Measured Measures

Technical Considerations for Developing a Local Assessment System

Prepared by the
Maine Comprehensive Assessment System
Technical Advisory Committee

June 2000

Maine Comprehensive Assessment System Technical Advisory Committee

Theodore Coladarci (Chair)
University of Maine

Judith L. Johnson
University of Southern Maine

Jeffrey Beaudry
University of Southern Maine

Michael Cormier
MSAD #9

Robert Ervin
Bangor School Department

Jill M. Rosenblum
Maine Mathematics and Science Alliance

David L. Silvernail
University of Southern Maine

Ex Officio Members:

Horace “Brud” Maxcy, Pamela G. Rolfe, and Paul “Randy” Walker
Maine Department of Education

Printed under appropriation 010 05A 0829 72. The Maine Department of Education insures equal employment, equal education opportunities, and affirmative action regardless of race, sex, color, national origin, religion, marital status, age, or handicap. Affirmative Action Coordinator, Maine Department of Education, 23 State House Station, Augusta, ME 04333.

Table of Contents

Acknowledgments

Measured Measures involved contributions from various individuals, and in various ways. Parts I, II, and III were written by me, Judith Johnson was the primary author of Part IV, and Jill Rosenblum crafted the "snapshots" and prepared Appendix A. Further, each draft of *Measured Measures* enjoyed extended—and rather animated—discussions at meetings of the Technical Advisory Committee (TAC). The thoughtful and constructive feedback of my TAC colleagues was invaluable in shaping *Measured Measures*, and I am indebted to them for their wise counsel and unfailing commitment to this project.

Measured Measures also benefited enormously from a focus group that was conducted in the fall of 1999. The Maine Department of Education convened a group of educators who were asked to read a draft of the document, respond in writing to feedback questions that the TAC had posed, and participate in a 11-17-99 focus group to share their perceptions and to offer suggestions for improvement. We were encouraged by the group's general endorsement of *Measured Measures*, and we welcomed their many recommendations for making the final publication inviting and pleasing to the eye. On behalf of the TAC, I extend sincere thanks to all focus group members: Karoldene Barnes (SAD 64), Sue Card (Auburn School Department), Madeline Clement (Union 34), Patti Duran (Union 34), Brenda Felch (Caribou School Department), Gail Gordon (Union 34), Frank McElwain (Caribou School Department), Louise Regan (SAD 63), Molly Schen (Auburn School Department), Sandra Schniepp (SAD 58), Ginny Secour (SAD 48), and Art Turner (SAD 17). Thanks also are due to Donna Asmussen and Mona Baker, who masterfully facilitated the focus group.

Last and decidedly not least, Amy Cates, showing the keen eye, impeccable judgment, and infectiously pleasant demeanor for which she is well known, was responsible for laying out and formatting the document now before you.

Ted Coladarci
University of Maine

PART I

Introduction

PART I
Introduction

The Nature of Measured Measures

When the Maine State Legislature established the *Learning Results* in 1996, it decreed that "[s]tudent achievement of the *Learning Results* . . . must be measured by a combination of state and local assessments to measure progress and ensure accountability." Similar language appeared in the 1997 legislation mandating that the State Board of Education develop a plan for funding programs and services deemed essential for achieving the *Learning Results*. "The plan," wrote the Maine lawmakers, "must include establishment of a *system* to measure and ensure that schools are held accountable of student *Learning Results*" (emphasis added).

What is meant by state assessment, local assessment, and a system of assessment? The Maine Educational Assessment (MEA), recently revised to align with the *Learning Results* (Maine Department of Education, 1997), constitutes the state assessment. Logically enough, local assessment involves assessment activity initiated at the classroom, school, and district level. And a local assessment system, as we will detail below, is a constellation of assessments that, together, yield data that document progress toward student mastery of the *Learning Results* and other learning targets.

A local assessment system is a constellation of assessments that, together, yield data that document progress toward student mastery of the Learning Results and other learning targets.

The purpose of *Measured Measures* is to describe several technical considerations for developing, using, and monitoring a local assessment system. To be useful for any educational purpose—informing instruction, providing student feedback, evaluating programs, certifying student achievement, and so on—an assessment must embody agreed upon levels of *validity* and *reliability*. Further, such assessments must carry explicit *performance standards* that are consistently employed. We have designed *Measured Measures* to help Maine educators gain an understanding of these technical qualities and, in turn, their applicability to the design and conduct of local assessment systems.

Why are these technical considerations—validity, reliability, and performance standards—essential qualities of a local assessment system? Documenting student mastery of the *Learning Results* is not an easy task, nor is developing a local assessment system. Both require that local assessments measure what they are intended to measure (validity), that they measure it consistently (reliability), and that the interpretation of assessment results is guided by clear definitions of acceptable student performance (performance standards). *Measured Measures* describes these technical features and, further, provides strategies for incorporating them into a local assessment plan.

We wish to emphasize several points before proceeding. First, a local assessment system has several key elements: (a) technically sound assessments of academic achievement; (b) adequate professional development regarding assessment principles and practices; and (c) explicit and well-coordinated mechanisms for managing assessments, assessment results, and professional development. *Measured Measures* deals only with the first of these key elements: the technical quality of assessments that make up a local assessment system.

Our intended audience admittedly is a somewhat select group: those educators who have expertise in the discipline of assessment and who are involved in assessment-related initiatives at the local or state level.

Second, our intended audience admittedly is a somewhat select group: those educators who have expertise in the discipline of assessment and who are involved in assessment-related initiatives at the local or state level. Although we believe that any educator can benefit from a careful reading of *Measured Measures*, we envision that, locally, this task will fall primarily on the shoulders of an individual or small team who, in turn, may use the document as a framework for staff development, local assessment initiatives, and related activities.

Measured Measures is not meant to be an exhaustive treatment.

Third, *Measured Measures* is not meant to be an exhaustive treatment. Indeed, no single volume can accomplish such an objective, particularly when the topic is as complex as the present one. Rather, we offer this document as a guide—a guide for educators as they develop, use, and monitor their local assessment system. Additional reading, reflection, and discussion will be necessary for any district to establish a local assessment system, regardless of the district's present level of assessment sophistication. Toward this end, many districts doubtless will profit from outside assistance (e.g., Maine Department of Education, private consultants), collaboration with neighboring systems, and other forms of support.

Fourth, because *Measured Measures* is prompted by the *Learning Results* legislation, you will find in the following pages many references to the *Learning Results*. However, one should not infer that only the *Learning Results* should be addressed in a local assessment system. Indeed, such a system also should accommodate locally defined objectives that exceed those included in the *Learning Results* framework. Further, the relevance of the technical issues we discuss below is not restricted to *Learning Results* assessment. "Validity," "reliability," and "performance standards" are important considerations in their own right, regardless of what learning targets have been identified.

Finally, the creation of a local assessment system requires considerable thought, effort, resources, and time. A local assessment system is not established quickly and in one fell swoop. It evolves. By attending to the considerations below, educators will be moving toward an assessment system—gradually and deliberately—that provides meaningful data for making inferences about student achievement and for guiding decisions about instructional practice and educational policy.

A local assessment system is not established quickly and in one fell swoop. It evolves.

We begin by describing several characteristics of a local assessment system. We then turn to the technical considerations of validity (Part II), reliability (Part III), and setting performance standards (Part IV).

What is a Local Assessment System?

A "system," *The American Heritage Dictionary* tells us, is "a group of interacting, interrelated, or interdependent elements forming a complex whole." Elements in a system must cohere in some fashion. What does this mean for a system of local assessment? A local assessment system is a coherent, coordinated plan for assessment. Clearly, a local assessment system is made up of individual assessments. But a collection of assessments does not entail a system any more than a pile of bricks entails a house. Thus the fundamental question is this: In what sense do local assessments constitute a *system* of assessments, rather than a mere collection of assessments? In our view, a local assessment system has six critical features:

A collection of assessments does not entail a system any more than a pile of bricks entails a house.

The assessments collectively are relevant to announced learning targets. A local assessment system provides evidence of student achievement regarding specified instructional objectives, or "learning targets." Learning targets ultimately are stated with sufficient specificity to communicate *measurable* outcomes. For example, the *Learning Results* "performance indicators" are stated at this level of specificity—they, for the most part, are directly amenable to measurement and assessment—whereas the corresponding "guiding principles" and "content standards" are not. To be sure, measurable outcomes can be *derived* from a guiding principle or content standard. But without this translation of the general to the measurable, assessment is exceedingly difficult at best. Whether locally defined or drawn from the *Learning Results*, then, learning targets must be in the form of measurable outcomes.

The assessments are conducted at multiple levels: classroom, school, district, and state. A local assessment system is made up of assessments that are initiated at the classroom, school, district, and state level.

SNAPSHOT

"Our district has barely begun to think about a local assessment system," said Theresa. "We've been working on our assessment program for over ten years," Alice added. The two women had met at a conference and were comparing notes.

Just two years ago, Theresa had been a middle school social studies teacher, and now she holds the position of curriculum coordinator for SAD 111. Alice serves as assistant superintendent in SAD 222. Although their backgrounds and experiences are different, and although their districts are in very different places in terms of assessment systems, they find themselves engaged in a conversation about technical considerations for local assessment systems. Both agree that with passage of *Learning Results* legislation, there came increasing pressure to use and report achievement data. "I want to be sure we're generating trustworthy data about performance on the *Learning Results*," said Theresa.

Alice noted that despite many years of developing, implementing, and refining assessments, her district hadn't had the benefit of clear strategies for establishing content validity with respect to standards. "We began our efforts before *Learning Results* came on the scene," she explained. "Now we need to work

continued

Classroom-level assessments reflect the day-to-day assessment practices of teachers, such as running records, unit exams, papers, projects, performances, and portfolios of work samples. School- and district-level assessments involve students across multiple classrooms. As an example, a district may administer a reading proficiency test to all second graders, or a science proficiency test to all eighth graders.

What is the role of the state-mandated MEA in a "local" assessment system? The realigned MEA is an important source of evidence regarding local progress toward student mastery of the *Learning Results*. While "MEA" and "local assessment" arguably is a contradiction in terms, "MEA" and "local assessment *system*" is not. Thus although state mandated, the MEA can be—and should be—considered an important component of any local assessment system.

The assessments are conducted at multiple grades. Teacher-initiated, classroom-level assessments occur at all grades. In contrast, school- and district-level assessments most likely occur at specific checkpoints for monitoring progress toward student mastery of the learning targets, as in a district-wide proficiency test administered at a particular grade.

The assessments draw on multiple methods—"traditional" and "alternative" alike. There are various ways to assess student learning, such as selected-response methods (e.g., multiple-choice, matching, or true-false items), constructed-response methods (e.g., worked problems, short answers, essays), and performance-based methods (e.g., projects, demonstrations). No one method is sufficient as a general assessment strategy. For example, selected-response items typically are superior to either constructed responses or performance measures for assessing recall and basic understanding of a large body of content, whereas the latter two methods are preferable over the former for assessing written, oral, or behavioral expression. Insofar as the content standards and performance indicators of the *Learning Results* represent a variety of outcomes, a local assessment system should comprise a variety of means for assessing those outcomes.

The assessment system allows for multiple opportunities to demonstrate knowledge, understanding, and skill development. A single administration of an assessment, whatever its form, typically provides an insufficient basis for making inferences about student proficiency with respect to identified learning targets. Inferences are more defensible when students have multiple opportunities to demonstrate proficiency. For instance, a performance assessment might be conducted at several points in time, or the learning targets might be assessed through a combination of assessment types.

The assessments have an announced rationale. Each assessment's *purpose*, *audience*, and *articulation* with other assessments in the system should be clearly stated. For example, consider teacher-initiated, classroom-level assessments. The announced purpose might be to monitor achievement and guide instructional decisions on a day-to-day basis, with students and parents/guardians serving as the primary audience. As for their articulation with other assessments in the system, classroom-level assessments perhaps are seen as yielding more detailed and contextualized information about student achievement than, say, a district-level assessment or the MEA. For another example, consider a reading proficiency test that a school district administers annually at the end of second grade. Here, the formative evaluation of the reading program perhaps is the stated purpose of the assessment, where the audience is primary-grade teachers, school board members, and the public. (This audience suggests a related purpose: accountability.) In comparison to the fourth grade MEA, this test might be seen as providing a more comprehensive portrait of a student's reading proficiency, and at a more critical point in development. Also, given the announced purpose of this test—program evaluation—its "standardized" nature would be seen as an important complement to the achievement information that derives from classroom-level assessments.

Although a system's assessments differ in their announced rationale, the individual assessments do not exist in isolation. Each should be used by educators to confirm their conclusions and inferences that derive from other measures in the local assessment system.

The faithful implementation of these six features will effectively work toward creating a *system*, rather than a mere collection, of local assessments. But for a local assessment system to be useful for monitoring student achievement, the assessments that it encompasses must be of demonstrated validity and reliability. Further, the interpretation of assessment results must be informed by clear and defensible performance standards. We now turn to the first of these technical considerations: validity.

SNAPSHOT

continued

backwards to document the alignment between our assessments and the *Learning Results*. I fully expect that the process will confirm most of our work, and it will feel good to have that kind of affirmation."

"The bright side of our situation," said Theresa, "is that we're in a position to build in technical quality from the beginning. We're going to start to consider content validity by creating a blueprint that defines which assessments will address which standards."

The two educators discussed the potential uses of *Measured Measures*. "It contains definitions and descriptions that will help me understand the issues better, and I plan to read and discuss various sections with committee members as the work progresses," said Theresa.

