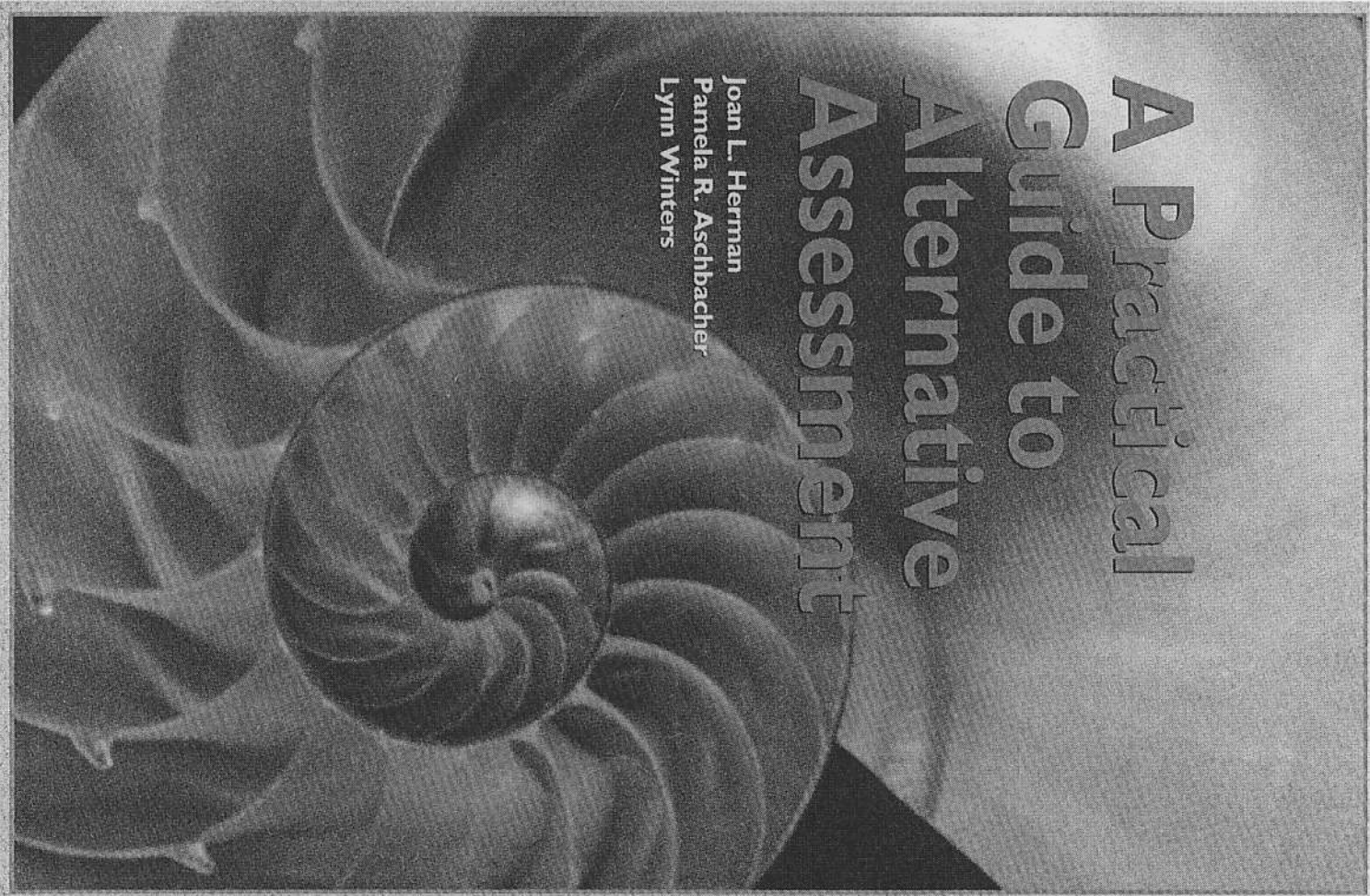


0095-1

A Practical Guide to Alternative Assessment

Joan L. Herman
Pamela R. Aschbacher
Lynn Winters



0095-1

5

Setting Criteria

The criteria used for judging student performance lie at the heart of alternative assessment. Although we have discussed selecting and describing assessment tasks separately from developing scoring criteria, these three aspects of assessment are intimately intertwined. In the absence of criteria, assessment tasks remain just that, tasks or instructional activities. Perhaps most important, scoring criteria make public what is being judged and, in many cases, the standards for acceptable performance. Thus, criteria communicate your goals and achievement standards.

Like "alternative assessment" itself, criteria for judging student performance have been called many things, including scoring criteria, scoring guidelines, rubrics, and scoring rubrics. For our purposes, we take all these terms to mean a **description of the dimensions** for judging student performance, a **scale of values** for rating those dimensions, and, when appropriate, the **standards** for judging performance.

Let's take a common example from social studies. You assign students a group presentation accompanied by individual written reports to assess their understanding of history. Because you wish to assess three skills—oral, written, and group process skills as they relate to history—you must consider scoring criteria for each skill. Figure 5.1 on pages

46–47 is a possible set of scoring criteria for just one of these skills, a history group process assessment developed by the California Assessment Program.¹

The group process exercise taps four **learning outcomes**: group learning, critical thinking, communication, and history knowledge. For each outcome, scoring **dimensions** are specified and levels of performance differentiated by a **scoring scale**. Finally, the scoring guide includes an **evaluation** of each performance level, labeling performance not only in terms of what was accomplished but how well, from minimal to exceptional achievement.

Understanding the Need for Criteria

Criteria are necessary because they help you judge complex human performance in a reliable, fair and valid manner. Scoring criteria guide your judgments and make public to students, parents, and others the basis for these judgments. Scoring a multiple-choice test does not require complicated judgment; nevertheless, human judgment is still a factor because the test developer phrases the questions and decides what constitutes the best answers. To the person who scores the test, a student either has or has not selected the correct answer; no judgment is needed. When we use selected-response tests, we are essentially corroborating the judgments about adequate performance built into the "answer key." Thus, all assessment, be it selected- or constructed-response, has a subjective or human judgment component.

Alternative assessments invite a wider range of possible responses. Instead of judging responses as right or wrong, alternative assessments judge the quality of, and sometimes the process of, arriving at a complex response. To make such judgments and to ensure their validity, consistency, and fairness, we need criteria or scoring guidelines. Scoring criteria must be well-conceived, explicitly defined, and consistently applied. Well-specified criteria help to ensure that everyone understands what is expected.

Well-articulated and publicly visible criteria for judging student responses are necessary and useful whether the results will be used in

¹Many of the examples we use throughout this book are from state assessment programs, especially those in California. Because of their pioneering work in developing curriculum frameworks reflecting current learning and curriculum theory, certain states have already field-tested promising prototypes for alternative assessment that can be adapted for classroom use.

