattern in information	[Identifies] some features of main pattern	[Identifies] very few features of the main pattern
	[Places] all in sequence	[Places] all significant items in sequence	[Identifies] all significant items but sequences them incorrectly or places only some items in correct sequence	[Identifies] some items but not in correct sequence
	[Identifies] all the significant items/details as well as subtleties	[Identifies] the significant items/details	[Identifies] some significant items/details	[Identifies] very few significant items/details
	Uncovers hidden/implied information	[Identifies] all information related to topic	[Identifies] information obviously related to topic	[Identifies] information not related to topic

CATEGORY	ADVANCED PROFICIENT	PROFICIENT	PARTIALLY PROFICIENT	NOVICE
Speed and Fluency	With fluency/ Automatically	With minimal hesitancy	With some hesitancy	With much hesitancy
	With ease	With minimal difficulty	With difficulty	With great difficult
	Independently as needed	Independently upon request	With assistance	With much assistance
Variety	An extensive variety	A variety	A few	One or two/ A limited number
	Use numerous [tools/strategies]	Use an appropriate number of [tools/strategies]	Use some [tools/strategies]	Use very few [tools/strategies]

The categories and descriptors provided in Table 12 were developed primarily from our work with districts and states, but also from consulting the works of experts in assessment, including Arter & McTighe (2001), Bailey & Guskey (2001), and Marzano (2000). A common concern regarding rubric development is that the descriptions should be internally consistent, and the differences between one score point and another should be clear (Tierney & Simon, 2004). Rubric guidelines must remain flexible to adequately address these issues, yet should be structured well enough so that the difference between one achievement level and the next is clear.

Rubrics with Consistency as the Focus. A rubric for consistency is useful when it is important to be sure that the student has clearly acquired knowledge or a skill and can use it over a period of time and/or in a variety of situations. Because the acquisition of a skill implies practice over time, rubrics for consistency are most commonly associated with skills or processes. Because it implies numerous observations (whether in person, or through multiple paper/pencil assessments), the consistency rubric is not useful for measuring a single performance, although the consistency rubric works well as a summative assessment of a skill or process. As an example, consider the skill of choosing the correct verb tenses in writing. A student might be able to apply the correct verb tense to a set of pre-written sentences directly after instruction on verb tenses, but fall back to using incorrect tenses at a later time in his or her own writing. Similarly, a student might use correct verb tenses in part, but not all, of a paper. In these cases a consistency rubric could be used to identify the skill level of a student who does not consistently use correct verb tenses, whether over time or throughout an entire assignment.

Rubrics with Difficulty as the Focus. A rubric for difficulty takes advantage of the fact that a well-designed taxonomy of educational objectives can be used to determine the level of difficulty appropriate for the mastery of the content, as well as to characterize the nature of difficulty above and below the desired level of performance. This is different from the other rubric types, which start with the assignment of a taxonomic level for proficient performance and characterize how student achievement might vary from that target in terms of consistency of execution, number of errors, level of detail, and the like. An important result of using the taxonomy itself to distinguish levels of a rubric is that it is unlikely all students could be scored using the same assessment. For example, if a student successfully described how plants depend on animals for their survival —thus achieving the Comprehension level selected for Proficient — it would not be clear from the student's description whether he or she could also have *classified* plants by their dependence on animals, thus reaching Advanced Proficient, which reflects the next higher level, Analysis: Classifying. Discrimination between score points when using the rubric for difficulty might often require different tasks for each score point. This might not be of great concern to teachers, who can use multiple assessments to make their judgments. For those who develop external assessments, scoring students on such rubrics might present a special challenge.

Rubrics with Error as the Focus. An error rubric may be applied to a knowledge or skill that has a definite right and wrong answer. Traditionally this type of content is assessed as either correct or incorrect. With a four-point rubric, however, the degree of proficiency can be determined by either the significance of the error made or by the overall number of mistakes made. For example, elementary math students who are learning the names of geometric shapes will either be able to name each shape correctly or not; but the total number of mistakes that a student makes while naming multiple shapes will indicate how well he or she has mastered the concept. On the other hand, high school math students who are learning proofs in geometry may make a small mistake in their calculations or a large mistake that reveals a lack of understanding. Such a student could be graded on the rubric continuum of working flawlessly, with minor errors, with some significant errors, or with so many errors that understanding is clearly not present.

Rubrics with Level of Detail as the Focus. Using a level of detail rubric will allow teachers to assess the depth of student knowledge about a particular topic. Rubrics about the level of detail work well on content that has many layers of information. In such cases, the teacher must determine what the desired degree of knowledge about the topic is and use that expectation as the basis for a proficient performance. An example of an appropriate topic for a level of detail rubric might be the Underground Railroad during the United States Civil War. A novice student might be able to identify the purpose of the Underground Railroad only; whereas an advanced student might be able to describe the role that the Railroad played in the abolitionist movement, citing specific events, influential individuals, and various impacts.