"I'm going to be able to use several of the tools in the appendices right away. We'll look at alignment, and I'm beginning the process of calculating reliability to report on the consistency of our teacher scoring," said Alice. "People always talk about rubric scoring as being subjective, and I want to produce statistics to refute that perception."

The two exchange e-mail addresses and agree to stay in touch as they each work to move forward with their local assessment systems.

PART II

Validity

PART II
Validity

In developing and monitoring a local assessment system, educators must remain mindful of certain technical aspects of assessments. In this section we discuss the subject of *validity*, which doubtless is the most important consideration in the design and use of assessments. In Part III, we turn to a close second—reliability.

Simply stated, a local assessment system should provide evidence regarding validity. The National Center for Research on Evaluation, Standards, and Student Testing defines validity as "the extent to which an assessment measures what it is supposed to measure, and the extent to which inferences and actions on the basis of tests scores are appropriate and accurate."[1] It is helpful to separate this definition into its two parts, which we will phrase as questions that should be asked of each assessment in a local assessment system. First, what *is* the assessment supposed to measure? Second, what inferences and decisions *do* data from the assessment permit? Importantly, the inverse of each question is equally instructive: What *isn't* the assessment intended to measure? As an example, how much should reading-comprehension skills influence performance on a mathematics problem-solving task? And what would be *inappropriate* data-based inferences and actions? For instance, do states with low SAT scores therefore have inferior schools (an inference), and should teachers' salaries be tied to student performance on a state-mandated achievement test (an action)?

What is the assessment supposed to measure?

What inferences and decisions do data from the assessment permit?

In our overview of validity considerations, we will focus on the concepts of content-related validity, fairness, and consequences.

[1] On-line CRESST assessment glossary (http://www.cse.ucla.edu/CRESST/pages/glossary.htm)

Content-Related Validity

Is the content of an assessment *relevant to* and *representative of* the learning targets? For example, does a teacher's end-of-unit social studies test adequately reflect the full range, or "universe," of instructional objectives that defined the unit of instruction? By virtue of Maine legislation, local assessment systems are required to "measure progress and ensure accountability" with respect to the *Learning Results*. Thus, the performance indicators that constitute the *Learning Results* should be represented among the learning targets that guide the design of local assessments. In other words, there must be demonstrable "alignment" between local assessments and the *Learning Results*. By constructing assessments that align with the *Learning Results* (along with other, locally defined objectives), educators ensure the content validity of their assessments with respect to the targeted criteria. Consequently, educators—and the communities they serve—can be more confident that local assessments indeed measure what they are "supposed to measure."

There must be demonstrable "alignment" between local assessments and the Learning Results.

Implications for local assessment systems. Given the 1996 and 1997 Maine legislation, school districts should work towards the development of local assessment systems that are valid with respect to the *Learning Results*. This by no means suggests that a local assessment system must explicitly target every performance indicator in the *Learning Results*. This would be a Herculean task, indeed! A more practical, evolutionary approach for each school district would involve at least two steps:

❶ *Identify the content standards and performance indicators that the local assessment system will initially assess.* This list should be revisited each year. Over time, the identified learning targets should become increasingly representative of the *Learning Results*. Importantly, a local assessment system also must accommodate locally defined learning targets—targets that may be of tangential relevance to the *Learning Results*.

❷ *Identify the sources of evidence that will be used for assessing the learning targets.* There are three general sources of evidence that speak to the attainment of identified learning targets in Maine schools: (a) classroom-level assessments initiated by teachers; (b) building- and district-level assessments, such as a reading proficiency assessment routinely administered in the second grade; and (c) the MEA. An assessment system may specify a reliance on the MEA for assessing performance indicators associated with some content standards, building- or district-level assessments for others, and classroom-level assessments for others still. And by relying on two or more sources of evidence for some learning targets, a system can provide important cross-checks for the attainment of instructional goals.

By relying on two or more sources of evidence for some learning targets, a system can provide important cross-checks for the attainment of instructional goals.

Table 1

Matrix of Learning Targets and Assessment Types: English Language Arts (Grades 3-4).

English Language Arts: Grades 3-4	Assessment Type (✓)			
Content standards, with targeted performance indicators 🡻	selected response	constructed response	performance assessment	conferencing
Process of Reading				
1. Determine the meaning of unknown words by using a dictionary, glossary, or other reference sources.			✓	
2. Adjust reading speed to suit purpose and difficulty of the material.			✓	
3. Recognize when a text is primarily intended to persuade.	✓	✓		
4. Select texts for enjoyment.				✓
5. Read a variety of narrative and informational texts independently and fluently.			✓	
Literature and Culture [targeted performance indicators are listed as above, and the chosen assessment types are noted to the right]				
Language and Images [targeted performance indicators listed here; assessment types noted to the right]				
Informational Texts [targeted performance indicators listed here; assessment types noted to the right]				
Processes of Writing and Speaking [targeted performance indicators listed here; assessment types noted to the right]				
Standard English Conventions [targeted performance indicators listed here; assessment types noted to the right]				
Stylistic and Rhetorical Aspects of Writing and Speaking [targeted performance indicators listed here; assessment types noted to the right]				
Research-Related Writing and Speaking [targeted performance indicators listed here; assessment types noted to the right]				

SNAPSHOT

CONTENT VALIDITY

"We think it's important to continue to administer a standardized test as part of the local assessment system. We want to compare our schools' scores nationally." Six members of the Greentown school board sat around the conference table, talking with the district's superintendent and assistant superintendent. "I totally agree," said Steve, the superintendent. "We get a lot of important data from that kind of testing, and we use the information in several different ways."

"What needs to happen to make that testing part of our local system?" asked a committee member.

"It seems to me we need to ensure that the tests are measuring the same things that our other assessments are, including the MEA," replied Alex, the assistant superintendent. "We should try to confirm the content validity in terms of Maine's *Learning Results* and our own locally defined standards. We sometimes see significant discrepancies between MEA scores and scores on those tests. I need to figure out if they're measuring different things."

"I think that the publisher will do that for us," suggested Steve. "They

continued

Most school districts in Maine annually administer a standardized achievement test. Such tests are another source of building- or district-level evidence, provided that the test (or the portion used) is aligned with the announced learning targets. With respect to the *Learning Results*, local alignment can be determined by judging the correspondence between (a) the items on the standardized test and (b) the targeted performance indicators. Because no standardized achievement test is constructed specifically with the Maine *Learning Results* in mind, establishing adequate correspondence in this regard—i.e., content validity—is mandatory whenever scores from such tests are treated as evidence of student progress toward mastery of the *Learning Results*.

For locally constructed assessments, educators may find it helpful to prepare a matrix showing the type of assessment that will be employed for the various learning targets. Table 1 displays such a matrix for the English Language Arts content standards, grades 3-4. We list different types of assessments across the top of this matrix (our entries are by no means exhaustive), with content standards appearing down the left side. For each content standard, the school district lists the targeted performance indicators. This hypothetical district has targeted all five performance indicators for the first content standard, "process of reading." Indicator #3 will be assessed with selected- and constructed-response items, performance measures will be used for assessing the indicators #1, #2, and #5, and indicator #4 will be assessed through teacher-student conferences. (Although we have not provided details for the remaining content standards in Table 1, they would be approached in a similar fashion.)

To be sure, the selection of performance indicators will vary from one district to another. And for any district, the selection doubtless will grow broader with time. Finally, the assessment type(s) chosen for a performance indicator should be capable of accommodating the cognitive and behavioral demands of that performance indicator. For example, a performance-based assessment is more appropriate than a multiple choice test for evaluating a student's ability to read independently and fluently, just as the latter form of assessment arguably is more efficient than the former for appraising a student's ability to identify a text's general purpose.

Implications for the design of assessments: The test blueprint. Entire volumes have been devoted to the design of assessments (see Linn & Gronlund, 2000; Popham, 1999; Stiggins, 1997), a complex topic to which we cannot do justice here. Rather, we simply wish to emphasize that local assessments should align—by design—with the identified learning targets.

Traditionally, alignment has been accomplished by using a test blueprint, or table of specifications, to guide the development of assessments. Consider the test blueprint in Table 2, which is for a decidedly fictitious multiple-choice test that will follow a unit on the

Table 2

Test Blueprint for a 30-Item Multiple-Choice Test on a Civil War Unit: Number of Test Items, by Content Area and Intended Action.

	Action (that is required by the test item)			
Content Area (to be assessed) ↓	*Know*	*Compare*	*Draw Inferences*	row total (items)
Causes:	2	2	1	5
Major Battles:	6	2	2	10
Effects:	7	4	4	15
column total (items)	15	8	7	30

(Adapted from Stiggins, 1997, p. 126)

Civil War. The content areas to be assessed appear along the left side of the table (the Civil War's *causes*, *major battles*, and *effects*), and the action required by the test item appears across the top (know factual information, *compare* elements of this knowledge, *draw inferences* from these elements). The values contained in this table signify the desired number of test items. For example, the row totals show that there will be a disproportionate number of items on the effects of the war, and the column totals reveal an emphasis on the assessment of knowledge-level understanding. Where a row and column intersect, you find the desired number of test items for assessing the particular combination of content area and action. For instance, there will be two items that assess the student's ability to make inferences regarding the major battles.

The test blueprint also can be used for constructed-response assessments, such as an essay exam on a series of short stories that were read in class (see Table 3). In this case, the test blueprint specifies the number of *points* (not items, as in Table 2) that will be assigned for each content-action combination. The column totals in Table 3 show that only 20% of a student's grade will reflect knowledge-level understanding, whereas the remaining 80% will be determined (in equal measure) by the student's ability to make inferences and form evaluative judgments. The row totals reflect a somewhat even balance across the three content areas, although the content/action combinations disclose a somewhat greater emphasis on assessing one's ability to make inferences and judgments about characters.

SNAPSHOT

continued

provide a blueprint that details the various categories within each discipline and the number of questions within each of those categories. We can just compare that documentation to the *Learning Results*."

"I don't think we should rely on that," asserted Alex. "The matches that you see may be superficial. It's likely that the topics we see on each list will be similar, but I want to review at least a sample of the questions in more depth. Content validity is defined by a correspondence between the content and cognitive level of the assessments and that of the standards. There are procedures outlined in the document *Measured Measures* that we can use. I think it would be great to get some teachers involved in the review process."

"An analysis of some of the test items is an excellent idea," replied the superintendent. "We'll have to do this review ourselves, though. Involving teachers' would compromise the security of the test."

The school committee chair asked if any members had further questions, and closed the topic by requesting a progress report from Steve and Alex at the first school committee meeting the following month.

Table 3

Test Blueprint for a 100-Point Essay Test on Short Stories: Number of Points Possible, by Content Area and Intended Action.

	Action (to be scored for)			
Content Area (to be assessed) ↓	*Know*	*Infer*	*Evaluation*	row total (items)
Setting:	10	10	10	30
Plot:	10	10	10	30
Characters:	0	20	20	40
column total (items)	20	40	40	100

(Adapted from Stiggins, 1997, p. 163)

A test blueprint contributes markedly to the content validity of an assessment.

A test blueprint contributes markedly to the content validity of an assessment, provided that at least two conditions are met:

❶ *The distribution of items/points is consistent with the announced learning targets.* If, in fact, the local objectives for the Civil War unit reflect five content areas (not merely the three in Table 2), then a test derived from Table 2 would be of questionable validity for making inferences about student attainment of these objectives. That is, the assessment would bite off less than the learning targets demand.

❷ *The assessment elicits the desired behavior.* As an example, if a learning target specifies the application of content knowledge, then the corresponding assessment should engage students in tasks that require them to *apply* what they know (rather than merely regurgitate factual information).

The challenge for educators is to design assessments-items, questions, writing prompts, performance measures, and so forth—that are of demonstrable "fidelity," or faithfulness, to the targeted outcomes.

Thus, the challenge for educators is to design assessments-items, questions, writing prompts, performance measures, and so forth—that are of demonstrable "fidelity," or faithfulness, to the targeted outcomes. Fidelity refers both to the conditions of an assessment (e.g., the wording of an item, the instructions preceding a performance assessment) and to how the students' responses are evaluated. *Above all else, an assessment should produce results that permit the desired inferences about what students know and are able to do.* It is the responsibility of educators to establish a sound justification—the "warrant," if you will—for their assessment-based inferences, and this is done by pointing to the thoughtful manner in which their assessments have been designed vis-à-vis the learning targets.

Implications for assessing the Learning Results. How does the notion of test blueprint apply to assessing student mastery of the *Learning Results*? In a sense, each performance indicator represents simultaneously the two dimensions of the test blueprint, which simplifies our task considerably. Take the Social Studies performance indicator for grades 3-4, “Describe the basic structure of local and state governments.” This performance indicator specifies both content (local/state government) and action (describe). Similarly, the secondary Mathematics performance indicator, “Create and interpret probability distributions,” specifies the content of probability distributions and the action of creating and interpreting.

In designing local assessments of the *Learning Results*, educators should take the following two steps to ensure that an assessment permits valid inferences with respect to the targeted performance indicators:

❶ *Identify the performance indicator(s) that the assessment is intended to address.* Table 4 displays a checklist that might be prepared by a fourth grade teacher who is designing a math assessment that would follow a unit on the meaning and determination of “area” (of a tabletop, playground, etc.). The 11 Mathematics content standards for grades 3-4 have 26 performance indicators. This teacher has identified 8 performance indicators, across 6 content standards, as being relevant to the unit on area and, therefore, the assessment that will follow. Such a checklist, by the way, can be of equal utility for building- and district-level assessments, and for mapping the agreement between performance indicators and the content of a standardized achievement test.