S-2900

Figure 5.1
California Assessment Program 1990
History-Social Science Grade 11
Scoring Guide: Group Performance Task

	Level I Minimal Achievement	Level II Rudimentary Achievement	Level III Commendable Achievement	Level IV Superior Achievement	Level V Exceptional Achievement
Group and Collaborative Learning 20	(1-4) Exclusive reliance on one spokesperson. Little interaction. Very brief conversations. Some students are disinterested or distracted.	(5-9) Strong reliance on spokespersons. Only one or two persons actively participate. Sporadic interaction. Conversation not entirely centered on topic.	(8-12) Some ability to interact. At least half the students confer or present ideas. Attentive reading of documents and listening. Some evidence of discussion of alternatives.	(13-16) Students show adeptness in interacting. At least 3/4 of students actively participate. Lively discussion centers on the task.	(17-20) Almost all students enthusiastically participate. Responsibility for task is shared. Students reflect awareness of others' views and opinions and include references to other opinions or alternatives in presentation and answer. Questions and answers illustrate forethought and preparation.
Critical Thinking 30	(1-6) Demonstrates little understanding and only limited comprehension of scope of problem or issues. Employs only the most basic parts of information provided. Mixes fact and opinion in developing a viewpoint. States conclusion after hasty or cursory look at only one or two pieces of information. Does not consider consequences.	(7-12) Demonstrates only a very general understanding of scope of problem focuses on a single issue. Employs only the information provided. May include opinion as well as fact in developing a position. States conclusion after limited examination of evidence with little concern for consequences.	(13-18) Demonstrates a general understanding of scope of problem and more than one of the issues involved. Employs the main points of information from the documents and at least one general idea from personal knowledge to develop a position. Builds conclusion on examination of information and some consideration of consequences.	(19-24) Demonstrates clear understanding of scope of problem and at least two central issues. Uses the main points of information from the documents and personal knowledge that is relevant and consistent in developing a position. Builds conclusion on examination of the major evidence. Considers at least one alternative action and the possible consequences.	(25-30) Demonstrates a clear, accurate understanding of the scope of the problem and the ramifications of the issues involved. Employs all information from the documents and extensive personal knowledge that is factually relevant, accurate, and consistent in the development of a position. Bases conclusion on a thorough examination of the evidence, an exploration of reasonable alternatives, and an evaluation of consequences.

46

Figure 5.1 (continued)

	Level I Minimal Achievement	Level II Rudimentary Achievement	Level III Commendable Achievement	Level IV Superior Achievement	Level V Exceptional Achievement
Communication of Ideas 20	(1-4) Position is vague. Presentation is brief and includes unrelated general statements. Overall view of the problem is not clear. Statements tend to wander or ramble.	(5-9) Presents general and indefinite position. Only minimal organization in presentation. Uses generalities to support position. Emphasizes only one issue. Considers only one aspect of problem.	(8-12) Takes a definite but general position. Presents a somewhat organized argument. Uses general terms with limited evidence that may not be totally accurate. Deals with a limited number of issues. Views problem within a somewhat limited range.	(13-16) Takes a clear position. Presents an organized argument with perhaps only minor errors in the supporting evidence. Deals with the major issues and shows some understanding of relationships. Gives consideration to examination of more than one idea or aspect of the problem.	(17-20) Takes a strong, well-defined position. Presents a well-organized, persuasive argument with accurate supporting evidence. Deals with all significant issues and demonstrates a depth of understanding of important relationships. Examines the problem from several positions.
Knowledge and Use of History 30	(1-6) Reiterates one or two facts without complete accuracy. Deals only briefly and vaguely with concepts or the issues. Barely indicates any previous historical knowledge. Relies heavily on the information provided.	(7-12) Provides only basic facts with only some degree of accuracy. Refers to information to explain at least one issue or concept in general terms. Limited use of previous historical knowledge without complete accuracy. Major reliance on the information provided.	(13-18) Relates only major facts to the basic issues with a fair degree of accuracy. Analyzes information to explain at least one issue or concept with substantive support. Uses general ideas from previous historical knowledge with fair degree of accuracy.	(19-24) Offers accurate analysis of the documents. Provides facts to relate to the major issues involved. Uses previous general historical knowledge to examine issues involved.	(25-30) Offers accurate analysis of the information and issues. Provides a variety of facts to explore major and minor issues and concepts involved. Extensively uses previous historical knowledge to provide an in-depth understanding of the problem and to relate it to past and possible future situations.

47

0095-3

the classroom or to make school level or national decisions. In all assessment settings, scoring criteria must:

- Help teachers define excellence and plan how to help students achieve it.
- Communicate to students what constitutes excellence and how to evaluate their own work.
- Communicate goals and results to parents and others.
- Help teachers or other raters be accurate, unbiased, and consistent in scoring.
- Document the procedures used in making important judgments about students.

Criteria and Instructional Planning

Scoring criteria clarify instructional goals. Along with the task description, the criteria define priority outcomes in terms of the content to be covered, the knowledge or skills to be demonstrated, and the context in which these are to occur. The complete alternative assessment specifications can guide selection and sequencing of relevant instructional activities.

Criteria and Students

The criteria for alternative assessments are often made public and are intended to be discussed with students. Public discussions help students to internalize the standards and "rules" they need to become independent learners. Alternative assessments and their criteria can be woven into the fabric of the curriculum so that they are transparent to the student and perceived as a natural part of the learning process. Such assessment is ongoing and takes many forms—journals, conferences, peer or teacher coaching episodes, critiques of products and exhibitions, and formal evaluations of individual works or a body of work. Examples of what constitutes good work engage students in the work itself and in judgments about their work. Public discussions of quality and criteria inform students during the formative period of instruction, not simply at the end of a unit or course when it is too late to make improvements. Furthermore, discussions of criteria also help students see the perspectives of their teachers, their peers, and sometimes even the experts in the field.

Criteria and Parent Involvement

Clearly articulated criteria also communicate to parents and others what the teachers and schools are trying to accomplish. Criteria operationalize learning goals and expectations for children. When parents know prior to grading what is expected, they can support their child's learning. For example, giving parents of kindergartners a copy of "Profile of Developmental Outcomes for Kindergarten" (Figure 5.2) allows them to work with their children at home on activities such as recognizing beginning letters or sight words. The road to literacy is well-marked; teachers who share the map with parents may find that more of their students reach their destinations in a timely manner.

Good criteria help both students and parents share some of the responsibility for learning. Parents and children who are familiar with the standards by which work is judged are less likely to ascribe poor performance to such external factors as not being told what was important or personality conflicts between teachers and students.

Criteria and Consistency

When guidelines for what constitutes good work are vague or unstated, it is difficult to be consistent, fair, and accurate in judging student responses. With selected-response tests, accuracy and consistency in scoring refers to whether the test score for an individual pupil remains fairly stable from one testing occasion to another, in the absence of intervening instruction or growth. This consistency is better known as reliability. For alternative assessments, reliability includes not only the idea of the stability of an individual student's performance over time but also the stability of a rater's judgments of that performance. Specifically, a reliable assessment that depends on human judgment must meet the following requirements:

- Several judges looking at a specific task would come to the same conclusion about a student.
- Each judge would rate the student's performance on a specific task about the same on a subsequent occasion.
- The student would perform the same task at about the same level on different occasions.
- If the task is meant to represent or generalize to some larger domain, the sample is representative of that domain.