Rubrics with Significance of Detail as the Focus. A significance of detail rubric should be used when content is somewhat complex, containing much information, some of which is important, some of which is not. This rubric describes the proficient level in terms of whether the most significant, though not necessarily all significant details, are identified. This rubric can also focus on the significant relationships within a concept, rather than details. Significance of detail rubrics are not effective for benchmarks whose scope can be described with a fixed or definitive list of items, when all items are of equal importance, or with a narrow topic. This rubric is not about the quality or extent of detail the student uses, but about understanding the relative significance of the details or relationships has identified. For example, a benchmark that requires students to summarize a text might easily lend itself to a significance of detail rubric, as students would be expected to differentiate between significant and insignificant information as they write the text.

Rubrics with Quality as the Focus. The quality rubric should be used when a student is expected to explain, design, produce, or create a product with no easily discernable "right" answer. Quality rubrics should not be used when there is an easily discernable "right" answer. The descriptions used to identify levels of performance in quality rubrics should be flexible and accurately describe the characteristics or content of the desired product. Quality rubrics make much use of adjectives to describe levels of performance. For example, a teacher might develop a quality rubric to assess the ideas in a piece of creative writing by choosing descriptive words for each level of performance, such as engaging, effective, limited, and unclear.

Rubrics with Speed and Fluency as the Focus. Rubrics that have fluency as their focus are most appropriate when the speed of recall and automaticity of a skill are of primary importance. (The automaticity of a skill refers to exercising a skill without conscious thought.) For instance, a student should be able to quickly remember multiplication facts, or automatically "Apply knowledge of letter/sound relationship when reading." Benchmarks most suited to fluency usually focus on basic knowledge and skills in math and reading. An important characteristic about such rubrics is that they are used in a one-on-one assessment setting since it is much easier to witness the fluency a student demonstrates than it is to make an inference about this from written work.

Rubrics with Variety as the Focus. A rubric that focuses on variety is appropriate when there is a set of ideas commonly associated with a topic, or a variety of contexts associated with a topic that students should know. For example, benchmarks best assessed using this rubric might use terms such as "different ways," "identify ways," or "some point made". In some instances, the content of the benchmark might actually refer to a very limited list of items, and thus a variety rubric would not be an appropriate choice. For example, a benchmark that requires students to compare or identify two or three items would not be appropriate for a variety rubric since the list of items is extremely limited.

SUMMARY

In this paper we have described the three critical steps to establishing a performance standard. The first is to develop or adopt a systematic method for describing levels of difficulty relative to the mastery of knowledge and skill. The second is to become informed about and reach consensus regarding the appropriate performance expectation for specific descriptions of student knowledge and skill. The third is to select and use a format that clearly communicates this performance to those we will benefit from its use.

We have established distinctions between the performance standard and performance expectation, described a method for identifying performance expectations in representative standards documents, and previewed a McREL resource that helps inform the process of determining appropriate expectations for student work. In addition, we have provided a new typology of rubrics that can be used to improve how educators communicate performance standards.

Clearly setting and communicating performance standards can be challenging. But it is a necessary step to clarify the expectations held for all students and to determine the appropriateness of classroom instruction and assessment for helping students meet desired outcomes.

REFERENCES

Anderson, J. R. (1993). *Rules of the Mind.* Hillsdale, NJ: Erlbaum

Anderson, L. W. , Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete ed.). New York: Longman.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: using performance criteria for assessing and improving student performance.* Thousand Oaks, CA: Corwin Press, Inc.

Bailey, J., & Guskey, T. (2001). *Developing grading and reporting systems for student learning.* Thousand Oaks, CA: Corwin Press, Inc..

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain.* New York: David McKay.

Council for Basic Education. (1998b). *Standards for excellence in education.* Washington, D.C.: Author.

Council of Chief State School Officers. (1998). *Tool kit: Evaluating the development and implementation of standards.* Retrieved from http://www.ccsso.org/tk98.html

Council of Chief State School Officers. (1992). *Provisional item specifications: 1994 National Assessment of Educational Progress in U.S. History.* Washington, DC: Author.

Hawaii Department of Education. (August 2005). *Content standards for Mathematics K–12.* Honolulu, HI: Author.

International Association for the Evaluation of Educational Achievement's Third International Mathematics and Science Study. (1998a). *TIMSS mathematics items: Released set for population 1 (third and fourth grade).* Retrieved March 1, 2000, from http://timss.bc.edu/TIMSS1/TIMSSPDF/AMitems.pdf

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kendall, J. S. (2000). Topics: A roadmap to standards. *NASSP Bulletin, 84*(620), 37–48.