❷ *Ensure that the conditions and scoring of the assessment are consistent with the language of the targeted performance indicator(s).* Let’s return to the Mathematics performance indicator, “Create and interpret probability distributions,” high school seniors could be asked to (a) identify and display all possible outcomes of tossing a coin four times and (b) determine the corresponding probability of various events (e.g., of obtaining three tails). In both scenarios, the assessment data arguably would contribute to valid inferences about the respective performance indicator.

Whether a multiple-choice item, an essay question, a performance task, or any other device, the assessment must be designed to yield data that permit inferences about the corresponding learning target(s).

We should emphasize that this second step is essential to the design of valid assessments, irrespective of assessment type. Whether a multiple-choice item, an essay question, a performance task, or any other device, the assessment must be designed to yield data that permit inferences about the corresponding learning target(s). And the notion of alignment is critical-to steps one and two alike. In Appendix A, we consider in

Table 4

Assessment/Performance Indicator Checklist for a Unit on "Area": Mathematics (Grades 3-4).

Content standards, with performance indicators ↓	*Is the performance indicator assessed?* (✓)
Numbers and Number Sense	
1. Read, compare, order, classify, and explain whole numbers up to one million.	
2. Read, compare, order, classify, and explain simple fractions through tenths.	
3. Demonstrate knowledge of the meaning of decimals and integers and an understanding of how they may be used.	✓
Computation	
1. Solve multi-step, real-life problems using the four operations with whole numbers.	✓
2. Solve real-life problems involving addition and subtraction of simple fractions.	
3. Demonstrate and explain the problem-solving process using appropriate tools and technology and defend the reasonableness of results.	
4. Develop proficiency with the facts and algorithms of the four operations on whole numbers using mental math and a variety of materials, strategies, and technologies.	✓
Data Analysis and Statistics	
1. Make generalizations and draw conclusions using various types of graphs, charts, and tables.	
2. Read and interpret displays of data.	
Probability	
1. Explain the concept of chance in predicting outcomes.	
2. Estimate probability from a sample of observed outcomes and simulations.	
Geometry	
1. Describe, model, and classify shapes and figures using applicable properties.	
2. Experiment with shapes and figures to make generalizations regarding congruency, symmetry, and similarity.	
3. Use transformations such as slides, flips, and rotations.	
4. Use the properties of shapes and figures to describe the physical world.	✓
Measurement	
1. Solve and justify solutions to real-life problems involving the measurement of time, length, area, perimeter, weight, temperature, mass, capacity, and volume.	✓
2. Select measuring tools and units of measurement that are appropriate for what is being measured.	✓
Patterns, Relations, Functions	
1. Use the patterns of numbers, geometry, and a variety of graphs to solve a problem.	✓
2. Use variables and open sentences to express relationships.	
Algebraic Concepts	
1. Develop and evaluate simple formulas in problem-solving contexts.	
2. Find replacements for variables that make simple number sentences true.	
Discrete Mathematics	
1. Create and use organized lists, tree diagrams, Venn diagrams, and networks.	
2. Give examples of infinite and finite solutions.	
Mathematical Reasoning	
1. Demonstrate an understanding that support for a claim should be based on evidence of various types (e.g., from logical processes, from measurement, or from observation and experimentation).	✓
Mathematical Communication	
1. Translate relationships into algebraic notation.	
2. Use statistics, tables, and graphs to communicate ideas and information in convincing presentations and analyze presentation of others for bias or deceptive presentation.	

more detail the various dimensions of alignment (Webb, 1997), which we believe provide a useful framework for local districts seeking to establish alignment as an indicator of content validity.

A final word on content-related validity. Although no single assessment, in and of itself, can be expected to be representative of the many content standards and performance indicators that compose the *Learning Results*, educators should strive for an assessment *system* that is. Clearly, the development of a local assessment system that adequately represents the *Learning Results* can only occur gradually, and thoughtfully, over a period of several years.

The development of a local assessment system that adequately represents the Learning Results can only occur gradually, and thoughtfully, over a period of several years.

Fairness

We now turn from content-related validity to the subject of fairness—specifically, the fairness of the inferences we make about students, based on the assessments in place and the data they provide.

Opportunity to learn. A fundamental aspect of fairness is "opportunity to learn." In this respect, an assessment is fair if students have been adequately exposed to, and engaged in, the subject matter being assessed. Put another way, there should be alignment between assessment and instruction. Even though an assessment (or assessment system) may be high in content validity vis-à-vis the learning targets, the assessment nonetheless would be unfair—at least to students—if instruction is poorly aligned with the learning targets. To be sure, the assessment data in this situation would be an *accurate* characterization of student knowledge vis-à-vis the learning targets. Further, these data doubtless would prove *useful* in identifying and rectifying the problem of poor instructional alignment. But the assessment would be unfair to students nevertheless. In short, students should not be expected to demonstrate proficiency when there are insufficient opportunities to learn.

A local assessment system should be able to demonstrate the fairness of its assessments.

This aspect of fairness is particularly relevant to the *Learning Results*, as Maine schools wrestle with the daunting challenge of implementation. Schools will vary considerably with respect to when, and to what degree, they achieve instructional alignment with the myriad objectives that make up the *Learning Results*. Assessment-based inferences must be made in full view of the degree of instructional alignment in this regard, whether the assessment is an informal observation, a unit exam, a district proficiency test, or the MEA.

Gender, ethnicity, socioeconomic status, and disability. The measures in a local assessment system should be fair with respect to gender, ethnicity, socioeconomic status (SES), and disability. In this context, fairness has important implications for the wording of assessments. An obvious injunction is that language in local assessments should be free of stereotypes regarding gender, ethnicity, SES, disability, and so on. Violations of this kind are easily caught. But fairness infractions can be subtle. For instance, fairness is compromised when problems are embedded in contexts that are more familiar to members of a particular group. Insofar as baseball is generally more popular among boys than girls, a mathematics word problem that is couched in batting averages could be questioned on grounds of gender fairness. Or consider a literacy assessment that takes passages from existing books. Children who have prior exposure to, say, the poetry of Robert Louis Stevenson have an unfair advantage in answering comprehension questions based on one of his poems. Because prior exposure reflects the literacy environment of the home (SES?), such an assessment could be questioned on grounds of fairness.

Fairness with respect to gender, ethnicity, SES, and disability also has important implications for how assessments are *administered* (e.g., extra time allowed for students with disabilities) and *scored* (e.g., performance ratings are free of bias).

A local assessment system, then, should be able to demonstrate the fairness of its assessments. This can be accomplished by a careful review of the alignment between instruction and the learning targets, as well as a careful review of the language, administration, and scoring of assessments. Toward this end, a district may wish to appoint a committee of stakeholders who would have oversight of this important responsibility.

Consequences

A recent addition to the traditional conceptualization of validity is that of "consequences" (e.g., Messick, 1995). As Robert Linn has argued,

> . . . it is not sufficient to provide evidence that the assessments are measuring the intended constructs. Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and, at the same time, not producing unintended negative outcomes. (Linn, 1994, p. 8)

In short, a local assessment system should include the collection of data that speak to both the *intended* and *unintended* consequences of the assessments it comprises.

An obvious intended consequence of a local assessment system is that, over time, there should be evidence of increased student mastery of the *Learning Results*. This consequence derives from other intended consequences of a local assessment system: greater alignment between instruction and the *Learning Results* and, in turn, between instruction and assessments.

SNAPSHOT

OPPORTUNITY TO LEARN

"It's not fair! My class hasn't covered all of the material on the district test, and I have to give it to them next week," Brad, a high school math teacher, complained. The assessment committee chair explained that the algebra on the test mirrored the expectations of Maine's *Learning Results,* and that every ninth grader would be taking the test at the same time, to be equitable. "That isn't equity," said Renee, a special educator and member of the committee. "Making all students take the same test at the same time might be quite inequitable."

"We designed an Algebra I course that spreads the material over two years to address the needs of students who need a different pace," said Brad. "That's equity. If we want all our students to take algebra and we offer them different ways to take the course, why can't we offer different ways to take the test?"

"I think that the test can stay the same, " said Renee. "But students should be able to take it when they finish their algebra course - whether that's in ninth grade or tenth."

"What about the kids who take Algebra I in eighth grade?" asked

continued

An example of an unintended, and decidedly negative, consequence of a local assessment system would be the excessive narrowing of instruction to mirror the demands of a particular assessment. This can happen with traditional and alternative assessments alike, as Linn and his colleagues remind us:

> It cannot be assumed that a more "authentic" assessment will result in classroom activities that are more conducive to learning. We should not be satisfied, for example, if the introduction of a direct writing assessment led to great amounts of time being devoted to the preparation of brief compositions following a formula that works well in producing highly rated essays in a 20-minute time limit. (Linn, Baker, & Dunbar, 1991, p. 17)

The parallels with the MEA are obvious: Classroom practice should not be narrowed or constrained by the format and demands of the MEA. As for local assessment systems, the legislated assessment of the *Learning Results* should not bring about the demise of locally defined learning targets, nor should the additional assessment activity erode instructional time.

Surveys of teachers and students can throw light on the intended and unintended consequences of assessments, as can observations of classroom practice. Assessments also may have consequences beyond the walls of the classroom, such as impacts on parents, future employers, and the community at large (Lane, Parke, & Stone, 1998). The use of focus groups can be helpful for appraising the intended and unintended consequences of assessments with respect to these constituencies (Chudowsky & Behuniak, 1998).

We now turn to the second of the three technical considerations: reliability.

SNAPSHOT

continued

Michelle, a middle school teacher. "Why not have the test available for them when they finish the course?"

The group agreed that the idea made a lot of sense. Equivalent forms of the district's algebra test would be made available to all teachers with an Algebra I course. Students would take the test at the end of their course, no matter what year, and the results would be used to certify their achievement of the algebra standards and to inform instructional decisions in their next math course.

A local assessment system should include the collection of data that speak to both the intended and unintended consequences of the assessments it comprises.

PART III

Reliability

PART III
Reliability

A local assessment system should provide evidence regarding the reliability of its assessments. Simply stated, reliability refers to the *consistency* of the scores, ratings, or judgments that derive from an assessment. For example, you would question the reliability of your weight scale if you obtained divergent results upon weighing yourself twice, identically clothed, within a 10-second interval. Similarly, you would question the reliability of proficiency judgments if two teachers, having examined the same exhibits, rendered identical judgments for only a handful of students.

Reliability refers to the consistency of the scores, ratings, or judgments that derive from an assessment.

As you will see below, reliability sometimes is reported in the form of a *coefficient*, which can range from 0 (no reliability whatsoever) to 1.00 (perfect reliability). Reliability coefficients for teacher-constructed tests typically fall between .60 and .85 (Linn & Gronlund, 2000), whereas those for standardized tests of achievement and aptitude tend to concentrate in the .80s and low .90s (e.g., see Salvia & Ysseldyke, 1998). Reliability also may take the form of a percentage, as when one reports the *percentage* of work samples for which two raters have reached similar judgments.

Whether expressed as a coefficient or a percentage, reliability can be conceptualized in several ways: consistency across *raters*, consistency across *equivalent forms*, consistency across *time*, and consistency across *items or tasks*. Not surprisingly, the purpose of an assessment, as well as practical considerations, dictate which conceptualization(s) is deemed appropriate.

An assessment high in reliability nonetheless can be low in validity for its announced purpose.

We wish to emphasize two points before elaborating upon these different conceptualizations of reliability. First, an assessment high in reliability nonetheless can be low in validity for its announced purpose. This would be the case if, say, a first-grade teacher used "student height" (a highly reliable measurement) for making judgments about reading ability (a patently invalid inference), or if two raters were in perfect agreement but were misapplying the scoring rubric in precisely the same fashion. In short, "reliability" does not entail "validity." Second, we do not expect, nor do we recommend, that teachers establish the reliability of their assessments as a matter of course. Indeed, this would be an unrealistic and unnecessary demand to impose on teachers. Nonetheless, teachers *should* understand the concept of reliability, the different ways it surfaces, and the factors that influence it. Such knowledge, even in the absence of formal calculations, invariably leads to more thoughtful assessment. And as thoughtfulness increases, so does the reliability of the product. We will return to this matter below.

Consistency Across Raters

For assessments that call for global ratings, such as extended responses, productions, and exhibits, reliability can be established by determining the amount of agreement among raters. This form of reliability is known as "inter-rater reliability."

As an example, suppose two raters independently evaluated 50 on-demand writing samples using a six-point scoring rubric. One reliability index, the "percentage of exact agreement," reports the percentage of students who receive the *same* score from both raters. A less stringent criterion is the percentage of agreement *within one point* (e.g., a student receives a 5 from one rater and a 4 from the other). Appendix B provides a worked example of both, as well as an application involving dichotomous judgments (e.g., "proficient" versus "not proficient").

High agreement is good news, of course: Raters are making similar judgments about the quality of student work. Where agreement is low, however, further study is required to identify and reconcile the source(s) of nonagreement. Generally, low agreement between raters indicates a problem with the *criteria* (ambiguous, unclear), the raters' *application* of the criteria (incorrect, nonsystematic), or both.