Figure 5.2
Profile of Developmental Outcomes for Kindergarten
Literacy and Numeracy Skills

Joan C. Hillard, Superintendent, Spreckels Union School District, Spreckels, California
 Elizabeth Jones, Professor, Pacific Oaks College, Pasadena, California
 Jane Meade-Roberts, Director and Owner, Power of Play Preschool, Salinas, California
 San Vicente School, Soledad Union School District, Soledad, California
 (Jones and Meade-Roberts 1990)

Oral language	Is nonverbal in school	Uses language to satisfy basic wants and needs	Often uses language in play and conversation with peers	Clearly describes real or imaginary situations using complex descriptive language	Speaks in whole sentences using a well-developed vocabulary
Drawing	Scribbles	Draws a face	Adds arms/legs	Adds body with arms/legs	Adds details (hair, ear, hands, etc.)
Writing	Scribbles and pretends to write	Uses letters or letter like signs to represent writing	Spontaneously writes own name including all letters	Spontaneously copies words	Can invent spelling of words using phonetic clues
Reading	Reads own name	Recognizes beginning letter of first name when written in other places	Recognizes own name, other letters and numerals	Recognizes and reads sight words, including signs, labels, key words, teacher-created word lists and/or words in books	Uses knowledge of letter sounds to sound out words

Figure 5.2 (continued)

Attitudes toward literacy	Not yet interested in books or writing	Demonstrates focused interest in picture books	Demonstrates interest in written language (e.g., asks about or reads signs, names, words in class, labels, words in books)	Spontaneously practices writing letters and numerals	Demonstrates interest in writing correctly
Problem solving using classification	Randomly manipulates objects	Spontaneously orders by likenesses and differences	Recognizes or creates simple (AB) patterns using a variety of materials and/or symbols	Recognizes or creates complex (e.g., AABAAB) patterns using a variety of materials and/or symbols	Can classify by more than one attribute at a time (e.g., size and color)
Problem solving using numbers	Calls numerals at random	Counts by rote	Demonstrates understanding of one to one correspondence (e.g., evaluates objects accurately)	Is able to use knowledge of counting to solve real problems	Demonstrates conservation of number (e.g., understands that number of objects remains constant)

10000

0095-6

Figure 5.2 (continued)

Curiosity	Watches silently	Asks cautious questions	Asks questions constantly	Asks questions appropriately	Uses resources to find answers to questions (e.g., experimenting, taking risks, solving problems)
Creativity	Waits to be told what to do	Explores available materials	Invents a simple dramatization or projects with provided materials	Asks or looks for not already available materials to accomplish project/play idea	Works competently on notably complex, creative, imaginative, self-initiated tasks
Social skills with peers	Usually observes play with others	Usually plays alone or is involved in parallel play	Is developing cooperative play skills	Socially self-confident; plays effectively with other children	Has well-developed skills of leadership and cooperation in play
Social skills with adults/groups	Accepts situations rather than ask for adult help	Communicates with adults primarily to get help	Speaks spontaneously and freely with adults	Participates in group activities and conversation	Is sensitive to and articulate about the needs of others

Figure 5.2 (continued)

Large motor skills	Runs	Jumps	Hops on one foot	Catches a ball with arms and chest	Can catch ball with hands only
Fine motor skills	Scribbles with crayon/pencil	Able to use scissors	Colors inside lines/cuts on lines	Draws/writes accurate lines	Consistently neat work
New learning	Chooses to observe	Prefers familiar tasks	Willing to try new tasks	Masters new tasks quickly	Masters new tasks independently
Social knowledge	Knows colors	Knows shapes	Knows personal information	Knows names of letters and numbers	Knows days of week, months
Attention span	Rapidly changing	Focuses on self-selected tasks	Focus on teacher-selected tasks	Works independently on self- and teacher-selected tasks	Can follow complex directions and maintain focused attention for long periods

(From E. Jones and J.M. Roberts, *Profile of Developmental Outcomes for Kindergarten, Literacy and Numeracy Skills*, San Vicente School, Soledad CA)

0095-16

0695-7

It is easy to see how these four requirements for reliable scoring demand a mechanism for creating rater agreement and for delineating clearly the domains of particular assessment tasks. Scoring criteria must meet this demand.

Criteria and Consequences

Specifying criteria is always important and becomes even more so when the consequences of an assessment are very serious, such as when results are used for retention, graduation, or placement in special programs. Clear guidelines for evaluating student work ensure appropriate consequences for students and the educational system as a whole. Furthermore, when alternative assessments are used for these high-stakes decisions, the scoring procedures and criteria must be legally defensible and adhere to the due process standards of a court of law.

Specifying Criteria

Different testing purposes require different kinds of scoring criteria. Many of the examples in this book were developed for state-level assessments with such high-stakes testing purposes as comparing schools, identifying low-performing schools, and evaluating individual schools. The California Assessment Program (CAP) history group process criteria (shown in Figure 5.1) are an example of the complex criteria used in high-stakes assessment. Because the criteria are used for a one-shot state assessment, the scoring guide was developed to extract the maximum amount of information possible during limited assessment time. We see that the criteria:

- List multiple learning outcomes.
- Divide each outcome into performance levels.
- Describe traits/characteristics for each level.
- Provide a numerical scale to rate the degree to which each level was attained.
- Evaluate the quality of student performance represented by the different levels using such descriptors as "minimal achievement" or "excellent achievement."

Your criteria will be less complex when your testing purposes are more focused and the decisions you wish to make about students are limited.

If you are using student academic journals to monitor their progress in making connections between science lessons and their daily lives, your scoring criteria may be to count the number of unprompted statements connecting classroom learning with out-of-class experiences. The number of connections you find will tell you whether you are achieving your goals. Your assessment purpose here may be formative—to improve your instruction and to identify students who need more help or a different approach.

Perhaps your assessment purpose is more traditional—you want to evaluate student progress toward meeting your goals in mathematics problem solving. Your scoring criteria might resemble the generalized rubric for essay-type mathematics problems developed by the CAP (shown in Figure 5.3). The criteria provide descriptions of each level of performance in terms of what students are able to do, assign values to these levels, then apply standards at certain cut points. Students rated 1-2 are evaluated as having "inadequate" responses; students rated 3-4 receive a "satisfactory"; and students receiving 5-6 are rated "competent."

While grading is a complex issue and the scores of any one alternative assessment may or may not be used to assign grades, it is possible to find or develop criteria linked specifically to letter grades. Researchers funded by the National Science Foundation have developed a grade-linked set of criteria to assess student's procedural knowledge in a hands-on science experiment (Baxter et al. 1992). The researchers determined which methods students could use to solve the problem posed by the experiment, judged which would produce the most logical and efficient solutions, then created grade-referenced criteria to reflect their evaluations of the solutions. A summary of how their criteria is linked to grades appears in Figure 5.4.

Regardless of the testing purpose, the sample criteria have four common elements. Each has

- One or more traits or dimensions that serve as the basis for judging the student response
- Definitions and examples to clarify the meaning of each trait or dimension
- A scale of values (or a counting system) on which to rate each dimension
- Standards of excellence for specified performance levels accompanied by models or examples of each level.

0095-8

Figure 5.3
CAP Generalized Rubric
 (California State Department of Education 1989)

Demonstrated Competence

Exemplary Response . . . Rating = 6
 Gives a complete response with a clear, coherent, unambiguous, and elegant explanation; includes a clear and simplified diagram; communicates effectively to the identified audience; shows understanding of the open-ended problem's mathematical ideas and processes; identifies all the important elements of the problem; may include examples and counterexamples; presents strong supporting arguments.

Competent Response . . . Rating = 5
 Gives a fairly complete response with reasonably clear explanations; may include an appropriate diagram; communicates effectively to the identified audience; shows understanding of the problem's mathematical ideas and processes; identifies the most important elements of the problem; presents solid supporting arguments.

Satisfactory Response

Minor Flaws But Satisfactory . . . Rating = 4
 Completes the problem satisfactorily, but the explanation may be muddled; argumentation may be incomplete; diagram may be inappropriate or unclear; understands the underlying mathematical ideas; uses mathematical ideas effectively.

Serious Flaws But Nearly Satisfactory . . . Rating = 3
 Begins the problem appropriately but may fail to complete or may omit significant parts of the problem; may fail to show full understanding of mathematical ideas and processes; may make major computational errors; may misuse or fail to use mathematical terms; response may reflect an inappropriate strategy for solving the problem.