Kendall, J.S. & Marzano, R.J. (2004). *Content Knowledge: A Compendium of Standards and Benchmarks for K–12 Education.* (4th ed., Online). Retrieved June 1, 2005 from http://www.mcrel.org/standards-benchmarks/

Laufer, B. & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning* 54: 469-523

Spearman, C. (1927). The Abilities of Man, Their Nature and Measurement: NY: Macmillan.

Lauer, P. A., Snow, D., Martin-Glenn, M., Van Buhler, R. J., Stoutemyer K., & Snow-Renner, R. (2005). *The influence of standards on K–12 teaching and learning: A research synthesis.* Aurora, CO: Mid-continent Research for Education and Learning.
Kendall, J. S. (2000). Topics: A roadmap to standards. *NASSP Bulletin, 84*(620), 37–48.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives.* Experts in Assessment Series, Guskey, T. R., & Marzano, R. J. (Eds.). Thousand Oaks, CA: Corwin Press.

Marzano, R. (2000). *Transforming classroom grading.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., & Kendall, J. S. (1996). *A comprehensive guide to designing standards-based districts, schools, and classrooms.* Alexandria, VA: Association for Supervision and Curriculum Development.

National Assessment for Educational Progress. (2005). *Released items for the 2003 NAEP assessment.* Retrieved July 1, 2005 from http://nces.ed.gov/nationsreportcard/ITMRLS/NQT_Search.asp

National Assessment Governing Board.(September 2004). *Mathematics framework for the 2005 National Assessment of Educational Progress.* Retrieved September 1, 2005 from http://www.nagb.org/pubs/m_framework_05/761607-Math%20Framework.pdf

National Center for History in the Schools. (1996). *National standards for history.* (Basic ed.). Los Angeles: Author

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author

The National Education Standards and Improvement Council. (1993). *Promises to keep: Creating high standards for American students. Report on the review of educational standards from the Goals 3 and 4 Technical Planning Group to the National Education Goals Panel.* Washington, DC: National Goals Panel.

New Standards. (1997). *Performance standards: English language arts, mathematics, science, applied learning, volume 1, elementary school.* Washington, DC: National Center on Education and the Economy.

Porter, A. C. (2002, October). Measuring the content of instruction: Uses in research and practice. *Educational Researcher 31*(7), 3–14.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Retrieved August 1, 2005 from http://PAREonline.net/getvn.asp?v=9&n=2

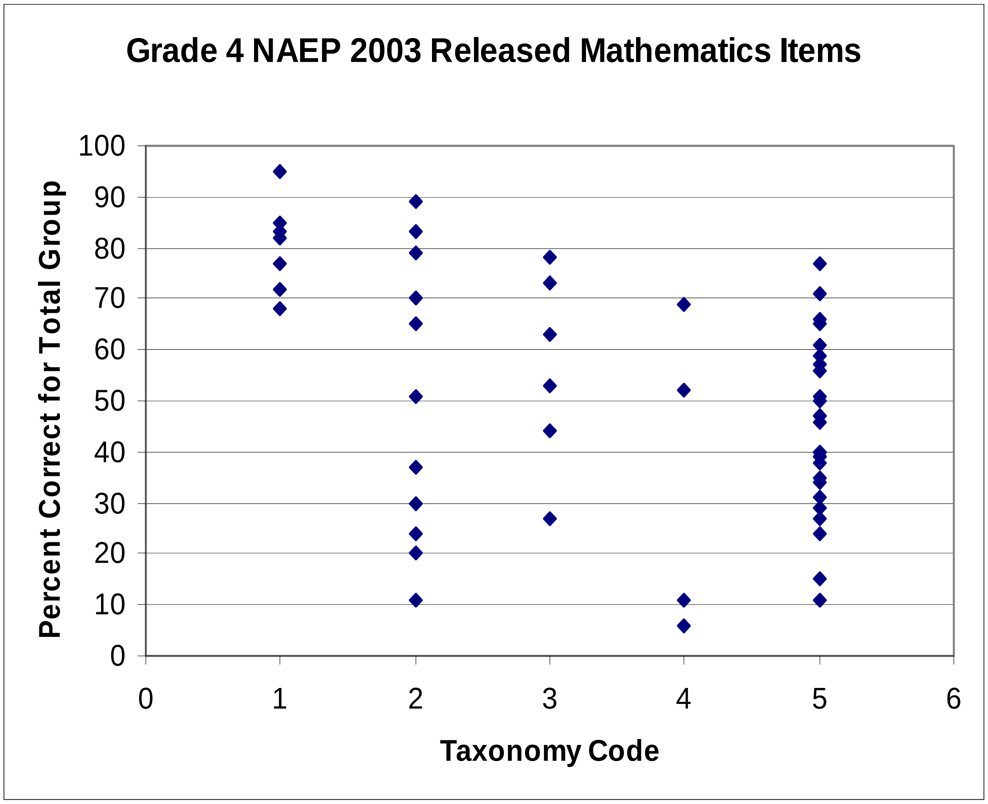
Webb, N. L. (1997, April). *Criteria for alignment of expectations and assessments in mathematics and science education.* Research monograph No. 8. Washington, DC: Council of Chief State School Officers.