Although we have used the percentage-agreement method to illustrate inter-rater reliability, you sometimes will encounter use of the Pearson correlation coefficient, or Pearson *r*, for this same purpose. For the example above, one could correlate the 50 six-point ratings of the first rater with those of the second rater. The resulting correlation is the reliability coefficient and, as we indicated earlier, will fall between 0 and 1.00. Calculating Pearson *r* by hand is tedious, particularly when based on large numbers of students (as would be the case here!). Fortunately, personal computers simplify this chore considerably. For the masochists among us, we provide a worked example of Pearson *r* in Appendix C.

SNAPSHOT

INTER-RATER RELIABILITY

"Are we supposed to add our scores together?" asked Elsa, a local game warden. Elsa sat with Stan, a biology professor from the university, and Vanessa, a high school teacher. The three had been invited to serve as panelists for a series of fourth grade science exhibitions addressing various ecology topics. They sat at a table with a stack of papers in front of them, listening to Ike, a fourth grade teacher, explain their role as panelists. "We're going to run through some training exercises to help you become familiar with the process and the scoring guides, " he explained. "You will each assign scores independently, and we'll ask you to confer and reach agreement on all the scores that you give."

"Why not have a conversation and assign the score as a group?" asked Stan.

"We need to be able to demonstrate that the scoring process isn't subjective," answered Paulette, the elementary school principal. "If we can show that three different panelists designate the same score, without consultation, that demonstrates the reliability of the scoring."

"What if we're within one score point of the other panelists?" asked the high school teacher. "Is that good enough?"

continued

Consistency Across Equivalent Forms

"Equivalent-forms reliability" is determined by administering two equivalent, or parallel, forms of an assessment to the same students and then establishing the similarity between the two sets of scores. The time interval between the two administrations should be kept to a minimum. Like inter-rater reliability, this form of reliability can surface either as a percentage or as a coefficient.

Let's consider the percentage first. Suppose a school district requires that students pass a computer-proficiency test before graduating from high school. Because students are allowed to retake this test if they initially fail, the district decides to develop two parallel assessments. Clearly, a student judged to be proficient (or not) on the basis of one form of the assessment should be similarly judged on the basis of the other. To examine reliability in this regard, the superintendent has a sample of 40 students take both forms of the proficiency test, and then calculates the percentage of students who receive the same proficiency judgment on the two occasions (see Appendix D). A high percentage indicates that the proficiency test yields similar judgments regarding proficiency, irrespective of which form of the test is taken. That is, the test is reliable.

The example above can be modified slightly to illustrate the use of Pearson *r* for establishing equivalent-forms reliability. (Again, it is best to use a personal computer for this task.) Imagine that this proficiency test was worth a total of 80 points. Each student in the superintendent's sample thus has two scores: one from each form. The correlation between the two sets of scores is an expression of equivalent-forms reliability: the higher the correlation, the greater the similarity in a student's relative performance on the two forms, and hence the greater the reliability of the proficiency test. For standardized tests of achievement and aptitude, equivalent-forms reliability invariably is expressed as a correlation coefficient.

Consistency Across Time

"Test-retest" reliability, as it is called, typically is established by administering a single assessment twice to the same individuals and then computing the correlation between the two sets of scores. There is an intentional interval of time between test and retest, an interval that can range from minutes to months. In this sense, test-retest reliability reflects the "stability" (over time) of performance. Standardized tests of achievement and aptitude frequently report test-retest coefficients. (The worked example in Appendix C easily can be modified to illustrate test-retest reliability: Simply consider *X* and *Y* to be the test and retest, respectively.)

SNAPSHOT

continued

"Being one point apart on a four point scale represents 25% error," answered Paulette. "That's especially hard to accept when one panelist gives a score that suggests the student is proficient and another score indicates otherwise. We need to strive for exact agreement."

Ike cautioned the three scorers about various kinds of bias that might interfere with reliable scoring. "Try to ignore neatness or prettiness unless it's specifically included in the scoring guide," he said. "It doesn't matter how many different colors a student uses in a graph. What matters is if the graph is technically correct and communicates effectively."

"These students have worked very hard in preparing their exhibitions," said Paulette. "But you are not to score them on their effort. You must look and listen carefully and match the evidence in their exhibitions to the descriptors on the scoring guides."

Ike and Paulette walked the three panelist through several video-taped sample exhibitions, asking them to score each one and then discuss the scores together. This process illuminated the details of the scoring guides, revealed important scoring decisions, and allowed the panelists to move into the live exhibitions with good strategies for scoring consistently.

A note on interpreting a correlation coefficient: Perfect test-retest reliability means that the two sets of student *rankings* are identical (e.g., Student A had the highest score on both occasions)—not that the two sets of scores are identical (e.g., Student A received a 98 on both occasions). This is true of inter-rater and equivalent-forms reliability coefficients, as well.

Consistency Across Items or Tasks

This form of reliability reflects the consistency of a student's performance across the items or tasks that make up an assessment. It is called "internal-consistency reliability" and, unlike equivalent-forms and test-retest reliability, requires the administration of a single assessment on a single occasion. In Appendix E, we provide a worked example of "KR-21," a convenient short-cut method for estimating the internal-consistency reliability of tests made up of *dichotomous* items (i.e., items scored correct-incorrect). A statistical cousin of KR-21, "Cronbach's alpha," can be used for assessments that are made up of—are you sitting down?—*polytomous* items or tasks. This is a fancy way of saying that the item or task is scored using multiple values, such as a short-answer question worth four points (rather than scored "right" or "wrong"). Although this coefficient requires rather unwieldy calculations, personal computers render the computation of Cronbach's alpha (almost) effortless.

KR-21 and Cronbach's alpha are appropriate for assessments having relatively "homogeneous" content—that is, all items or tasks tap a similar skill, competency, or ability.

KR-21 and Cronbach's alpha are appropriate for assessments having relatively "homogeneous" content—that is, all items or tasks tap a similar skill, competency, or ability. What does one do when the content of an assessment is more heterogeneous? One option is to calculate either coefficient for homogeneous *subsets* of items (e.g., items dealing only with mathematical calculations, items dealing only with vocabulary, items dealing only with problem solving). Another option is to employ the "split-half" method, an alternate measure of internal-consistency reliability that is not affected by heterogeneous content. This method requires dividing the test in half (e.g., odd items vs. even items), scoring the two halves separately for each student, and correlating the two sets of scores. This correlation, after a minor adjustment, is the "split-half reliability" of the complete test (see Linn & Gronlund, 2000).

We encourage building- and district-level administrators to develop and conduct their assessments with reliability in mind and, in turn, to calculate reliability indices on a routine basis.

Implications for Building- and District-Level Assessments

We encourage building- and district-level administrators to develop and conduct their assessments with reliability in mind and, in turn, to calculate reliability indices on a routine basis. For example, equivalent-forms reliability should be examined periodically where assessments have alternate forms; inter-rater agreement should be monitored where assessments rely on global ratings; and internal-consistency reliability should be inspected where assessments contain selected-response items.

Further, insofar as the subject of reliability doubtless exceeds the comfort level for many teachers, building- and district-level administrators should provide staff development to provide the necessary understanding of reliability and its applications.

Implications for Classroom-Level Assessments

Reliability indices commonly are reported for standardized tests, and you may also encounter them in association with building-, district-, and state-level assessments. However, we must concede that the transfer of these methods to classroom practice does not come easily. For instance, it is unreasonable to expect teachers to routinely give a test twice, or develop parallel forms of a test, in order to estimate reliability. Teachers nonetheless can take important steps to enhance the reliability of their assessments—whether or not any calculations are ever made. In short, reliability is generally higher when:

- the assessment has an ample number of items or tasks;
- the assessment has clear language—clear essay questions, clear writing prompts, clear multiple-choice items, clear instructions for performance assessments, and so forth;
- the criteria for evaluating student performance are stated clearly; and
- the criteria are applied consistently and with fidelity.

Teachers can take important steps to enhance the reliability of their assessments—whether or not any calculations are ever made.

Nevertheless, we encourage teachers to take the next step by actually examining the reliability of their assessments. As an example, a teacher might occasionally enlist the services of a colleague to independently rate a sample of student work in order to calculate inter-rater reliability. Or, if selected-response tests are given, the teacher could periodically calculate internal-consistency reliability (e.g., KR-21).

A teacher might occasionally enlist the services of a colleague to independently rate a sample of student work in order to calculate inter-rater reliability.

How High Should Reliability Be?

Unfortunately, there is no straightforward answer to this question—other than to say “the higher, the better.” The importance of reliability clearly depends on the nature of the decision to be made. For example, Salvia and Ysseldyke (1998, p. 163) specify a minimum reliability of .90 for assessments that are used for tracking and placement decisions, and .80 for screening decisions (e.g., recommending a student for further testing). Although providing no numbers, Linn and Gronlund (2000) argue that high reliability is mandatory when assessment-based decisions are important, final, irreversible, unconfirmable by other data, concern individuals, and have lasting consequences. In contrast, lower reliability can be tolerated when decision-making is in the early stages and the decision is of minor importance, reversible, confirmable by other evidence, concerns groups, and has temporary effects. The reliability of teacher-constructed assessments (.60-.85) is generally adequate for the

The importance of reliability clearly depends on the nature of the decision to be made.

day-to-day decisions routinely made by teachers (e.g., whether to provide additional review of a topic, move on to another unit, give an individual student further assistance).

We now turn to our final technical consideration: setting performance standards.

PART IV

Setting Performance Standards

PART IV
Setting Performance Standards

In this final section , we provide an overview of setting performance standards for local assessment systems. In addition to establishing the validity and reliability of assessments within the system, educators must identify the amount and quality of evidence necessary to demonstrate proficiency on assessments. In other words, they must establish performance standards. First, we provide definitions of key terms in the standard setting process. Second, we describe strategies for defining performance standards for the tools and instruments within the local assessment system. Finally, we discuss how to establish performance standards for local systems, combining information from local and external assessments to make decisions about students' achievement of the *Learning Results*.

Educators must identify the amount and quality of evidence necessary to demonstrate proficiency on assessments.

What Are Standards?

Standards are statements of expectations for student learning. *Content standards* are "broad descriptions of the knowledge and skills that students should acquire" (Maine Department of Education, 1997, p. iii). Content standards answer the question, "What do students need to know and be able to do?" Each content standard specifies a number of *performance indicators* that "define in more specific terms the stages of achievement, or checkpoints toward meeting the content standard within each of four grade spans" (Maine Department of Education, 1997, p. iii). As an example, consider the following content standard taken from the Mathematics portion of the *Learning Results*: "Students will understand and apply concepts of data analysis." A corresponding performance indicator (grades 5-8) is "Organize and analyze data using mean, median, mode, and range." Table 5 presents, for your information, a sample task that is aligned with this content standard and performance indicator.

In contrast to content standards, performance standards are "explicit definitions of what students must do to demonstrate proficiency at a

Table 5
“Who is the Best?”

Rick, Mike, and Sarah are all on their school’s golf team. They have been practicing their chipping. Each player thinks she/he is the best chipper on the team. To decide who is right, they have a contest. Each player chips 10 balls onto the same green. The balls are different colors so they can tell them apart. When they finish, they measure the distance from each ball to the cup in inches. Here are the results, in no particular order:

Rick:	40, 60, 100, 120, 312, 320, 152, 105, 95, 46
Mike:	52, 76, 184, 288, 230, 120, 64, 60, 88, 188
Sarah:	84, 99, 130, 135, 200, 165, 120, 129, 136, 152

When the contest was over, the kids still couldn’t decide who was the winner. The balls were all spread out. No one was close every time. They asked the coach for advice. He said, “In the game of golf, getting close and being consistent is important. So, you should consider who is closest and most consistent. Don’t just consider who had the best shot. You’re the math whizzes—I’m sure you can figure it out.”

Help the kids decide who won. Analyze the results in as many different ways as you know. Present a mathematical argument to back up your decision about who the winner was and why she/he won.

***Remember that in golf, the closer the ball is to the cup, the better the shot.

(Source: 1998-1999 Maine Assessment Portfolio Pilot Middle Level Math Anchor #1)

specific level on the content standards.”[2] Performance standards answer the familiar question, “How good is good enough?” Performance standards might apply to the amount of emissions from a factory, the job performance necessary to get a raise or promotion, the requirements for receiving a driver’s license, and so on. For example, students might demonstrate their proficiency on the Mathematics performance indicator above by (a) organizing the set of scores in Table 5, (b) correctly calculating two of the three measures of central tendency, and (c) accurately considering the range of scores. In summary, once the task is defined, the performance standard is set based on judgment regarding how well a student must perform in order to indicate proficiency.

Performance standards answer the familiar question, “How good is good enough?”

As you can see, then, a performance standard is a level against which something is measured. In the case of a driver’s license, the equivalent of content standards would establish age, driving ability, and knowledge of rules of the road as important criteria in judging an applicant’s eligibility for receiving a driver’s license. In this case, the established performance standards include a specific age (usually 16), a certain score on a written exam (usually 60%), and a certain number of points on a driving test (usually 60%).

[2] On-line CRESST assessment glossary (http://www.cse.ucla.edu/CRESST/pages/glossary.htm).

Performance standards in a criterion-referenced system. The *Learning Results* spell out content standards, and students will be expected to demonstrate the skills and knowledge that these content standards delineate. This is a decidedly criterion-referenced enterprise. That is, educators will compare a student's performance to pre-established performance standards, rather than engage in norm-referenced judgments of student performance where students are compared to each other. In other words, the standards are absolute, not relative. Again, the driver's license: "Whether any single applicant passes or fails has nothing whatsoever to do with how the other applicants do on the test because applicants' scores are not compared to one another. Instead, each applicant is compared to the predefined standard . . . to determine whether he or she passes. . . . In this system, it is possible for all or none of the applicants to pass" (Airasian, 1994, p. 295).