Inadequate Response

Begins, But Fails to Complete Problem . . . Rating = 2
 Explanation is not understandable; diagram may be unclear; shows no understanding of the problem situation; may make major computational errors.

Unable to Begin Effectively . . . Rating = 1
 Words do not reflect the problem; drawings misrepresent the problem situation; copies parts of the problem but without attempting a solution; fails to indicate which information is appropriate to problem.

No Attempt . . . Rating = 0

Figure 5.4
Linking Criteria to Grades

Grade	Criteria for Determining Grades
A	Student selects method. Student saturates towels. Result determines result so as to answer question. Result logically follows from method used to saturate towel. Measurements are accurate/carefully done. Conclusions are correct.
B	Meets all requirements of an "A" but measurement is careless.
C	Meets all requirements of "A" but may be deficient in some areas. Must attempt to control saturation by putting the same amount of water on each towel. Towels not saturated (key dimension for determining a "C" or below grade).
D	Student fails to saturate towels or control for saturation. Result is logically inconsistent with method used to saturate towels.
F	Student did not conduct the investigation Or, equipment manipulated without purpose Or, towels not wet Or, conclusions based on how towels felt.

*Criteria abridged from Baxter et al. (1992, p. 5).

Considerations in Selecting Dimensions

The dimensions you use to assess student performance in a certain domain should reflect the essential qualities of good performance in that domain. Where do you find these essential qualities? The qualities or dimensions can be provided by non-educator experts, colleagues in your department, grade level teachers, district curriculum committees, research literature, and national, state, or local subject area standards committees. If you are creating criteria for your own classroom, focus your criteria on those aspects of student performance that reflect your highest priority instructional goals and represent teachable and observable aspects of performance.

0095-9

One way to uncover dimensions for scoring criteria is to ask yourself the following kinds of questions:

- What are the attributes of good writing, of good scientific thinking, of good collaborative group process, of effective oral presentation? More generally, by what qualities or features will I know whether students have produced an excellent response to my assessment task?
- How does completing this task relate to my goals for students? What will they do that shows me we are working towards or achieving some of these goals?
- What do I expect to see if this task is done excellently, acceptably, poorly?
- Do I have samples or models of student work, from my class or other sources, that exemplify some of the criteria I might use in judging this task?
- What criteria for this or similar tasks exist in my state curriculum frameworks, my state assessment program, my district curriculum guides, my school assessment program?
- What dimensions might I adapt from work done by national curriculum councils, by other teachers?

In addition to describing your judgments about performance, the dimensions you use for your criteria need to be written so that all audiences who use them will understand them in the same way. Perhaps you are judging an interdisciplinary art project designed to reflect social studies understanding of the relationship of Native Americans to their environment. Your criteria for assigning grades or judging levels of performance should be clear to students, parents, and other teachers who depend on your judgments about content mastery; be they others at your grade level or those teaching your students next year.

Clear descriptions of performance dimensions can be achieved in several ways:

1. You could write definitions in terms of the behaviors or elements you will see when judging students. For example, instead of saying, "Acceptable performance means students show an understanding of living in harmony with the land," you could say, "Acceptable performance means that student drawings depict an environment that is almost unchanged from its original state. Few trees are cut; grassland is undisturbed except for small sustenance patches; no large waste dumps exist, and so on."
2. You could provide models or examples for each dimension. This

is commonly done in direct writing assessments. Teachers are given copies of student essays exemplifying each point in the score distribution. The essays illustrate such dimensions as, "the essay is well organized; it begins and ends effectively." From these, teachers and others can articulate precise definitions of each dimension.

3. If you are assessing informally, you could clarify your dimensions as a set of questions. For example, when you are assessing journals to see what kinds of help students need in developing fluency in writing, your criteria for deciding what to work on next could include the following questions: Which students are using some pre-writing strategies such as clustering, drawing, listing, or free-writing? Which students are keeping a log of writing ideas? Which students are having spelling problems that block the flow of ideas?

Unambiguous scale definitions usually consist of a description of the dimension to be rated, plus examples of student work illustrating acceptable responses. These models or work samples are crucial in developing a consensus about the meaning of criteria when used for rater training in formal assessments. Models also provide students with concrete examples of what acceptable or excellent work can look like. Figure 5.5 details one of several dimensions in a scoring rubric developed by CRESSST to assess the depth of high school students' understanding of history as revealed in their essays. Note that dimensions and scale points are thoroughly operationalized: key terms, such as "concept," are defined and examples of basic points, such as statements of opinion, are provided.

In most cases, your performance dimensions, particularly for classroom assessment, will reflect your views of what constitutes excellence or expertise and will be moderated by your expectations for students at different grade levels and by your instructional goals at different points in the school year. Because your criteria help students focus on what's important instructionally, you may use different criteria at different times during the school year. For example, while you may feel that organization and mechanics are an important part of expressing discipline-based knowledge in history or science, at the beginning of the year you may particularly want to encourage fluency. Thus, your criteria at the beginning of the semester will stress the number of ideas presented, number of examples or definitions for each idea, and so on. As students become more fluent and able to substantiate their views, you can expand your criteria to include organization and mechanics. To take an example from figure skating, you may believe in the Olympic criteria of "technical

0095-10

Figure 5.5
CREST Content Area Explanation
Essay Scoring Guidelines
 (Baker, Aschbacher, Niemi, and Sato 1992)

CREST Scoring Rubric Scales:
General Impression—Content Quality
Number of Principles or Concepts
Prior Knowledge: Facts and Events
Argumentation
Misconceptions
Text details

Example of Guidelines for the Number of Principles or Concepts Scale:

Number of Principles/Concepts

This is a measure of the number of different social studies concepts or principles that the student uses with comprehension.

A *concept* is an abstract, general notion, such as "inflation." It does not refer to particular events or objects (such as one particular period of inflation), but instead represents features common to a category of events or objects. "Imperialism," for example, does not refer to any specific facts or events; it is a heading that characterizes a class of behaviors and beliefs. "Industrialization" likewise identifies a class of activities and events that share common properties. It must be clear that the student is using a term conceptually, not just as a label.

A *principle* is a rule or belief used to justify an action or judgment, as in the statement "Slavery is immoral," where "morality" serves as a justifying principle.

It should be evident that the student understands the concept and means to discuss it. The concept should not simply be mentioned within a quotation from the text with no indication that the student grasps the concept. To earn a score point, the concept or principle need not be named explicitly, such as, "Constitutionality was an important principle that influenced the debate over slavery," but the idea should be stated clearly, for example, "One problem was determining what the constitution said about slavery."

Score point guidelines:

- 0—no response
- 1—one concept/principles
- 2—two concepts/principles
- 3—three concepts/principles
- 4—four or more concepts/principles

Example: "One great factor that held us back from war was our economy. It was not known what would happen to our economy without the safety of Britain. Britain could defend our commerce and coasts. Also, with Britain there was a great advantage with exportation. It seemed our economy could only suffer without the aid of Britain."

merit" and "artistic expression" but at different points in your teaching you may want to differentially emphasize one or the other.

Dimensions for Complex Tasks

As we mention in Chapter 4, it is entirely possible to create a complex assessment with multiple intended outcomes. Multiple outcomes require multiple criteria, a set for each outcome. Multidimensional criteria are unavoidable when you are doing interdisciplinary assessment or judging complex learning goals. You may either formulate separate criteria for each of these outcomes or create a multidimensional set of criteria. Connecticut's state assessment in science incorporates two approaches to assessing the same task by providing criteria for assessing group process and individual accomplishment (see Figures 5.6 and 5.7). Another perspective on student performance is provided by the subskills within the individual and group assessments. When examining group process skills, we are interested in scientific process, communication, and group collaboration. Separate criteria attend to each of these skills. The multiple dimensions on the individual scale include content and communication outcomes.