APPENDIX

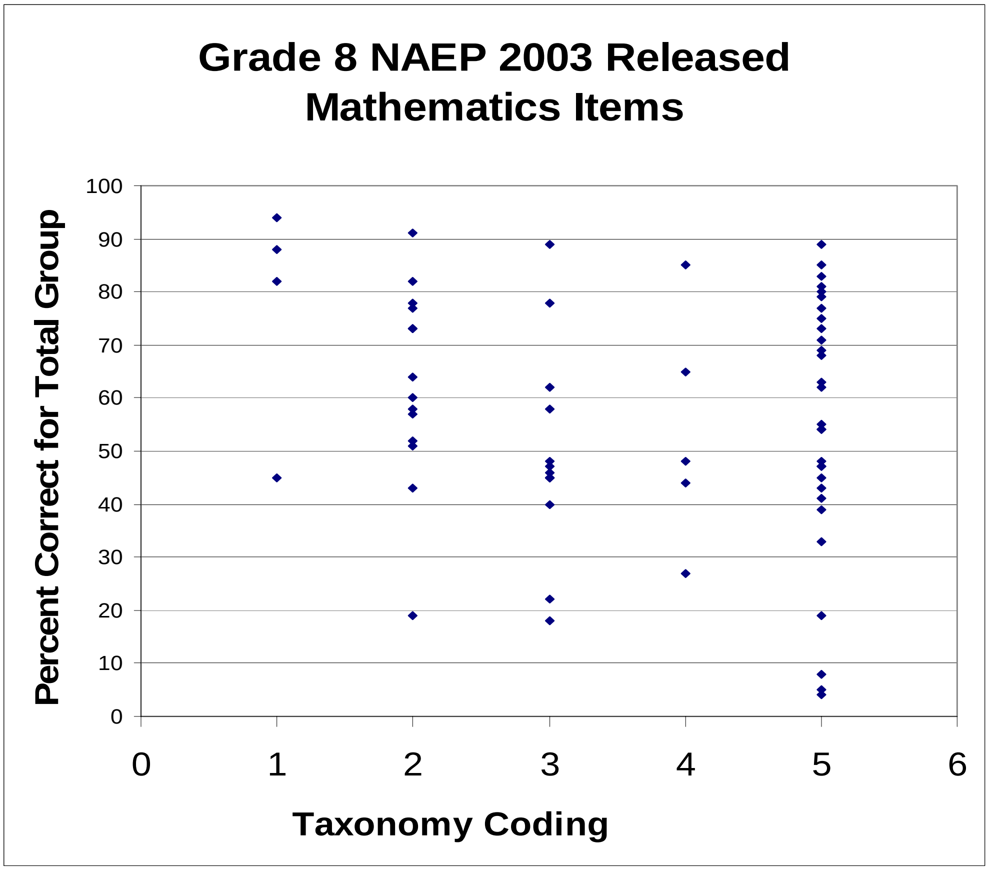
Content analysts reviewed a set of released test items from grade 4 and 8 of the NAEP 2003 mathematics assessment. These items were assigned a "Taxonomy Code" based upon Marzano's Taxonomy. For the purposes of analysis, Level 1 of the Taxonomy was divided into two levels, corresponding to recognition and recall. There is empirical support that these two are distinct in terms of level of difficulty, (Laufer & Goldstein, 2004), and they will likely be incorporated into the next version of Marzano's Taxonomy (personal communication, 2005). The following taxonomic codes are used in Table 1:

1. Recognition
2. Recall
3. Comprehension
4. Analysis
5. Utilization

Level 6, Goal setting and monitoring, and level 7, Self-system monitoring, did not occur and were not assigned to any items.



Based on 54 items, the prediction of item difficulty from taxonomy level is significant at the p<0.01 level. The multiple correlation of R= 0.44 indicates that approximately 19% of the variance in item difficulty for this set of items is accounted for by the coded taxonomic levels of the items.



Based on 65 items, the prediction of item difficulty from taxonomy level is not significant.