Standards are absolute, not relative.

Within the context of a local assessment system, every student can achieve at the highest level, just as every student can lack proficiency. It all depends on where student performance falls with respect to the standard.

Standard setting requires judgment. Standard setting is a judgmental process. That is, judgment is used to define the acceptable level of student performance. But questions arise: "Who should make the judgments? How should these judgments be elicited? Should judgments be based on information about tests, test items, student work, student ability, or a combination of these factors?" (Jaeger, 1993, p. 492).

At all levels—one teacher in a classroom, several department members in a school, or a representative group of stakeholders in a state—judgments establish performance standards.

Setting performance standards requires judgment whether for an individual teacher's classroom assessment, a department's final exam, or a state's testing program. As the purpose for the assessment changes, and as the number of students involved and the magnitude of the decisions that might be made increases, the judgments must be made in more formal, public ways.

Teachers generally set their own grading standards. That is, each teacher ultimately makes his or her own judgment about how good is good enough. In contrast, when the faculty of a high school department meet to set standards for a final exam, they must seek consensus in making a collaborative judgment. Because the consequences of passing or failing such an exam are significant, the faculty must be prepared to explain and justify their performance standards. When the Maine Department of Education sets performance standards for the MEA, the stakes are quite high. School achievement data are made public, and the judgments that establish performance levels must be defensible. This is accomplished by involving many stakeholders in the process and by employing an accepted, credible procedure. In summary, at all levels—one teacher in a classroom, several department members in a school, or a representative group of stakeholders in a state—*judgments* establish performance standards.

The purpose of the assessment informs the standard setting procedure. The formality and inclusiveness of a standard setting process differ from one level to the next because of the variety of purposes for which assessments may be used. In the classroom, the purpose may be to improve instruction. But across a grade level, department, or system, the assessment may be used for certification—that is, to document student achievement of state and local standards. For example, to determine whether students have an understanding of a particular concept taught in class, a teacher might observe students during classroom discussions, or she might devise a quiz or performance task. A teacher is likely to do this independently. In the case of a final exam in a high school course that serves as a prerequisite for more advanced studies, all of the teachers teaching that course would need to be involved in developing the exam and establishing the performance standard for a passing grade. Finally, when a school system seeks to certify student achievement of the *Learning Results* as a qualification for graduation, many stakeholders would collectively look at the performance of students across many assessments, in all content areas.

Standard setting must be aggregated up; that is, in the classroom a teacher may make an independent decision but it should be consistent with standards set at the grade level.

Where assessment information is intended to serve as part of a comprehensive system, classroom performance standards must be consistent with grade level standards and with school-wide standards. Standard setting must be aggregated up; that is, in the classroom a teacher may make an independent decision but it should be consistent with standards set at the grade level. Teachers at each grade level should talk with each other and come to consensus about what the standard should be across the school. Schools come to consensus across the district, and districts across the state.

Two General Methods for Setting Performance Standards

There are two general approaches to setting performance standards. One is based on a particular test or assessment, which we will call the assessment-based method for standard setting. The other is based on the examinee or on student work, which we will call the examinee-based method for standard setting. In each case, judgments are made about what constitutes proficiency.

Assessment-based methods for setting standards. A common, informal assessment-based method occurs when teachers get together and look at a local test to determine (a) the test's degree of alignment with the *Learning Results* and (b) the score that indicates proficiency or meeting the standard. For example, teachers may declare that "Students will need to get 80% correct on this test to meet the standard." The standard therefore is 80%. In the case of reading, teachers assign levels to literature or reading passages according to established text criteria and make a judgment about what level meets the standard for acceptable reading performance. Whenever teachers review a test, task, or question and determine a passing grade or acceptable performance, they are employing an assessment-based method to establish performance standards.

A higher-stakes assessment requires a more formal standard setting procedure that documents both the judgments and the decision rules used to reach them. The performance standards must be tied explicitly to the assessments. A common assessment-based protocol for establishing performance standards for large scale assessments is the "Angoff method." Here, individuals who are involved in the standard setting process individually examine each item on the assessment, and they estimate the probability that students with knowledge and skills within each performance level would answer the item correctly. For instance, how likely is it that a fourth grader deemed "proficient" in mathematics would select the correct choice on the following question from the MEA?

Whenever teachers review a test, task, or question and determine a passing grade or acceptable performance, they are employing an assessment-based method to establish performance standards.

Susan's age is a multiple of 3. Which number could represent her age?

A. 5 B. 10 C. 15 D. 16

Judges likely would specify high probabilities for such a simple question (e.g., .90 or greater), just as they would assign lower probabilities to more challenging items. The average of the probabilities—across all judges and all items—becomes the score necessary to "meet the standard" on that test. If the average probability across all judges and all items were, say, .75, then a score of 75% or higher corresponds to proficiency, or, meeting the standard.

As an example, imagine that three judges review a social studies test consisting of five multiple-choice items. Item by item, each judge estimates the percentage of students categorized as "meeting the standard" who would select the correct response on the item. Judge 1 estimates 75%, 95%, 60%, 85%, and 90% for the respective items; Judge 2 estimates 70%, 90%, 50%, 80%, and 90%; and Judge 3 estimates 70%, 95%, 70%, 90%, and 100%. These ratings total 1210, which is divided by 15 (i.e., the number of judges multiplied by the number of items). Thus, 81% becomes the performance standard for proficiency, or meeting the standard, on this test.

A simpler version of the Angoff method allows judges to choose from a list of probabilities. In the case of the item above, each judge might be asked to select a probability from the following choices:

.10 .20 .30 .40 .50 .60 .70 .80 .90 1.00

Table 6

Local Uses of Angoff-Related Methods.

To implement the Angoff (or modified Angoff) method
in a local assessment system:

a. Identify an appropriate population of judges (think about all of the groups or stakeholders who should be represented).

b. Select a representative sample to actually carry out the standard setting.

c. Have judges review the selected content standards and performance indicators that the test or assessment is designed to measure, and performance level definitions (as have been written at the state level for the MEA performance levels).

At this point, the procedure will differ depending on whether it is being applied to selected- or constructed-response items:

IF SELECTED RESPONSE (e.g., multiple-choice test) ↓	IF CONSTRUCTED RESPONSE (e.g., 4-point scoring rubric) ↓
Each judge considers individual items and estimates the percentage of students, within each performance level, who would get the item correct.	Each judge considers individual items and estimates the percentage of students, within each performance level, who would score 4.
All estimates for a performance level are averaged to establish the performance standard for that level of performance.	All estimates for a performance level are averaged, expressed as a decimal, and multiplied by 4 to establish the performance standard for that level of performance.
The average of the percentage correct for “meeting the standard” becomes the performance standard for “meeting the standard,” and so on.	The results represent performance levels in terms of mean (average) rubric score points. The figure for “meeting the standard” becomes the performance standard for “meeting the standard,” and so on.

As you can imagine, this procedure facilitates the judgmental process considerably. A further modification allows judges to discuss their estimates and, in turn, alter their choices based on this discussion.

Although Angoff methods were designed to set standards on selected response tests (e.g., multiple choice items), they can be altered to accommodate constructed response items or tasks. Suppose three judges review a science test comprising three constructed response items. They each estimate the likelihood that a student categorized as “meeting the standard” will score a 4 on a 4-point rubric. They respectively assign 80%, 70%, and 90%; 85%, 75%, and 90%; and 70%, 75%, and 90% to the three items. This totals 725. Convert to .73 and multiply by 4 (on a 4 point rubric) = 2.9. Thus a mean score of 2.9 across the 3 constructed response items constitutes proficiency, or meeting the standard, on this test.

SNAPSHOT

STANDARD SETTING

“Let’s start with introductions,” said the woman standing in the front of the room. My name is Kathleen Potok. I’m the assistant principal here at the middle school, and I’ll be facilitating this standard setting session.” The 18 people seated at tables around the school library introduced themselves, revealing that the group included parents, an employee of the local moving company, a geography professor, a journalist, and teachers from both the middle and high schools.

Kathleen went on to explain that the group’s charge was to set performance standards for the district’s eighth grade social studies assessment. “We’ll be looking directly at samples of student work to help you set the standards.”

Sally, a middle school social studies teacher, stood before the group. She reviewed the *Learning Result’s* content standards and performance indicators addressed by the district assessment. She explained that those standards were the focus of the middle school’s social studies curriculum, and that the faculty believed their students were prepared for the assessment.

Kathleen reviewed the four performance level descriptors that they would be using, similar to those defined for the MEA. These included “not meeting standards,” “partially meeting standards,” “meeting standards,” and “exceeding standards.”

continued

Among other things, these brief examples suggest the value of having as many test items and judges as possible in the standard-setting process. This guards against an individual judge or a single item affecting the overall performance standard disproportionately. In Table 6, we summarize the basic steps of Angoff-related methods.

Examinee-based methods for setting standards. The second approach for setting performance standards frequently involves the review of student work. The work represents the student (examinee) and allows judgments to be based on first-hand evidence of performance. As a familiar example, elementary teachers sometimes assign ratings of ✓, ✓-, or ✓+ to student projects. They base their judgments on the quality of the projects and, in the process, define the characteristics of a ✓, or acceptable level of performance. This represents an informal examinee-based standard setting process.

One example of an examinee-based approach is the "student-based constructed-response method." Panels of judges match actual student work to definitions of performance levels. This method also is called the "body of work" or "bookmark" method. In any case, judgments are based on a review of students' complete sets of responses across the items on an assessment. A more complete description of this process, along with some suggestions for implementing it at the local level, can be found in the section below entitled "Setting Performance Standards for the MEA."

The "contrasting groups method" is another example of an examinee-based approach for setting performance standards. This method requires teachers first to consider all that they know about their students, based on classroom observation, interaction, performance, and so forth. Teachers then predict each student's level of performance on the non-classroom assessment (e.g., school, district, or state) for which performance standards are to be set. These predictions, in turn, are compared to actual scores on the assessment to establish the performance standards. This process, too, is elaborated upon in the section entitled "Setting Performance Standards for the MEA."

Before a district establishes performance standards, it may choose to engage educators in exercises to elucidate the issues and processes involved in standard setting. If a district currently is administering a district-wide assessment, for example, district leaders may wish to follow a procedure such as that in Table 7. This exercise assigns performance standards solely on examination of student work and does not establish numerical cut points to represent those standards. Nevertheless, this procedure allows a group to make judgments and assign performance levels. It may also serve as useful professional development activity, engaging educators in an experience that can inform more formal standard setting procedures.

SNAPSHOT

continued

The panelists began by looking at selected samples of student work across the various parts of the social studies exam. They looked at all of the work that an individual child did on the test: the multiple choice answers, the responses to several open-ended questions, and the essay and bibiographical information on an extended research question. After reading through each student's "body of work" on the test, the group discussed which performance level descriptor fit it best. After several such discussions, Kathleen moved them on to making independent judgments about many more sets of student responses. Each panelist assigned one of the four performance levels to each of the some 20 selected "bodies of work." The collections had been selected to represent a range of score points to allow the specific "cut point" to be identified.

Karen explained the final steps in the process to the members of the group. "We will look for the specific score point where half of you assigned the collection one performance level and half assigned it the next level. That total score, across the "body of work," becomes our cut score and determines which performance level the student has achieved."

"We are delighted to have you all here to do this important work," Kathleen said. "It's very important that these decisions be made by a group from both inside and outside of the school system, and that we use a process that we'll be able to explain and justify to others. Thanks again for being here."

Table 7

An Exercise for Assigning Performance Levels by Examining Student Work.

If a district (or school) administers and scores a series of performance tasks and/or some other combination of assessments, and it is feasible for a panel to review all of the student work, performance levels can be assigned in the following way:

a. A standard-setting panel is assembled.

b. The panel separates student work into two piles: "proficient" and "not proficient."

c. The two piles are further subdivided into two more piles each. "Proficient" student work is designated as either "meets" or "exceeds" the standard, and "not proficient" work is labeled either "partially meets" or "does not meet" the standard.

Writing rubrics to describe performance standards. Writing rubrics combines elements of assessment-based and examinee-based methods for establishing performance standards. Initially, rubric writers review an assessment task, the content standards and performance indicators it addresses, and the generic descriptor for "proficient" performance. Based on that review, they then describe the evidence necessary for a student to demonstrate proficiency on that standard (content standard and performance indicator) for that particular assessment. Thus far, the process has been assessment-based, and the initial performance standard has been set. After the assessment has been used with students, however, rubric writers revisit their descriptors to ensure that the evidence cited concurs with the evidence found in actual student work. At this point, descriptors are often modified to better reflect the characteristics of student responses at various performance levels. In this way, rubric writing represents a combination strategy for establishing performance standards.

Writing rubrics combines elements of assessment-based and examinee-based methods for establishing performance standards.

Educators intending to develop rubrics to establish performance standards must first consider the format or type of rubric and scoring guide that they wish to generate. Appendix F presents more information about rubric types, along with examples that can serve as templates. Further, technical quality and utility must be considered in the process of drafting rubrics and scoring guides. Appendix G provides suggestions that may be helpful in this regard.