The dimensions for each scale require a lot of inference. Both teachers and students would need further descriptions of such dimensions as "draw reasonable conclusions" or "collaborate effectively" in order to use the scales. In fact, these scales are used in classrooms only after teachers have had inservice training to discuss the meaning of the dimensions, review examples, and practice using the criteria. Through classroom discussion and examples students and teachers come to a mutual understanding of the dimensions of the individual scale.

A less complex example of multidimensional criteria appears in Figure 5.1. The criteria assess four group performance outcomes: collaboration, critical thinking, communication, and history knowledge. The criteria include sub-criteria for deciding at which of five performance levels we should place students for each outcome. The entire set of group process criteria may be viewed as a compendium of four sets of criteria, one for collaboration, one for critical thinking, one for communication, and one for history knowledge.

Using Rating Scales

All sample scoring criteria included in this chapter contain some type of scale, either numerical, qualitative, or both. The criteria in Figure 5.1,

Figure 5.6
PART II: Objectives Rating Form — Group

Student I.D. #'s
 1. _____
 2. _____
 3. _____
 4. _____
 5. _____

Title of the Task: _____ Task # _____

Teacher ID #: _____ Date: _____

The group should be able to...	Where to Find Evidence				E	G	N.I.	*
	Group Report (Page #)	Oral Presentation	Teacher Observation	Other (Specify)				
1. Identify and apply physical and/or chemical properties for the purpose of identification.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
2. Formulate predictions based on prior knowledge.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
3. Identify information and steps needed to solve a problem.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
4. Test predictions.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
5. Gather data pertinent to a problem.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
6. Make inferences based on pertinent data.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
7. Draw reasonable conclusions and defend them rationally.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
8. Communicate the strategies and outcomes of a study through written means.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
9. Orally communicate the strategies and outcomes of a study.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
10. Collaborate effectively.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
* Check if students' work is a strong and clear example of rating given.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

(Connecticut Department of Education 1990) E = Excellent G = Good N.I. = Needs Improvement

62

012000

Figure 5.7
PART II: Objectives Rating Form — Individual

Title of the Task: _____ Task # _____ Student ID # _____

Teacher ID #: _____ Date: _____

The group should be able to...	Where to Find Evidence				E	G	N.I.	*
	Group Report (Page #)	Oral Presentation	Teacher Observation	Other (Specify)				
1. Identify and apply physical and/or chemical properties for the purpose of identification.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
2. Identify information and steps needed to solve a problem.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
3. Communicate the strategies of a study through written means.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
* Check if students' work is a strong and clear example of rating given.					<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

(Connecticut Department of Education 1990) E = Excellent G = Good N.I. = Needs Improvement

63

0095-11

0095512

the history group process, and Figure 5.3, the mathematics problem, contain both numerical and qualitative rating scales. Figure 5.4, the hands-on science criteria, and Figures 5.6 and 5.7, the group and individual science experiment, have qualitative ratings only, such as letter grades or evaluations such as "excellent" or "needs improvement."

Why scales? How do you know whether to use numerical or qualitative ratings? What about using a checklist instead of a rating scale? Whether you rate the presence or absence of a performance, as in a checklist, or use numbers or qualitative evaluations will depend on your testing purpose. There are three major types of scales: checklists, numerical ratings, and qualitative (either descriptive or evaluative) ratings. If your purpose is to **describe** what students can do, perhaps for parent conferences or to compare student performance to certain developmental standards, you may be able to use the simplest rating scale of all, the checklist. If you need more information than simply whether or not a student is engaged in specific aspects of a task, you will need a more fully developed rating scale. When you want to know the **extent** to which dimensions were observed or the **quality** of the performance, you need more elaborate scales. Rating scales, beyond the yes-no checklist format, reflect aspects of student performance other than mere accomplishment of an activity.

Checklists

A checklist is a list of dimensions, characteristics, or behaviors that are essentially scored as "yes-no" ratings. A check indicates that either the characteristic or behavior was present or absent. Checklists often contain more dimensions to be scored than do rating scales, but those dimensions are often quite narrow and concrete.

Checklists can be useful in assessing processes, an important purpose for teachers concerned with the how as well as the what of learning. A process checklist for a hands-on experiment could resemble Figure 5.8, which asks the rater to note the presence of specified behaviors.

Primary school teachers find checklists useful because they must often determine how students are developing according to some theory of skills acquisition. For example, current language acquisition theory suggests that this skill cluster supports a child's ability to read:

- Ability to draw or depict an idea
- Ability to recognize sound-letter correspondence
- Ability to recognize that words stand for something

- Knowledge of left to right and up-to-down page orientation
- Ability to recall and retell favorite stories

Figure 5.8
Process Checklist

Procedure	Check if Observed	Comments
Selected approach		
Correct equipment used		
Measurement accurate		
Sought peer help if needed		
Recorded observations		
Cleaned up after experiment		

The teacher can document acquisition of these readiness skills with a checklist. There is no need to judge how well each of these behaviors are displayed, only that they are in place. Figure 5.2 demonstrates a developmentally-based profile for kindergartners created by teachers of the Soledad Union School District in California, with consultation from Pacific Oaks College in Pasadena, California. This is an example of a theory-based profile. The profile development process was designed to help staff better understand constructivism, the developmental learning theory on which it is based. The behaviors identified in Figure 5.2 are sequenced from left to right in the order that the kindergarten staff predicted that those behaviors are acquired. This document was designed to be re-analyzed each year as teachers observe children's behaviors from a developmental point of view.

Numerical Scales

A numerical scale uses numbers or assigns points to a continuum of performance levels. The length of the continuum or the number of scale points can vary, three points, four points, five points, seven points—any number is possible. How many divisions or scale points should a good

scale include? While there's no single answer to this question, our experience suggests that you consider these issues.

The number of points or divisions on a scale can and should vary depending on what decisions you will be making about students and whether the scale will be used in the classroom or in a formal scoring session with several raters involved in judging performance. In general, the larger the scale, the more difficult it is to clearly differentiate among the score points. Consider how quickly you can sort essays into stacks worthy of zero points, one point, or two points; essentially a decision among low, medium, and high. Why use a ten-point scale if you really only want to distinguish two or three groups of students, such as those who need additional instruction on writing a well-organized essay and those who don't?

A scale with only a few points does have some disadvantages. More scale points enable you to identify small differences between individual students and may provide more diagnostic information than a reduced scale. For example, a longer scale may be needed if you want to use one scale for all students K-12 and you also want to differentiate among students in a single grade. Also, if your scale will be used for formal assessment purposes where several readers will be rating each performance, any statistics you have to calculate, such as rater agreement, will be affected by the scale range. Using a shorter scale will result in a high percent agreement, but it will be more difficult to achieve a high correlation between raters' scores (two different ways of figuring inter-rater reliability).

It takes longer to arrive at consensus about how to assign scale points when there are more points to consider. With a five- or six-point scale, raters often refer to prior experience and assign the lowest points to off-task or truly terrible performances, the highest to stellar examples, reserve the middle for "passing," "acceptable," or model performances, then allocate those not fitting into the three anchor points to the remaining scale values. An eleven- or seventeen-point scale makes it more difficult for raters to anchor their judgments in prior experience. However, you will often see scales in multiples of five, such as ten, fifteen, or twenty point scales, which allow readers to "chunk" the points into five-point intervals. Initial rating distinctions are then really made between a five and a ten rather than a four and a seven with examples not clearly fitting into the increments receiving the intermediate points.

Another consideration related to scale size concerns multidimensional criteria. If you are rating the same performance with several criteria, each assessing a different outcome, you may want to use the same number of scale points for each outcome. Not only does this make it possible to aggregate or compare the results of several scales, but it

eases the rating task. For example, using a four-point scale for coherence and a five-point scale for supporting facts could slow the rating process while raters mentally shift to different scale points. Students trying to understand their relative strengths and weaknesses can also have difficulty comparing different scales. However, if you want some outcomes to count more than others for a total score, you can use different size scales to reflect relative value or weight. A good example of this strategy appears in Figure 5.1, the history group process task. The scoring guide uses two different scales with one set of outcomes "weighted" up to twenty points and the other up to thirty.

Qualitative Scales

A qualitative scale uses adjectives rather than numbers to characterize student performance. These scales are of two general sorts, descriptive and evaluative. Descriptive scales label student performance but don't necessarily make explicit the standards underlying the judgment; they use fairly neutral terms to characterize performance. Judgments about task completion, task understanding, or the appearance of certain elements in the performance are typical descriptors. Figure 5.9 provides three examples of descriptive scales that do not evaluate the worth of student performance.

Figure 5.9
Descriptive Scales

No evidence...Minimal evidence...Partial evidence...Complete evidence.
Task not attempted...Partial completion...Completed...Goes beyond.
Off task...Attempts to address task...Minimal attention to task...Addresses task but no elaboration...Fully elaborated and attentive to task and audience.

Evaluative scales incorporate judgments of worth anchored in underlying standards of excellence. The most commonly used evaluative scales are grades (see Figure 5.4). Scales using descriptors of "excellence" (Figures 5.1, 5.6, and 5.7) or judging competence (Figure 5.3) are

evaluative in nature. Evaluative scales require higher levels of inference to interpret than descriptive scales. The inferences are made by referring directly to the scoring criteria. The criteria themselves embed notions of excellence, competence, or acceptable outcomes.

Numerical-Qualitative Scales

Numerical scales are often easier for people to remember, to aggregate, and to average, but are difficult to interpret in the absence of good descriptors. After all, a score of "4" on a six-point scale may connote different levels or qualities of attainment to different people. Good criteria often include both descriptive and numerical values. For example, Figure 5.3 displays a draft of a scale used by the California Assessment Program for judging open-ended math problems. Note that it is both numeric and descriptive. Performance is rated numerically, but each numerical score is attached to an evaluation ranging from "inadequate" to "competent."

Whether your scale values are numerical, descriptive, or both, it is important to make sure that scales help parents, students, teachers, administrators, and policymakers understand the meaning of the performance in the same way. This common understanding helps ensure reliable and fair judgments.

The Link with Standards

Nearly all criteria, even descriptive checklists, are linked in some way to standards—the expectations for student performance. Grades or qualitative ratings reflect teacher judgment, or in the case of the hands-on science criteria in Figure 5.4, the consensus of the rating team. The standards underlying different scales may reflect either criterion-referenced or norm-referenced approaches to judging quality. The mathematics criteria (Figure 5.3) with descriptors for "inadequate response," "satisfactory response," and "demonstrated competence," reflect an absolute standard or mastery approach to standard setting. The descriptors clearly indicate good or desired performance levels, "satisfactory and above," versus poor levels, "inadequate." The levels are referenced to discipline-based standards, mathematics teachers' conceptions of adequate problem-solving strategies.

Another example is Illinois' six-point writing assessment scale, which employs an absolute scale and is designed to be used across grade

levels. A score of six represents an extremely high level of writing, and few if any elementary students are expected to score above a "3." This type of scale is especially useful in measuring growth over years. The limitation of an absolute scale for multigrade/age assessment is that because elementary students all tend to score near the bottom of the scale, there is little variability in their scores so it is impossible to tell much about them individually from their scores. They all "look alike."

Other evaluative scales reflect norm-referenced approaches to standard setting. When grades or points are assigned by comparing students' relative status, such as, "Maria's essay was better than the class average," "Gary's video was among the best in the class," the standards are norm-referenced. Developmental checklists or scales demonstrate another common use of norm-referenced scales in alternative assessment. The sequencing of behaviors in these scales rests on what educators and others have observed over time to be typical performance at specified ages. For example, children who score "average" in reading readiness demonstrate behaviors typical for their age or grade level. "Below average" or "developmentally delayed" refers to performance typical of children in a younger age group than those being assessed.

It is possible to anchor standards in both criterion- and norm-referenced information for the same assessment. You start with a criterion-referenced scale, a scale describing performance relative to a clearly defined set of behaviors, then gather or otherwise obtain data about how a national, state, or local sample of students performed on the same measure. You can then say "Maria wrote a well-organized essay, receiving a '4' in organization; her performance was described as better than 75 percent of the students in the state." Or, on a more informal level, in your classroom, you can always describe an individual student's performance level in comparison to the rest of the class's performance: "Maria's score put her among the best in the class."

Some scales may look like absolute or criterion-referenced scales but might actually incorporate both norm- and criterion-referenced information. An age- or grade-related scale defines student performance in terms of benchmarks or expectations for a particular grade level. Benchmarks for 5th grade mathematics problem solving will differ from those for the 7th grade. What constitutes excellence in essay organization at the 8th grade will not do so at the 11th. Despite their "criterion-referenced" appearance, scales tied to an age or grade level curriculum have an underlying norm-referenced interpretation. The dimensions themselves were derived from what students were able to do at particular grades, not from absolute standards of performance across ages and grades. For practical purposes, these grade level scales are considered criterion-referenced because their primary use is to decide what students can do

vis-a-vis particular content and skills rather than to compare them to each other.

How can you get the best of both worlds? By determining appropriate standards according to your assessment purposes. For classroom or schoolwide assessment use, you'll probably lean toward criterion-referenced or absolute standards. For selection decisions in which there are more candidates than available space, you will probably use absolute standards for inclusion in the candidate pool, but normative standards for the final selection. For example, if you are selecting horn players for the after-school honors band, you will choose only the top 2 percent of the candidates.

We have not discussed how standards are set. How do you know where to set the acceptable level of performance? How good is competent? What is the cut point between barely satisfactory and satisfactory? High-stakes assessments, such as graduation certification, use formal standard-setting procedures. These may include using a group of judges, provided with norm- and criterion-referenced information, to determine a passing score. In district or schoolwide assessment, passing scores or labels describing poor and excellent performance are determined by consensus of those using the assessment. In the classroom, teachers set standards based on their experiences, their knowledge of what students have done in the past, their familiarity with expectations in a discipline, the current performance of students, and the purpose of the assessment.

Considering Other Choices: Holistic or Analytic Criteria*

Based on experience with direct writing assessment, we offer two more choices in specifying criteria: holistic and analytic. Holistic criteria require raters to assign a single score based on the overall quality or to one aspect of the student's response. An analytic scale requires that raters give separate ratings to different aspects of the work. Criteria incorporating several outcomes are analytic.

*You may be familiar with the term "Primary Trait Scoring." When Primary Trait criteria focus on only one trait, they are holistic; when expanded to two or more traits, they become analytic.

0095-15

Which Is Better?