Setting Performance Standards for the MEA

While *Measured Measures* is intended to inform the development of local assessment systems, we include an overview of the standard setting methods that have been applied to the MEA. We do so primarily to illustrate and elaborate upon our general description of examinee-based methods above. Reflecting on the standard setting process within the

Table 8

Performance Levels, Definitions, and Evidence Statements for the MEA.

Exceeds the Standard	*Definition:* The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results*. The student's body of work demonstrates exemplary knowledge of content and skills such as analysis, problem solving, and communication.	*Evidence:* This student's responses on the MEA demonstrate an in-depth understanding of the content knowledge and application skills. The evidence indicates that the student grasps major concepts, draws connections among ideas, and communicates complex concepts effectively (often creatively). The student's responses demonstrate an ability to solve challenging problems in a correct and exemplary manner.
Meets the Standard	*Definition:* The quality of a student's work at this level of proficiency meets the standards of performance for Maine's *Learning Results*. The student's body of work demonstrates consistent knowledge of content and skills such as analysis, problem solving, and communication.	*Evidence:* This student's responses on the MEA demonstrate a consistent understanding of content knowledge and application skills. These responses are characterized by their clarity, comprehensiveness, effectiveness, and correctness.
Partially Meets the Standard	*Definition:* The quality of a student's work at this level of proficiency partially meets the standards of performance identified for Maine's *Learning Results*. The student's body of work demonstrates partial and/or inconsistent knowledge of content and skills such as analysis, problem solving, and communication.	*Evidence:* This student's responses on the MEA demonstrate a partial understanding of content knowledge and application skills. These responses are characterized by some inconsistency in clarity, comprehensiveness, effectiveness, and correctness. Some of the student's responses are incomplete or exhibit some gaps in content knowledge or application skills.
Does Not Meet the Standard	*Definition:* The quality of a student's work at this level of proficiency does not meet the standards of performance as identified by Maine's *Learning Results*. The student's body of work demonstrates a limited knowledge of the content and skills such as analysis, problem solving, and communication.	*Evidence:* The student's responses on the MEA demonstrate a limited understanding of content knowledge and application skills. These responses are characterized by lack of clarity, comprehensiveness, effectiveness, and completeness. Many of the student's responses are incomplete and exhibit gaps in content knowledge and application skills.

(Source: Maine Department of Education)

familiar context of the MEA brings to the surface some of the issues that must be considered at the local level.

As readers doubtless know, the Maine Department of Education has developed a set of performance levels to be used in conjunction with the MEA. There are four performance levels, corresponding definitions, and evidence statements (see Table 8). Through the standard setting process, each of performance levels in Table 8 is assigned a value between 501 and 580 (the MEA scale scores). For example, a student who places in the highest performance level (exceeds the standard) must score between 561 and 580 on the MEA. Similarly, a student who scores between 501 and 519 will be placed in the lowest performance level (does not meet the standard). These scales correspond to specific ranges of raw score point values, identified through standard setting as representing the levels of performance. These performance standards have been set by the state, using various statistical and judgmental procedures.

Student-based constructed-response method. The first approach used in MEA standard setting, the aforementioned student-based constructed-response (SBCR) method, brought together a large group of stakeholders: teachers, parents, business people, and policy makers. As individuals, the groups reviewed many samples of student work across the questions on one subject area test from the MEA. For instance, a group reviewed all of the fourth grade reading questions (multiple choice, short answer, open response) and answers from a large sample of students. After initial training, group members were asked to make independent judgments about the performance level to which each "body of work" belonged. That is, panel members examined a student's responses to all questions and then categorized that body of work as exceeding, meeting, partially meeting, or not meeting the standard. Cut points, or performance standards, were set where half of the readers judged a body of work at one level of performance and half judged it at the next level. These judgments were then translated into score points (the number of raw score points that the student received) in order to establish the raw score that would serve as the cut point—which determines where a student is classified among the four performance levels. The raw scores were placed on the now familiar 501-580 scale to enable all of the grade levels and subject areas to report in a consistent format. For example, Table 9 presents the cut points for the reading portion of the 1998-1999 MEA.

Panel members examined a student's responses to all questions and then categorized that body of work as exceeding, meeting, partially meeting, or not meeting the standard.

How can this process be applied to local assessment? Districts could replicate the SBCR method by following these steps:

❶ Score all student work prior to standard setting, but provide samples to judges without scores or notations.

❷ Organize sample collections by total score points, ensuring that a broad range of score points are represented by multiple examples.

❸ Review content-specific performance standards and descriptions of evidence with the judges.

❹ Engage judges in a discussion to reach consensus on the performance level of several bodies of work (full sets of responses, across a test or assessment, from one student).

❺ Ask judges independently to assign each of their assigned collections one of the four performance levels. Judges will each review different sets of student work, but within the same range of score points.

❻ Identify the score points where half of the judges rated the work at one performance level, and half of the judges rated the work at the next level. These represent cut points for assigning performance levels to individual students on the basis of raw score points.

The contrasting groups method. The second approach used in MEA standard setting, the contrasting groups method, engaged a sample of Maine teachers for the purpose of making judgments about student performance levels. This method is based on teachers' ratings of students. After the MEA was administered, selected teachers were asked to review their class lists and designate one of the four MEA performance levels for each student, based on the student's performance in class. These judgments, in turn, were linked to MEA raw scores so that cut points could be identified. Again, cut scores were set where half of the judges (teachers) designated one performance level, and half designated the next level. If, among students scoring 33 points, half of the teachers classified their classroom performance as "meeting the standard" and half as "exceeding the standard," then the cut point would be set at 33 points.

Teachers were asked to review their class lists and designate one of the four MEA performance levels for each student, based on the student's performance in class.

Table 9
Raw-Score Cut Points for the MEA Reading Test.

		Performance Level			
Grade	Total possible points	"Does not meet the standard" (501-520)	"Partially meets the standard" (521-540)	"Meets the standard" (541-560)	"Exceeds the standard" (561-580)
4 8 11	53 52 53	0-21 0-21 0-23	22-33 22-33 24-37	34-46 34-44 38-47	47-53 45-52 48-53

Basing judgments on classroom performance and linking it to MEA performance established an important connection between classroom performance standards and external assessment standards, which enhances coherence and consistency within the assessment system.

Basing judgments on classroom performance and linking it to MEA performance established an important connection between classroom performance standards and external assessment standards, which enhances coherence and consistency within the assessment system.

Again, how can this process be applied to local assessment? Districts might replicate the contrasting groups method by asking teachers to make judgments about students' performance levels and then linking them to particular scores on district-wide assessments. In such a case, teachers would draw on informal observations, grades, scores on standardized tests, and other relevant sources of evidence to inform their judgments. For instance, if a district administers a district-wide sixth grade science test, all sixth-grade teachers might participate in a contrasting groups process to set performance standards for the test. Each teacher would assign a selected sample of their students a performance level based on recent grades and other academic considerations. Linking these ratings with raw scores on the test identifies the cut scores differentiating the performance levels. This process can enhance the coherence of the assessment system, including the procedure for assigning grades.

General Considerations for Setting Performance Standards for Local Assessment

Before engaging in formal standard setting procedures, educators must consider several general matters for defining and structuring the standard setting process. While these issues are not specifically technical, they outline a framework within which standard setting takes place.

Who should be involved in standard setting, and what role should they play? Where appropriate and feasible, individuals who have a stake in the results of the assessment should be asked to participate in setting standards. In the case of local assessment, this may include school committee members, principals, teachers, students, parents, community members, or any combination of these stakeholders. These individuals would be told what the assessments are designed to measure, the population of the pupils to be assessed, the decisions to be made from the results, and the possible consequences of these decisions. Schools may wish to establish an advisory group to offer guidance regarding the best way to approach standard setting and to serve as a standard-setting group over time.

Individuals who have a stake in the results of the assessment should be asked to participate in setting standards.

What does it mean to achieve the Learning Results? Throughout this discussion, we have used the phrase "meeting the standard." This phrase must be clearly defined at the local level in order to proceed with standard setting. Does it mean that each student must attain all content standards of the *Learning Results*? Does it mean that each student must achieve a certain percentage of the *Learning Results*? Does it mean attaining a subset of content standards that are, in some way, more significant or relevant? Or does it mean something else? Performance standards will be structured around the resolution of these important questions.

What performance levels describe local performance standards? Local districts have a number of choices when selecting or developing performance levels. The MEA performance levels and definitions may be used to establish the local performance standards. This helps to ensure consistency and coherence between local assessments and state assessments when combined in an assessment system. Alternatively, local performance levels may be based on the MEA levels, but with added emphasis or specificity to highlight local priorities. They may be tailored for specific disciplines or for the various grade spans. Local districts may also choose to create their own performance level descriptors, without reference to the MEA levels. In any case, all standard setting activities depend on clearly defined performance levels.

Standards, once set, are not necessarily set forever; they need to be revisited.

Revisiting standards once they are set. Judgments are imperfect, and sometimes incorrect. Standards, once set, are not necessarily set forever; they need to be revisited. If a standard is found to be unfair or incorrect, it should be revised before final decisions are made. For example, if an item on a test is found to be confusing or without a correct answer, a teacher probably will decide not to count the item when assigning grades. However, it is inappropriate to discount an item just because the majority of students failed to answer it correctly. “Lowering standards to guarantee high grades discourages pupil effort and seriousness and diminishes the validity of the [test or assessment]. Fairness means teaching pupils the things on which they are assessed, using assessment procedures that are clear and suited to the pupils’ level and classroom experiences, and establishing performance standards ... that are realistic if pupils work hard. These are the teacher’s responsibilities in instruction, assessment, and grading” (Airasian, 1994, p. 298). Performance standards must permit decisions—decisions about individual learning, decisions about certification of individuals, and decisions about school accountability.

The clear, public nature of all expectations is central to the concept of a standards-based system.

Holding performance standards steady. After allowing sufficient opportunity to refine or correct performance standards, assessment systems must commit to holding the standards steady for several years. Without this consistency, progress will be difficult to document. Further, without time to communicate and illustrate the performance standards, and time for students and teachers to become oriented to the goal of achieving standards, the standards’ capacity to present an unequivocal target will be diminished. The clear, public nature of all expectations is central to the concept of a standards-based system. The *Learning Results* provide a clear, public answer to the question, “What do we want Maine students to know and be able to do?” We must provide an equally clear and public answer to the question, “How good is good enough?”

Establishing Performance Standards for a Local Assessment System

In addition to establishing performance standards for individual items, tasks, or instruments, local districts must set standards for performance across their system of assessment. In order to demonstrate proficiency on any standard or set of standards, performance across the various measures in the system must be evaluated. Consider the Mathematics standards in the *Learning Results*. A student might be asked to complete a variety of classroom assessments, compile a portfolio, present a data analysis project, take a district test, and sit for the MEA. The student's performance across all of these measures must be considered to draw a conclusion about proficiency. There various ways this might be accomplished. If each assessment yields a numerical score or rating of some kind, the resulting scores can be expressed in equivalent scales and averaged. Similarly, they might be converted to equivalent terms and weighted to reflect the varied scope or importance of each source of evidence. In another scenario, a certain level of performance on one or more specified measures (e.g., the MEA, a district test, and a portfolio) might serve as sufficient evidence of proficiency. Finally, local educators may develop strategies for reviewing evidence and data generated by a suite of assessments in combination to identify patterns of performance considered adequate to demonstrate proficiency.

A final word. In developing a local assessment system, the first task of educators will be to select, adapt, and develop the array of assessments necessary to provide adequate evidence that their students are making progress toward mastering the *Learning Results*. An important next step challenges us to identify a meaningful and manageable method for establishing performance standards for that system. This undertaking must proceed over time, carefully considering the theory and methodology of traditional standard setting and applying it in new ways to serve an assessment *system*.

SNAPSHOT

"Can you recommend a consultant or someone from the Maine Department of Education to help my assessment committee review our new reading assessments for reliability? I understand inter-rater agreement, but the other reliability methods in *Measured Measures* are beyond me," e-mailed Theresa to Alice.

The two had stayed in touch since meeting at a conference some months earlier, and they occasionally e-mailed or phoned one another for advice as they grappled with the task of developing a local assessment system. Theresa went on to say that, before speaking to a consultant, she'd like additional resources to shore up her understanding of internal-consistency reliability. She asked Alice to recommend articles or books to read.

"I can give you a couple of readings," Alice replied. "In the meantime, have you read anything new on standard setting? The section in *Measured Measures* is a useful overview, but I'd now like to see more of the details. Didn't you mention something that was due out soon?" "Yes," Theresa responded, "there's a piece by Ron Hamilton that is very good. I think it's Hambleton, Jaeger, and Mills (2000)."

References

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). New York: McGraw Hill.

Chudowsky, N., &, Behuniak, P. (1998). Using focus groups to examine the consequential aspect of validity. *Educational Measurement: Issues and Practices, 17*(4), 28-35.

Hambleton, R. K., Jaeger, R. M., & Mills, C. N. (2000). *Handbook of standard-setting methods*. Washington, DC: Council of Chief State School Officers.

Jaeger, R. M. (1993). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.; pp. 485-514). New York: Macmillan.

Lane, S., Parke, C. S., Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practices, 17*(2), 24-28.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and evaluation in teaching* (8th ed.). New York: Macmillan.

Maine Department of Education (1997). *State of Maine Learning Results*. Augusta, ME: The author.