By this time you know that we're going to say "it depends on the purpose of the assessment." The pattern of results from an analytic scale provides useful feedback about the strengths and weaknesses of the individual student and the classroom instructional program. Unfortunately, because student performance on different dimensions of an analytic scale may be related in complex ways, the results may not be as clearly diagnostic as desired. Despite the fact that one of the qualities of a good analytic scale, from an efficiency and measurement perspective, is that each dimension be distinct, the subscale scores are often highly interrelated and not well differentiated. CRESSST research on analytic scoring scales found high correlations among scores for overall essay and paragraph organization, and between organization, support, and a general competence score. Under such circumstances, the diagnostic value of subscale performance is greatly diminished.

Holistic scoring is usually simpler and faster than analytic; an important concern when teacher time is involved. Unless assessment's purpose is not to provide data to guide program improvement, a quick overview of achievement may be particularly suitable for program evaluation, for flagging students who need more help, and for assigning final evaluations.

Concurrent use of analytic and holistic strategies can optimize both diagnostic value and efficiency. One approach emerging from minimum competency testing is to score all essays holistically then rate analytically those essays that were scored below minimum competency. Another strategy, used in the Maine statewide assessment, is to score essays holistically, but to note analytic dimensions that are particularly strong or weak in an individual's work as a kind of generic "comment" on the performance.

Opinions differ considerably regarding the value of these different approaches, and research is ongoing. The important point is not so much the correct labeling of scales, but that a variety of approaches exist and can prove useful.

What About Portfolio Assessment?

Portfolio assessment is often the first strategy that comes to mind when people think of alternative assessments. In some respects, portfolio assessment is a misnomer for "assessment of a body of work." In other instances, the portfolio assessment is really the assessment system.

Portfolios are collections of student work that are reviewed against criteria in order to judge an individual student or a program. The portfolio or collection of work does not constitute the assessment; it is simply a receptacle for work (essays, videotapes, art, journal entries, and so on) that may or may not be evaluated. The "assessment" in portfolio exists only when (1) an assessment purpose is defined; (2) criteria or methods for determining what is put into the portfolio, by whom, and when, are explicated; and (3) criteria for assessing either the collection or individual pieces of work are identified. Deciding what should be included is really a task description, not a scoring guideline problem. What goes in, who chooses, when samples are taken—these are dimensions of the assessment task that define the setting and kinds of work that will be considered. (See Chapter 7 for more discussion of portfolio assessment.)

There are two issues related to selecting the dimensions of scoring criteria for portfolio assessment: (1) What are the criteria for selecting the samples that go into the portfolio, and (2) What are the criteria for judging the quality of the samples? Prior to considering criteria for judging portfolios, you will need to determine whether the portfolio should be rated as a whole or as individual samples. Second, you need to decide which dimensions reflect the intent or purpose of your assessment. When looking at a body of work, many issues arise, for example:

- Will progress or improvement be assessed?
- How or will progress be evaluated?
- How will different tasks, videos, art work, essays, journal entries, and the like be compared or weighted in the assessment?
- What is the role of student reflection in the assessment? Parental input?

Once these issues are settled, defining the dimensions of portfolio scoring criteria is the same as defining multidimensional criteria. Perhaps the best known example of portfolio assessment criteria is provided by the Vermont Mathematics portfolio, which is summarized in Figure 5.10. A body of mathematics work is evaluated on two major dimensions, problem-solving and communication skill. Within each dimension, several subdimensions further define each of the larger skills. Ratings are given for the subskills under the two dimensions, problem solving and communication. You can see how this example of portfolio assessment criteria resembles the multidimensional examples in Figures 5.1 and 5.7.

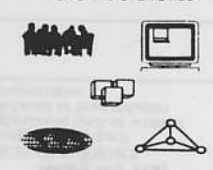
Figure 5.10
Mathematics Rating Form

Student: _____ ID Number: _____ School: _____ Grade: _____ Date: _____ Rater: _____	A1 Understanding of Task SOURCES OF EVIDENCE <ul style="list-style-type: none"> • Explanation of task • Reasonableness of approach • Correctness of response leading to inference of understanding 	A2 How—Quality of Approaches/Procedures SOURCES OF EVIDENCE <ul style="list-style-type: none"> • Demonstrations • Descriptions (oral or written) • Drafts, scratch work, etc. 	A3 Why—Decisions Along the Way SOURCES OF EVIDENCE <ul style="list-style-type: none"> • Changes in approach • Explanations (oral or written) • Validation of final solution • Demonstration
ENTRY 1 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 2 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 3 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 4 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 5 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 6 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ENTRY 7 Title: _____ P Puzzle I Investigation A Application O Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
OVERALL RATINGS →	UNDERSTANDING OF TASK FINAL RATING <ul style="list-style-type: none"> 1 Totally misunderstood 2 Partially understood 3 Understood 4 Generalized, applied, extended 	HOW—QUALITY OF APPROACHES/PROCEDURES FINAL RATING <ul style="list-style-type: none"> 1 Inappropriate or unworkable approach/procedure 2 Appropriate approach/procedure some of the time 3 Workable approach/procedure 4 Efficient or sophisticated approach/procedure 	WHY—DECISIONS ALONG THE WAY FINAL RATING <ul style="list-style-type: none"> 1 No evidence of reasoned decision-making 2 Reasoned decision-making possible 3 Reasoned decisions/adjustments inferred with certainty 4 Reasoned decisions/adjustments shown/explicated
Comments: _____			

0095-16

41-28000

Figure 5.10 (continued)

A4 What—Outcomes of Activities	B1 Language of Mathematics	B2 Mathematical Representations	B3 Clarity of Presentation	CONTENT TALLIES
SOURCES OF EVIDENCE • Solutions • Extensions—observations, connections, applications, syntheses, generalizations, abstractions	SOURCES OF EVIDENCE • Terminology • Notation/symbols	SOURCES OF EVIDENCE • Graphs, tables, charts • Models • Diagrams • Manipulatives	SOURCES OF EVIDENCE • Audio/video tapes (or transcripts) • Written work • Teacher interviews/observations • Journal entries • Student comments on cover sheet • Student self-assessment	Number Sense—Whole No./Fractions (4) Number Relationships/No. Theory (2)
				Operations/Place Value (4) Operations (8)
				Estimation (4/8)
				Patterns/Relationships (4) Patterns/Functions (8)
				Algebra (8)
				Geometry/Spatial Sense (4/8)
				Measurement (4/8)
				Statistics/Probability (4/8)
				TASK CHARACTERISTICS 
WHAT—OUTCOMES OF ACTIVITIES FINAL RATING <input type="checkbox"/> 1 Solution without extensions <input type="checkbox"/> 2 Solution with observations <input type="checkbox"/> 3 Solution with connections or applications <input type="checkbox"/> 4 Solution with synthesis, generalization or abstraction	LANGUAGE OF MATHEMATICS FINAL RATING <input type="checkbox"/> 1 No or inappropriate use of mathematical language <input type="checkbox"/> 2 Appropriate use of mathematical language some of the time <input type="checkbox"/> 3 Appropriate use of mathematical language most of the time <input type="checkbox"/> 4 Use of precise, elegant, appropriate mathematical language	MATHEMATICAL REPRESENTATIONS FINAL RATING <input type="checkbox"/> 1 No use of mathematical representation(s) <input type="checkbox"/> 2 Use of mathematical representation(s) <input type="checkbox"/> 3 Accurate and appropriate use of mathematical representation(s) <input type="checkbox"/> 4 Perceptive use of mathematical representation(s)	CLARITY OF PRESENTATION FINAL RATING <input type="checkbox"/> 1 Unclear (e.g., disorganized, incomplete, lacking detail) <input type="checkbox"/> 2 Some clear parts <input type="checkbox"/> 3 Mostly clear <input type="checkbox"/> 4 Clear (e.g., well organized, complete, detailed)	EMPOWERMENT COMMENTS Motivation Risk Taking Confidence Curiosity/Interest Flexibility Reflecting Perseverance Value Math

74

0095-17

Developing and Evaluating Scoring Criteria

Beginning the Development Process

The process for developing your own criteria is straightforward:

- Investigate how the assessed discipline defines quality performance.
- Gather sample rubrics for assessing writing, speech, the arts, and so on as models to adapt for your purposes.
- Gather samples of students' and experts' work that demonstrate the range of performance from ineffective to very effective.
- Discuss with others the characteristics of these models that distinguish the effective ones from the ineffective ones.
- Write descriptors for the important characteristics.
- Gather another sample of students' work.
- Try out criteria to see if they help you make accurate judgments about students.
- Revise your criteria.
- Try it again until the rubric score captures the "quality" of the work.

You probably noticed how recurrent this development process is. Initial ideas about important and scorable aspects of student performance become refined through use. Your criteria may focus on process—how a student approaches and solves a problem—as well as on the product or outcomes.

For example, we can refer to the development process for the criteria in Figure 5.5 (Baker, Aschbacher, Niemi, and Sato 1992). CRESST developed its rubric for rating depth of content understanding in history by collecting and examining the differences in essays written by history experts (university professors and graduate students in history) versus those written by novices (high school students). CRESST researchers looked for dimensions that seemed to differentiate the performance of these two groups. In a number of subject areas, the researchers observed differences between the students and the experts in the application of prior knowledge, the use of organizing concepts and principles, and misconceptions. These traits defined the first draft of scoring criteria. The criteria were then tried out on samples of student work and further clarified and refined to ensure that the scales were clearly defined, were

appropriate for the range of student responses likely to be encountered, and enabled teachers or other raters to distinguish between essays that deserved adjacent points on a scale.

While undertaking the task of developing criteria, don't forget to take advantage of others' work. Quite often you can import or modify criteria from state and local assessment programs, curriculum experts, or colleagues who have grappled with similar assessment problems. Research literature on alternative assessment also provides examples of pilot alternative assessments similar to the one appearing in Figure 5.4, which can be adapted for classroom use. There is also a small but growing literature on the nature of expertise in various disciplines, such as how an historian reads and uses primary source documents.

Evaluating Criteria

Your criteria for judging students' work shape the decisions you eventually make about programs and students. Regardless of whether you are developing your own criteria or using those provided by others, it is important to review the quality of the scoring guidelines. We conclude this chapter with a proposed set of "criteria for criteria"—a checklist you can use to rate the quality of scoring criteria you borrow or develop. Our proposed criteria appear in Figure 5.11.

Now let's look at a set of dimensions for assessing the worth of your own criteria.

Keyed to Important Outcomes

At a minimum, criteria for judging student performance need to address all the student outcomes you are trying to measure. For example, your criteria for judging student drama productions should encompass all the important drama and art that you want to be able to assess, and no others. If originality and logical presentation are part of the desired outcomes, you will want to include scales for judging these aspects of student work. If they are not an important outcome, omit them.

Sensitive to Purpose

What educational decisions will you make on the basis of your assessment? The answer to this question should guide your decisions about

whether to use a checklist or rating scales, how many scales, which traits, what types of scale, and so forth. Do you need a global, holistic view of student achievement or an analytical one that gives you information about several specific aspects of students achievement? Do you need the information in the form of a number for ease of reporting and aggregation at the expense of detail, or do you need the richness of qualitative description, or perhaps both?

Figure 5.11
How Do You Evaluate Scoring Criteria?

- All important outcomes are addressed by criteria.
- Rating strategy matches decision purpose: holistic for global, evaluative view; analytic for diagnostic view.
- Rating scale provides usable, easily interpreted score.
- Criteria employ concrete references, clear language understandable to students, parents, other teachers.
- Criteria reflect current conceptions of "excellence" accepted in the field.
- Criteria have been reviewed for developmental, ethnic, gender bias.
- Criteria reflect teachable outcomes.
- Criteria are limited to feasible number of dimensions.
- Criteria are generalizable to other similar tasks or larger performance domain.

Meaningful, Clear, and Credible

The criteria by which you judge a performance need to be meaningful to students, parents, raters, teachers, administrators, policymakers, and the public. If the criteria are not credible, the results will probably be ignored or may be misused. Examples of student work that illustrate criterion traits can help make the criteria concrete for others. Involving others in the development of criteria increases their credibility.

Because one of the tenets of performance assessment is public and discussed criteria, your criteria need to make sense to students so that

they will be able to apply them easily to their own work and become self-regulated learners. Although judgments of student performance tend to be subjective by their nature, they are more reliable and credible when they rely less on high inference and more on observable, concrete characteristics.

Fair and Unbiased

Not only do assessment tasks need to be fair, but so do the criteria by which you define excellence. Unrecognized biases can seep into your definitions of traits, your specifications for what kind of performance earns which scale point, and your application of those criteria to individual pieces of student work. When you want your criteria to have diagnostic value, they must be sensitive to instruction and students' opportunities to learn the skills that are assessed. In contrast, you do not want them to reflect variables over which educators have no control, such as a child's culture, sex, or socioeconomic background.

Feasible

Several reasons exist to limit the number and complexity of the performance dimensions to be judged. First, the time, effort, and money available for judging performance are always limited, sometimes severely so. Second, raters find it difficult to address too many different aspects of a work at once. In our experience at CRESSST, raters were frustrated when asked to use more than six or seven scales for rating student essays. It became an onerous task and a less reliable process. Third, students will probably find it difficult to deal with too many aspects of their work at once. And finally, administrators and policymakers usually need information in as brief a form as possible. Separate scores for a large number of traits or for complex characteristics may make it more difficult to use the results effectively.

Generalizable

Although we recognize that criteria for performance are strongly linked to discipline-based notions of excellence, rating can be more efficient when a single set of "generic" criteria can serve multiple topics, tasks, or disciplines. For example, we could develop a common set of criteria

0095-19

for assessing student understanding of science concepts through journals, hands-on experimentation, computer simulation, and oral presentation. We could also use a common set of criteria for judging student essays in social studies, science, and math? As disparate as these situations may seem, it is possible to envelop generic criteria for some purposes. If we could conceptualize excellence in consistent ways across assessment methods and disciplines, our criteria could have a more powerful impact on learning and instruction. Our example of the CRESSST history-social studies rubric (Figure 5.5), which has also been applied to science and economics, shows one strategy for developing cross-discipline criteria. Like all good criteria, these proposed dimensions are subject to revision and refinement.

References

- Baker, E.L., P.R. Aschbacher, D. Niemi, and E. Sato. (1992). *CRESSST Performance Assessment Models: Assessing Content Area Explanations*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.
- Baxter, G., R.J. Shavelson, S. Goldman, and J. Pine. (Spring 1992). *Journal of Educational Measurement* 29, 1: 1-17.
- Jones, E., and J.M. Roberts. (1990). *Profile of Developmental Outcomes for Kindergarten Literacy and Numeracy Skills*. Soledad, Calif.: San Vincente School District.
- Vermont Department of Education. (1992). *Looking Beyond "the Answer": The Report of Vermont's Mathematics Portfolio Assessment Program, Pilot Year 1990-91*. Montpelier: Vermont Department of Education.