Maine Mathematics and Science Alliance (1997). *A user's guide for reaching the standards*. Augusta, ME: The author.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Minium, E. W., Clarke, R. C., & Coladarci, T. (1999). *Elements of statistical reasoning* (2nd ed.). New York: Wiley.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed.). Boston: Allyn & Bacon.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. Columbus, OH: Merrill.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: National Institute for Science Education (NISE), University of Wisconsin-Madison.

Wiggins, G. (1998). *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass.

Appendices

Appendix A

Developing Assessments for the Learning Results: Dimensions of Alignment

Alignment describes the relationship between standards and assessments. For any standards-based assessment, such as an instrument intended to measure student mastery of the *Learning Results*, the degree of alignment defines the content validity of the assessment. As is the case with validity in general, the question of alignment includes a number of dimensions. Webb (1997) identified various dimensions of the alignment between standards and assessment. We believe that these dimensions provide a useful framework for local districts seeking to establish alignment as an indicator of content validity.

Here, we illustrate the applicability of some of Webb's dimensions to the development of assessments that are to be aligned with the *Learning Results*. With respect to the *Learning Results*, performance indicators provide the level of specificity necessary to establish alignment. Therefore, the considerations below emphasize the alignment between performance indicators and assessments.

For each of the selected dimensions, we provide a brief overview, an example drawn from the performance indicators in Maine's *Learning Results*, and a suggested procedure for addressing that dimension of alignment in the development and selection of assessments. These procedures are by no means prescriptive. Rather, we offer them to help educators envision practical ways for addressing alignment in local assessment systems.

Categorical Concurrence

The reporting categories for the assessment system should have a one-to-one correspondence with the content standards being assessed. That is, the content standards serve as organizers for reporting student performance.

Example. The reporting categories for the content area Visual and Performing Arts should include Creative Expression, Cultural Heritage, and Criticism and Aesthetics (because these are the content standards for this content area).

Suggested procedure. A procedure—probably self-evident—for addressing this particular dimension of alignment is to use content standards as reporting categories. This requires that, to allow for reporting on each content standard, the system include a sufficient number of assessments aligned with the performance indicators associated with each content standard.

Balance of Representation

Assessments should reflect the balance of representation of skills and knowledge found in the standards. Assessments should be weighted (a) to represent the relative importance of individual performance indicators to their discipline and (b) to take into account the range in the scope of indicators (some being broad and inclusive, others being focused and precise). Assessments should be selected or developed to mirror the emphasis interpreted from the content standards and performance indicators.

Example. Consider the following performance indicator for the Mathematics content standard, Computation (grades 5-8): "Compute and model all four operations with whole numbers, fractions, decimals, sets of numbers, and percents, applying the proper order of operations." This arguably should receive greater emphasis than the Probability performance indicator, "Explain the idea that probability can be represented as a fraction between and including zero and one."

Suggested procedure. Here is an example of a procedure for addressing this particular dimension of alignment:

- As a group, educators at each identified grade level analyze content standards: What is the relative scope, importance, or priority of each content standard? In making decisions, consider the taught curriculum, developmental issues, and other discipline documents (e.g., national standards) as resources.

- Divide 100 points among the content standards to reflect the appropriate weighting in curriculum, instruction, and assessment.

- Within the value assigned to each content standard, consider the performance indicators: What is the relative scope, importance, or priority of each indicator?

- For each content standard, distribute the content standard's assigned point value among its performance indicators. As an example, if a particular content standard is intended to receive 20% of the assessment weighting, these 20 points would be doled out among the several performance indicators for that content standard. Again, establish priorities among the indicators by virtue of their scope or importance within the discipline/content standard at that grade level.

- Reach consensus on the "rating" for each indicator.

- Develop assessments in priority and in proportion to that weighting.

Depth of Knowledge Consistency

Assessments should require the same level of cognitive demand as is indicated by the verb chosen to state the performance indicator. This ensures that the rigor of assessments corresponds to the difficulty intended by the performance indicator. While a word list from Bloom's taxonomy might be a useful tool for determining this alignment, one must ensure that the alignment is substantial and not merely semantic.

Example. The following is a performance indicator for the Science content standard, Classifying Life Forms (grades 5-8): "Compare systems of classifying organisms, including systems used by scientists." This performance indicator would not align with an assessment that, say, asks students to "Describe the system that scientists use to classify organisms." This question is posed at a lower level of thinking and, consequently, does not require the analysis implicit in the performance indicator.

Suggested procedure. Here is an example of a procedure for addressing this particular dimension of alignment:

- Review the performance indicators and the assessment being considered: Is there alignment in the cognitive demand? Is the assessment requiring the same level of thinking or rigor that is included in the performance indicator?

- If there is not alignment in this regard, consider what revision would be necessary to bring the assessment into alignment with the indicator, or what additional assessments might be created or selected to address the indicator's cognitive demand.

- Consider whether the revision or additional assessment development is "worth it," or if the performance indicator is more appropriately measured through other forms of assessment.

- If appropriate, revise or develop additional assessments to address the cognitive demand described by the performance indicator.

Range of Knowledge Correspondence

Assessments should be developed to encompass as much of the range indicated in a performance indicator as possible.

Example. If a performance indicator refers to use of a "variety of types of graphs," then the assessment (or combination of assessments) should require the use of the same "variety of types of graphs" as is inferred from the indicator.

Suggested procedure. Here is an example of a procedure for addressing this particular dimension of alignment:

- Review the performance indicators and the assessment being considered: Is there alignment between the range of knowledge in the performance indicator and the range of knowledge encompassed by the assessment?

- If there is not alignment in this regard, then consider broadening the scope of the assessment being reviewed, or creating or selecting additional assessments to adequately address the scope of the performance indicator.

Fairness

Assessments should clearly address the expectations specified by the performance indicators so that all students are afforded a fair opportunity to demonstrate corresponding skills and knowledge. Assessments should be prescriptive enough to require the demonstration of the expected skills and knowledge so that student interpretation will not dilute the intended demand of the performance indicator.

Example. The following performance indicator is for the Career Preparation content standard (grades 3-4), Integrated and Applied Learning: "Identify the major components of a technological system (input, process, output, feedback) and cite examples in the school and/or community." An assessment addressing this indicator should specify the use of the appropriate terms, and the assessment should make explicit that multiple examples will yield a better score.

Suggested procedure. Here is an example of a procedure for addressing this particular dimension of alignment:

- Review the performance indicators and the assessment being considered: Are all expectations that are specified by the performance indicator clearly communicated in the assessment so that students are afforded a fair opportunity to demonstrate those skills and knowledge?

- If expectations are not sufficiently clear in the assessment, revise the assessment accordingly (e.g., make requirements more explicit, make directions more understandable).

Cognitive Soundness

Assessments developed at each grade level should be developmentally appropriate. Further, they should reflect a continuum of intellectual development and sophistication.

Example. Consider the following performance indicator for the English Language Arts content standard, Literature and Culture (grades 3-4): “Use literary pieces to better understand and appreciate the actions of others.” The reading passages selected for an assessment should be of an appropriate reading level, and these passages should deal with situations that are understandable and relevant to Maine students in the targeted grades. (The suggested procedure for addressing this dimension of alignment is combined with the dimension that follows.)

Cumulative Growth in Content Knowledge

Assessments for each grade level test should be checked against the relevant performance indicators at other grade levels to ensure a deliberate increase in cognitive expectations that parallels research about how students learn at various grade levels.

Example. One of the Social Studies content standards is Human Interaction with Environments, where performance indicators address the relationship between a culture and its environment. At grades 5-8, for example, we find the following performance indicator: “Explain how cultures differ in their use of similar environments and resources.” At the secondary level, we find this one: “Analyze the cultural characteristics that make specific regions of the world distinctive.” While the content of the two performance indicators is similar, assessments at the secondary level should be more demanding—perhaps by virtue of the expected scope of response, the use of independent research, the requirement of abstraction or generalization supported by specific examples, and so on.

Suggested procedure. Here is an example of a procedure for addressing the two dimensions of alignment, cognitive soundness and cumulative growth:

- Review the comparable performance indicators at grade ranges above and below the performance indicator: Is the assessment developmentally appropriate? Is the context familiar and of possible relevance? Is the assessment appropriately demanding?

- Confirm that the developmental level of the assessment aligns with the performance indicator at the appropriate grade level and, if not, adjust the difficulty and/or context of the assessment as necessary.

Appendix B

Calculating Inter-Rater Reliability

Suppose that the writing samples of 50 students are independently rated by two teachers using the same scoring rubric, which classifies performance on a scale from 1 to 6. Table 10, which we have adapted from Linn and Gronlund (2000), summarizes the resulting data.

Table 10
Establishing Inter-Rater Reliability (Two Raters, Six-Point Scale).

		Score Assigned by Rater 2						
	Score →	1	2	3	4	5	6	row
	↓							total
	6	0	0	0	1	(1)	3	5
Score Assigned by Rater 1	5	0	0	1	(2)	4	(3)	10
	4	0	1	(2)	4	(2)	1	10
	3	0	(2)	5	(3)	2	0	12
	2	(1)	7	(1)	0	0	0	9
	1	3	(1)	0	0	0	0	4
	column total	4	11	9	10	9	7	50

Notice that the row and column totals are the frequencies with which each rater assigned the various score values. For example, Rater 1 (row totals) assigned the highest score to five writing samples whereas Rater 2 (column totals) assigned this score to seven writing samples. (A quick examination of row and column totals can reveal differences between raters in scoring leniency, which also can be detected by calculating the mean rating for each rater.)

The shaded portion of this table—the "diagonal"—represents the cases of exact agreement between the two raters. For instance, both raters assigned a 1 to the same writing sample on three occasions, a 2 on seven occasions, a 3 on five occasions, and so on. *The percentage of exact agreement* is obtained by summing the shaded values, dividing by the total number of writing samples, and multiplying by 100: (26/50)100 = 52%. *The percentage of agreement within one point* is calculated in a similar fashion, but the numerator now also includes the parenthetical values in Table 10: (44/50)100 = 88%.

This procedure can be applied to summative judgments, such as “proficient” versus “not proficient,” “meets standard” versus “does not meet standard,” and so forth. As an example, imagine that a student must obtain a score of at least 5 on this writing sample in order to meet a local standard. These dichotomous judgments are summarized in Table 11. As before, the percentage of exact agreement is based on the shaded portion of the table: (41/50)100 = 82%.

Table 11
Establishing Inter-Rater Reliability (Two Raters, Dichotomous Scale).

		Judgment of Rater 2		
	Judgment → ↓	“Does Not Meet Standard”	“Meets Standard”	row total
Judgment of Rater 1	“Meets Standard”	4	11	15
	“Does Not Meet Standard”	30	5	35
	column total	34	16	50

Appendix C

Calculating the Pearson Correlation Coefficient (r)

Imagine you want to examine the inter-rater reliability of a particular end-of-unit exam. Toward this end, you give the test to six students and have the six exams scored by two teachers: yourself and a willing accomplice. (Although highly unrealistic, this small number of students simplifies our example considerably!) Let's refer to the scores that you assigned as "*X*," and the scores that your colleague assigned as "*Y*." The scores for *X* and *Y* appear in Table 12, along with the necessary terms for calculating the correlation between the two sets of scores.

As you see from the column headings of Table 12, we need to square each *X* score and each *Y* score (third and fifth columns, respectively). Consider Student A: This student's score on *X* is 5, which when squared

Table 12
Calculating the Correlation (Pearson r) Between Two Scores, X and Y.

Student	X (you scored)	X^2	Y (colleague scored)	Y^2	XY
A	5	25	6	36	30
B	1	1	2	4	2
C	7	49	10	100	70
D	9	81	8	64	72
E	3	9	1	1	3
F	4	16	4	16	16
$n = 6$	$\Sigma X = 29$	$\Sigma X^2 = 181$	$\Sigma Y = 31$	$\Sigma Y^2 = 221$	$\Sigma XY = 193$

gives us 25. We also need to obtain the product of each pair of scores (last column). For instance, the product for Student D is (9)(8) = 72.

Now look at the last row of Table 12, where you see the Greek symbol "Σ." This symbol stands for the operation of summation. $\sum X$, for example, instructs you to sum the X scores (5 + 1 + 7 + 9 + 3 + 4 = 29). The six sums appearing in the last row of this table are needed to calculate the correlation coefficient, Pearson r. Although the formula for Pearson r is somewhat daunting, it is easily negotiated if you take it one term at a time.[3] We encourage you to study our calculations that follow, and verify that you arrive at the same answer we do!

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

$$= \frac{193 - \frac{(29)(31)}{6}}{\sqrt{\left(181 - \frac{(29)^2}{6}\right)\left(221 - \frac{(31)^2}{6}\right)}}$$

$$= \frac{193 - 149.83}{\sqrt{\left(181 - \frac{841}{6}\right)\left(221 - \frac{961}{6}\right)}}$$

$$= \frac{43.17}{\sqrt{(40.83)(60.83)}} = \frac{43.17}{\sqrt{2483.69}} = \frac{43.17}{49.84} = .87$$

Thus, .87 is the inter-rater reliability for this test.

[3] We have taken this formula from Minium, Clarke, and Coladarci (1999, p. 118).

Appendix D

Calculating Equivalent-Forms Reliability

Suppose your school annually administers a reading proficiency test at the end of the second grade, and there are two forms to allow for retakes. You wish to determine the equivalent-forms reliability of this test, so you select a sample of 40 second graders and have each child complete both forms of the test (with a juice break in between). The two sets of exams are then scored and proficiency judgments made (at your school, proficiency is defined as a score of at least 70%). As illustrated in Table 13, each student thus has two proficiency judgments: one from each form of the test. The frequencies in the shaded portions of this table show where the same judgment was made from both forms of the proficiency test. A measure of equivalent-forms reliability is obtained by summing the shaded values, dividing by the total number of students, and multiplying by 100: (36/40)100 = 90%.

For an example of equivalent-forms reliability that involves Pearson *r*, simply consider *X* and *Y* in Table 12 as equivalent assessments (rather than different raters for a single assessment).

Table 13
Establishing Equivalent-Forms Reliability.

		Judgment Based on Form 2		
	Judgment → ↓	"Not Proficient"	"Proficient"	row total
Judgment Based on Form 1	"Proficient"	3	25	28
	"Not Proficient"	11	1	12
	column total	14	26	40

Appendix E

Calculating Internal-Consistency Reliability: KR-21

KR-21 is appropriate for estimating the internal-consistency reliability of a test having items that are scored dichotomously (e.g., "correct" vs. "incorrect"). As an example, suppose you have given your six students a selected-response test having four items, and for each student you calculate a total score, X. We arrange the data as follows:[4]

	X	X^2
Student	(total score)	
David	1	1
Jeff	2	4
Jill	2	4
Judy	4	16
Mike	3	9
Sandy	0	0
$n = 6$	$\sum X = 12$	$\sum X^2 = 34$

[4] This example is adapted from Nitko (1996, p. 460). As in Table 12, we wish to minimize computational details. This scenario, of course, specifies an unrealistically small number of students and an unacceptably small number of test items. By the way, "KR" refers to the statistic's inventors—Kuder and Richardson—and "21" refers to the version of this particular formula.

One convenient formula for KR-21 is:

$$\text{KR-21} = \left(\frac{k}{k-1}\right)\left(1-\frac{\overline{X}\left(k-\overline{X}\right)}{(k)(S^2)}\right)$$

This formula involves three terms:

1. *k*, the number of items on the test.
   In the present example, $k = 4$.

2. $\overline{X}$, the arithmetic mean of the scores on the test.
   The arithmetic mean is determined by dividing the sum of scores ($\sum X$) by the number of scores (*n*). Stated mathematically, $\overline{X} = \sum X/n$. In the present example, $\overline{X} = 12/6 = 2$.

3. S^2, the variability among scores on the test.
   S^2, technically known as the "variance," simply reflects how much spread or dispersion there is in a group of scores. For example, if everyone received the same score, then $S^2 = 0$. The variance is obtained using the formula,

$$S^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}{n}$$

Using data from the present example,

$$S^2 = \frac{34 - \frac{(12)^2}{6}}{6} = 1.667$$

With $k = 4$, $\overline{X} = 2$, and $S^2 = 1.667$, you now can calculate KR-21:

$$\begin{aligned}
\text{KR-21} &= \left(\frac{k}{k-1}\right)\left(1-\frac{\overline{X}\left(k-\overline{X}\right)}{(k)(S^2)}\right)\\
&= \left(\frac{4}{4-1}\right)\left(1-\frac{2(4-2)}{(4)(1.667)}\right)\\
&= (1.333)\left(1-\frac{4}{6.668}\right)\\
&= (1.333)(1-.600)\\
&= .53
\end{aligned}$$

Thus, .53 is the internal-consistency reliability of this test. You are not alone if you find this value unacceptably low. If you obtained this reliability for an assessment of your own, your first task would be to identify the reason(s) why. As we indicate in the section, “Implications for classroom-level assessments,” reliability is related to the number of items or tasks on an assessment. Our paltry reliability coefficient doubtless is due, in part, to the brevity of this fictitious test (4 items).

KR-21, as we stated earlier, is a “shortcut” method for estimating internal-consistency reliability. It assumes that all items on the test are equally difficult (i.e., an equal proportion of students gets each item correct). If this assumption does not hold, then KR-21 will underestimate the reliability of the test.

Appendix F

Models for Rubric Development

Performance-based assessments require rubrics and scoring guides to produce reliable data about student performance. Such tools must also contribute to content validity by accurately reflecting the intended learning targets of the assessment. Rubrics present criteria and levels of performance, whereas scoring guides provide specific descriptions of performance that contribute to consistent scoring decisions. Rubrics and scoring guides serve as performance standards for the tasks, exhibits, or collections to which they are applied.

Rubrics come in many different shapes and forms. Table 14 describes four categories of rubrics.

Table 14
Four Categories of Rubrics, and Their Characteristics.

	Generic	*Task Specific*
Holistic	"Generic Holistic" describes overall levels of performance for any task or product	"Task Specific Holistic" describes overall levels of performance for a particular task, item, or project
Analytic	"Generic Analytic" describes levels of performance on more than one dimension for any task or product	"Task Specific Analytic" describes levels of performance on more than one dimension for a particular task, item, or project

(Source: Maine Mathematics and Science Alliance, 1997)

In the instructional setting, generic rubrics supply an image of "good work" to teachers and students. The descriptors found at the "proficiency," or "meeting the standard," level convey an image of satisfactory achievement. A generic rubric, used consistently, can illuminate expectations and requirements and feedback based on such a rubric guides students toward improvement. Generic rubrics can also serve as templates for task specific rubrics. In this case, specific descriptors, detailing the particular evidence at each level, are inserted within the framework (criteria and performance levels) of a generic rubric. A highly reliable scoring process calls for this specificity to produce consistent scores. Employing the same generic rubric(s) within, and across, an assessment system contributes to consistency in scoring procedures and standard setting.

The generic holistic rubric in Table 15 is used to score the MEA. It might serve as a template for task specific generic rubrics.

A holistic rubric, whether generic or task specific, produces a single score. This is sufficient for assessments addressing individual performance indicators. If, on the other hand, an assessment is complex and provides students with the opportunity to demonstrate several performance indicators, an analytic rubric allows scorers to assign separate values to performance on different dimensions of the task. A common grading system in English composition provides one grade for content and another grade for technical aspects of writing (grammar, usage, and mechanics). This is an example of analytic scoring, which breaks performance into its component parts. Likewise, an analytic rubric might produce separate scores for several aspects of a science project in which students pose a question about the motion of objects, design an experiment to investigate their question, carry out the experiment, collect and analyze data, and draw a conclusion about their question. Using an analytic rubric, a score reflecting the student's demonstrated content knowledge of motion is independent of the score describing his or her ability to pose a question and design an experiment (scientific inquiry skills) and of scores indicating command of scientific vocabulary and notation (scientific communication) and ability to support conclusions with data (scientific reasoning).

Again, a generic analytic rubric might serve as a template for the development of task-specific analytic rubrics and scoring guides. You find in Table 16 an example of a generic, analytic rubric for mathematics, which provides for scoring on four dimensions of a problem solving task: computation and problem solving, communication, reasoning, and mathematics content. (Note: Entries must align with at least one of the performance indicators listed beneath each criterion, and the full collection should address the range of performance indicators.)

Table 15

The Generic Holistic Rubric Used in Scoring the MEA.

Category 4	1. The student completes all important components of the task and communicates ideas clearly. 2. The student demonstrates in-depth understanding of the relevant concepts and/or processes. 3. Where appropriate, the student chooses more efficient and/or sophisticated processes. 4. Where appropriate, the student offers insightful interpretations or extensions (generalizations, applications, analogies).
Category 3	1. The student completes most important components of the task and communicates clearly. 2. The student demonstrates understanding of major concepts even though he/she overlooks or misunderstands some less important ideas or details.
Category 2	1. The student completes some important components of the task. 2. The student demonstrates that there are gaps in his/her conceptual understanding.
Category 1	1. Student shows minimal understanding. 2. Student addresses only small portion of the required task(s).
Category 0	1. Response totally incorrect or irrelevant.
Blank	Blank/no response

(Source: Maine Department of Education)

Table 16

Mathematics Rubric and Scoring Guide: Maine Assessment Portfolio.

	1 *Attempted Demonstration* (little evidence)	2 *Partial Demonstration* (some evidence)	3 *Proficient Demonstration* (evidence meets standards)	4 *Sophisticated Demonstration* (evidence exceeds standards)
Computation and Problem Solving (*B*) 1. Compute and model all four operations with whole numbers, fractions, decimals, sets of numbers, and percents, applying the proper order of operations. 2. Create, solve, and justify the solution for multi-step, real-life problems including those with ratio and proportion.	Employs inappropriate strategies and inaccurate or inappropriate application of computation skills.	Employs appropriate strategies, but includes some inaccurate and/or inappropriate application of computation skills.	Employs appropriate strategies and includes accurate, appropriate application of computation skills.	Employs sophisticated or or efficient strategies and includes accurate, appropriate application of computation skills.
Mathematical Reasoning (J) 1. Support reasoning by using models, known facts, properties, and relationships. 2. Demonstrate that multiple paths to a conclusion may exist.	Explanation lacks coherence, is not relevant, and/or relies on information not necessary to complete the task.	Explanation used to justify and explain solution is not connected to information generated while completing the task, focuses on concrete aspects and does not generalize.	Explanation used to justify and explain solution is supported by evidence gathered while completing the task. Offers generalized conclusions, and identifies relevant and irrelevant information.	Explanation used to justify and explain solutions includes relevant information from student's experience beyond the requirements of the task, meaningfully generalizes conclusions beyond the scope of the task, and identifies relevant and irrelevant information and the impact of each on completing the task.
Mathematical Comunication (K) 1. Translate relationships into algebraic notations. 2. Use statistics, tables, and graphs to communicate ideas and information in convincing presentations and analyze presentations of others for bias or deceptive presentation.	Includes little or no mathematical terminology, symbols, or visual representation incorrectly to report, explain, enhance, or clarify.	Includes mathematical terminology, symbols, or visual representation incorrectly, or inconsistently used to report, explain, enhance, or clarify.	Includes clear, accurate appropriate communication including mathematical terminology, symbols, and/or visual representation to report, explain, enhance, and clarify.	Includes clear, elegant communication using sophisticated mathematical terminology, symbols, and/or visual representation to report, explain, enhance, and clarify.
Mathematical Content Students will understand and demonstrate math content including: A. numbers and number sense; C. data analysis and statistics; D. probability; E. geometry; F. measurement; G. patterns, relations, and functions; H. algebra concepts; I. discrete mathematics.	No significant demonstration of any performance indicators from content standards A, C, D, E, F, G, H, or I for this grade span.	Some demonstration of performance indicator # ____ from content standard ____ for this grade span.	Accurate, appropriate demonstration of performance indicator(s) # ____ from content standard ____ for this grade span.	Exceeds expectations in demonstrating performance indicator(s) # ____ from content standard ____ for this grade span.

Definitions: *Sophisticated*—exceeding the expectation of an age or developmental level, applying skill/concepts in novel way. *Efficient*—demonstrating unusual insight through use of a more direct approach than is typical. *Elegant*—concise and precise. (Source: Maine Assessment Portfolio, Maine Mathematics and Science Alliance, and Maine Department of Education.)

Appendix G

Guidelines for Developing Quality Rubrics

The following descriptions of rubrics are taken directly from Wiggins (1998). Readers may find them helpful in developing rubrics for assessments contributing to local assessment systems.

Rubrics are best when they . . .

- *are sufficiently generic to relate to general goals* beyond an individual performance task, but specific enough to enable useful and sound inferences about the task.
- *discriminate among performances validly*, not arbitrarily, by assessing the central features of performance, not those that are easiest to see, count, or score.
- *do not combine independent criteria in one rubric.*
- *are based on analysis of many work samples* and on the widest possible range of work samples, including valid exemplars.
- *rely on descriptive language* (what quality or its absence looks like) as opposed to merely comparative or evaluative language, such as "not as thorough as" or "excellent product", to make a discrimination.
- *provide useful and apt discrimination* that enables sufficiently fine judgments, but do not use so many points on the scale (typically more than six) that reliability is threatened.
- *use descriptors that are sufficiently rich* to enable student performers to verify their scores, accurately self-assess, and self-correct.
- *highlight judging the impact of performance* (the effect, given the purpose) rather than over-reward processes, formats, content, or the good-faith effort made.

(Source: Wiggins, 1998)

Rubrics that meet technical requirements are . . . ↓	
continuous	The change in quality from score point to score point is equal: the degree of difference between a 5 and a 4 is the same as between a 2 and a 1. The descriptors reflect this continuity.
parallel	Each descriptor parallels all the others in terms of the criteria language used in each sentence.
coherent	The rubric focuses on the same criteria throughout. Although the descriptor for each scale point is different from the ones before and after, the changes concern variance of quality for the (fixed) criteria, not language that explicitly or implicitly introduces new criteria or shifts the importance of the various criteria.
aptly weighted	When multiple rubrics are used to assess one event, there is an apt, not arbitrary, weighting of each criterion in reference to the others.
valid	The rubric permits valid inferences about performance to the degree that what is scored is what is central to performance, not what is merely easy to see and score. The proposed differences in quality should reflect task analysis and be based on samples of work across the full range of performance; describe qualitative, not quantitative, differences in performance; and not confuse merely correlative behaviors with actual authentic criteria.
reliable	The rubric enables consistent scoring across judges and time. Rubrics allow reliable scoring to the degree that evaluative language (“excellent”, “poor”) and comparative language (“better than”, “worse than”) is transformed into highly descriptive language that helps judges to recognize the salient and distinctive features of each level of performance” (Wiggins, 1998).

(Source: Wiggins, 1998)